



## Volume 8, Issue 6, June 2018

The Constraint Satisfaction Problem: Complexity and Approximability (Dagstuhl Seminar 18231)	
<i>Martin Grohe, Venkatesan Guruswami, and Stanislav Zivny</i> .....	1
High-Performance Graph Algorithms (Dagstuhl Seminar 18241)	
<i>Henning Meyerhenke, Richard Peng, and Ilya Safro</i> .....	19
Secure Routing for the Internet (Dagstuhl Seminar 18242)	
<i>Phillipa Gill, Adrian Perrig, and Matthias Wählisch</i> .....	40
Database Architectures for Modern Hardware (Dagstuhl Seminar 18251)	
<i>Peter A. Boncz, Goetz Graefe, Bingsheng He, and Kai-Uwe Sattler</i> .....	63
Ubiquitous Gaze Sensing and Interaction (Dagstuhl Seminar 18252)	
<i>Lewis Chuang, Andrew Duchowski, Pernilla Qvarfordt, and Daniel Weiskopf</i> .....	77
Discipline Convergence in Networked Systems (Dagstuhl Seminar 18261)	
<i>Yungang Bao, Lars Eggert, Simon Peter, and Noa Zilberman</i> .....	149
10 Years of Web Science: Closing The Loop (Dagstuhl Perspectives Workshop 18262)	
<i>Susan Halford, James A. Hendler, Eirini Ntoutsi, and Steffen Staab</i> .....	173

ISSN 2192-5283

*Published online and open access by*

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

*Publication date*

January, 2019

*Bibliographic information published by the Deutsche Nationalbibliothek*

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

*License*

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



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

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

The copyright is retained by the corresponding authors.

*Aims and Scope*

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

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

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

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

*Editorial Board*

- Gilles Barthe
- Bernd Becker
- Daniel Cremers
- Stephan Diehl
- Reiner Hähnle
- Lynda Hardman
- Hannes Hartenstein
- Oliver Kohlbacher
- Bernhard Mitschang
- Bernhard Nebel
- Bernt Schiele
- Albrecht Schmidt
- Raimund Seidel (*Editor-in-Chief*)
- Emanuel Thomé
- Heike Wehrheim
- Verena Wolf

*Editorial Office*

Michael Wagner (*Managing Editor*)  
Jutka Gasiorowski (*Editorial Assistance*)  
Dagmar Glaser (*Editorial Assistance*)  
Thomas Schillo (*Technical Assistance*)

*Contact*

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

Digital Object Identifier: 10.4230/DagRep.8.6.i



# The Constraint Satisfaction Problem: Complexity and Approximability

Edited by

Martin Grohe<sup>1</sup>, Venkatesan Guruswami<sup>2</sup>, and Stanislav Živný<sup>3</sup>

1 RWTH Aachen, DE, [grohe@informatik.rwth-aachen.de](mailto:grohe@informatik.rwth-aachen.de)

2 Carnegie Mellon University – Pittsburgh, US, [guruswami@cmu.edu](mailto:guruswami@cmu.edu)

3 University of Oxford, GB, [standa.zivny@cs.ox.ac.uk](mailto:standa.zivny@cs.ox.ac.uk)

---

## Abstract

Constraint satisfaction has always played a central role in computational complexity theory; appropriate versions of CSPs are classical complete problems for most standard complexity classes. CSPs constitute a very rich and yet sufficiently manageable class of problems to give a good perspective on general computational phenomena. For instance, they help to understand which mathematical properties make a computational problem tractable (in a wide sense, e.g., polynomial-time solvable, non-trivially approximable, fixed-parameter tractable, or definable in a weak logic). In the last decade, research activity in this area has significantly intensified and hugely impressive progress was made. The Dagstuhl Seminar 18231 “The Constraint Satisfaction Problem: Complexity and Approximability” was aimed at bringing together researchers using all the different techniques in the study of the CSP so that they can share their insights obtained during the past three years. This report documents the material presented during the course of the seminar.

**Seminar** June 3–8, 2018 – <http://www.dagstuhl.de/18231>

**2012 ACM Subject Classification** Theory of computation → Problems, reductions and completeness

**Keywords and phrases** Constraint satisfaction problem (CSP); Computational complexity; CSP dichotomy conjecture; Hardness of approximation; Unique games conjecture; Parameterised complexity; Descriptive complexity; Universal algebra; Logic; Semidefinite programming.

**Digital Object Identifier** 10.4230/DagRep.8.6.1

**Edited in cooperation with** Peter Fulla


## 1 Executive Summary

*Martin Grohe*

*Venkatesan Guruswami*

*Dániel Marx*

*Stanislav Živný*

**License**  Creative Commons BY 3.0 Unported license

© Martin Grohe, Venkatesan Guruswami, Dániel Marx, and Stanislav Živný

The *constraint satisfaction problem*, or CSP in short, provides a unifying framework in which it is possible to express, in a natural way, a wide variety of computational problems dealing with mappings and assignments, including satisfiability, graph colourability, and systems of equations. The CSP framework originated 30–35 years ago independently in artificial intelligence, database theory, and graph theory under three different guises, and it was



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

The Constraint Satisfaction Problem: Complexity and Approximability, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 1–18

Editors: Martin Grohe, Venkatesan Guruswami, and Stanislav Živný



Dagstuhl Reports

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

realised only in the late 1990s that these are in fact different faces of the same fundamental problem. Nowadays, the CSP is extensively used in theoretical computer science, being a mathematical object with very rich structure that provides an excellent laboratory both for classification methods and for algorithmic techniques; while in AI and more applied areas of computer science, this framework is widely regarded as a versatile and efficient way of modelling and solving a variety of real-world problems, such as planning and scheduling, software verification, and natural language comprehension, to name just a few. An instance of the CSP consists of a set of variables, a set of values for the variables, and a set of constraints that restrict the combinations of values that certain subsets of variables may take. Given such an instance, the possible questions include (a) deciding whether there is an assignment of values to the variables so that every constraint is satisfied, or optimising such assignments in various ways, or (b) finding an assignment satisfying as many constraints as possible. There are many important modifications and extensions of this basic framework, e.g., those that deal with counting assignments or involve soft or global constraints.

Constraint satisfaction has always played a central role in computational complexity theory; appropriate versions of CSPs are classical complete problems for most standard complexity classes. CSPs constitute a very rich and yet sufficiently manageable class of problems to give a good perspective on general computational phenomena. For instance, they help to understand which mathematical properties make a computational problem tractable (in a wide sense, e.g., polynomial-time solvable, non-trivially approximable, fixed-parameter tractable, or definable in a weak logic). One of the most striking features of this research direction is the variety of different branches of mathematics (including universal algebra and logic, combinatorics and graph theory, probability theory and mathematical programming) that are used to achieve deep insights in the study of the CSP. In the last decade, research activity in this area has significantly intensified and hugely impressive progress was made.

The recent flurry of activity on the topic of the seminar is witnessed by four previous Dagstuhl seminars, titled “Complexity of constraints” (06401) and “The CSP: complexity and approximability” (09441, 12541, 15301), that were held in 2006, 2009, 2012, and 2015 respectively. This seminar was a follow-up to the 2009, 2012, and 2015 seminars. Indeed, the exchange of ideas at the 2009, 2012, and 2015 seminars has led to ambitious new research projects and to establishing regular communication channels. There is clearly the potential for further systematic interaction that will keep on cross-fertilising the areas and opening new research directions. The 2018 seminar brought together 47 researchers from different highly advanced areas of constraint satisfaction and involved many specialists who use universal-algebraic, combinatorial, geometric, and probabilistic techniques to study CSP-related algorithmic problems. The participants presented, in 24 talks, their recent results on a number of important questions concerning the topic of the seminar. One particular feature of this seminar is a significant increase in the number of talks involving multiple subareas and approaches within its research direction – a definite sign of the growing synergy, which is one of the main goals of this series of seminars.

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

## Description of the Topics of the Seminar

**Classical computational complexity of CSPs.** Despite the provable existence of intermediate problems (say, between P and NP-complete, assuming  $P \neq NP$ ), research in computational complexity has produced a widely known informal thesis that “natural problems are almost always complete for standard complexity classes”. CSPs have been actively used to support and refine this thesis. More precisely, several restricted forms of the CSP have been investigated in depth. One of the main types of restrictions is the *constraint language* restriction, i.e. a restriction on the available types of constraints. By choosing an appropriate constraint language, one can obtain many well-known computational problems from graph theory, logic, and algebra. The study of the constraint language restriction was driven by the *CSP Dichotomy Conjecture* of Feder and Vardi which states that, for each fixed constraint language, the corresponding CSP is either in P or NP-complete. There are similar dichotomy conjectures concerning other complexity classes (e.g., L and NL). Recent breakthroughs in the complexity of the CSP have been made possible by the introduction of the universal-algebraic approach, which extracts algebraic structure from the constraint language and uses it to analyse problem instances. The above conjectures have algebraic versions which also predict in algebraic terms where the boundary between harder problems and easier problems lies. The algebraic approach has been applied to prove the Dichotomy Conjecture in many important special cases (e.g., Bulatov’s dichotomy theorems for 3-valued and conservative CSPs), culminating in two independent proofs of the general conjecture announced in 2017 by Bulatov and Zhuk.

- Bulatov and Zhuk gave detailed talks on the main insights into their proofs.
- Kolmogorov described an algorithm for Boolean CSPs under the restriction that every variable appears in exactly two constraints and all constraints are even  $\Delta$ -matroids.

The valued CSP (VCSP) is a significant generalisation of the CSP that involves both feasibility and optimisation aspects. While the computational complexity of finite-domain VCSPs is by now well understood, the infinite-domain VCSPs are fairly unexplored.

- Viola gave a talk on submodular VCSPs on infinite domains.
- Kazda presented his results on the structure of weighted clones, which are intimately related to the computational complexity of VCSPs.

**Approximability of CSPs.** The use of approximation algorithms is one of the most fruitful approaches to coping with NP-hardness. Hard optimisation problems, however, exhibit diverse behavior with respect to approximability, making it an exciting research area that is by now well-developed but far from fully understood.

An emerging topic bridging the complexity of the CSP with approximation aspects is *promise constraint satisfaction* (PCSP). The PCSP is a generalization of the CSP in which the constraints come in pairs of “stricter” and “weaker” versions. In a PCSP instance, the task is to find an assignment satisfying the weaker constraints under the promise that there is an assignment satisfying the strict constraints.

- Brakensiek gave an introductory talk to this exciting research direction and also presented a dichotomy classification for symmetric Boolean PCSPs.
- Opršal explained the very recently introduced algebraic approach to the computational complexity of PCSPs.
- Barto presented his results on PCSPs and cyclic operations.

Many approximation algorithms for CSPs are based on convex relaxations.

- Berkholz gave an overview on relaxations for Boolean CSPs based on algebraic methods.
- Schramm explained the power of semidefinite programming relaxations for random CSPs.
- Tulsiani presented results on the limits of linear programming relaxations for CSPs.
- Makarychev showed how to obtain an integrality gap for the Călinescu-Karloff-Rabani linear programming relaxation of the Multiway-Cut problem.
- Austrin established the currently best known inapproximability result for Min UnCut, which is a special Boolean CSP.

Some of the most exciting developments in approximability in the last decade revolve around the *unique games conjecture*, or UGC, of Khot (2002). This bold conjecture asserts that, for CSPs with a certain constraint language over a large enough domain, it is NP-hard to distinguish almost satisfiable instances from those where only a small fraction of constraints can be satisfied. This conjecture is known to imply tight inapproximability results for many classical optimisation problems. Moreover, if the UGC is true, then, as shown by Raghavendra in 2008, a simple algorithm based on semidefinite programming provides the best possible approximation for *all* CSPs (though the exact quality of this approximation is unknown).

- Moshkovitz presented recent developments on the so-called 2-to-2 PCP theorem, which covers important special cases of the UGC.

**Logic and the complexity of CSPs.** Logic has been used in two distinct ways in the study of the CSP. One of them, starting from earlier work of Kolaitis and Vardi, is *descriptive complexity*, where one tries to classify CSPs as classes of instances with respect to definability in a given logic. The other way is to use logic to specify CSP instances, which can be done very naturally. The latter direction leads to generalisations such as the quantified CSP (QCSP), as well as to the study of CSPs over infinite domains, where important links with the algebraic approach were found.

- Roy presented a dichotomy theorem for the inverse satisfiability problem.
- Bodirsky gave a talk on two methods of reducing infinite-domain CSPs to finite-domain CSPs.
- Pinsker explained recent results on the algebraic approach to infinite-domain CSPs. These results are related to the so-called loop conditions, which were in more detail discussed by Kozik.
- Kompatscher presented a proof of the equivalence of two dichotomy conjectures for infinite-domain CSPs.
- Mottet gave a new proof of the dichotomy for MMSNP and discussed consequences for infinite-domain CSPs.
- Martin described recent results for temporal and spatial problems, which are special cases of infinite-domain CSPs.

**Exact exponential complexity of CSPs.** The area of parameterised complexity is closely related to the area of exact exponential complexity, in which the goal is to design the most efficient exponential-time algorithms. There has been significant progress on the exact exponential complexity of CSPs.

- Golovnev presented results that give optimal lower bounds on the running time of algorithms for deciding if there is a homomorphism from one graph to another.
- The complexity of counting solutions for CSPs and related problems from statistical physics were presented by Goldberg and Jerrum.

## 2 Table of Contents

### Executive Summary

*Martin Grohe, Venkatesan Guruswami, Dániel Marx, and Stanislav Živný . . . . .* 1

### Overview of Talks

Improved NP-hardness of approximating Max Cut <i>Per Austrin . . . . .</i>	7
Cyclic operations in promise constraint satisfaction problems <i>Libor Barto . . . . .</i>	7
On algebraic and semi-algebraic relaxations for Boolean CSPs <i>Christoph Berkholz . . . . .</i>	7
Reducing Infinite-Domain CSPs to Finite-Domain CSPs <i>Manuel Bodirsky . . . . .</i>	8
Promise Constraint Satisfaction <i>Joshua Brakensiek . . . . .</i>	8
CSP Dichotomy <i>Andrei A. Bulatov . . . . .</i>	9
Boolean approximate counting CSPs with weak conservativity <i>Leslie Ann Goldberg . . . . .</i>	9
Tight Bounds for the Graph Homomorphism Problem <i>Alexander Golovnev . . . . .</i>	10
Approximately counting list H-colourings: a complexity classification <i>Mark R. Jerrum . . . . .</i>	10
Reflections for VCSP <i>Alexandr Kazda . . . . .</i>	11
Even Delta-Matroids and the Complexity of Planar Boolean CSPs <i>Vladimir Kolmogorov . . . . .</i>	11
The Equivalence of Two Dichotomy Conjectures for Infinite Domain CSPs <i>Michael Kompatscher . . . . .</i>	12
Finite pseudoloops <i>Marcin Kozik . . . . .</i>	12
Integrality Gap For Minimum Multiway Cut <i>Yury Makarychev . . . . .</i>	12
Classification transfer for qualitative reasoning problems <i>Barnaby Martin . . . . .</i>	13
Small Set Expansion in The Johnson Graph <i>Dana Moshkovitz . . . . .</i>	13
A universal-algebraic proof of the dichotomy for MMSNP <i>Antoine Mottet . . . . .</i>	14
Algebraic view on PCSP and hardness of coloring a $d$ -colorable graph with $2d - 1$ colors <i>Jakub Opršal . . . . .</i>	14

Algebraic structure of polymorphism clones of infinite CSP templates	
<i>Michael Pinsker</i> . . . . .	14
A Dichotomy Theorem for the Inverse Satisfiability Problem	
<i>Biman Roy</i> . . . . .	15
Refuting Random CSPs: Survey + NAE-3SAT	
<i>Tselil Schramm</i> . . . . .	15
From Weak to Strong LP Gaps for all CSPs	
<i>Madhur Tulsiani</i> . . . . .	16
Submodular cost functions and valued constraint satisfaction problems over infinite domains	
<i>Caterina Viola</i> . . . . .	16
An algorithm for Constraint Satisfaction Problem on finite set	
<i>Dmitriy Zhuk</i> . . . . .	17
<b>Participants</b> . . . . .	18

### 3 Overview of Talks

#### 3.1 Improved NP-hardness of approximating Max Cut

*Per Austrin (KTH Royal Institute of Technology – Stockholm, SE)*

**License** © Creative Commons BY 3.0 Unported license  
© Per Austrin

**Joint work of** Per Austrin, Mårten Wiman

We show that the Min Uncut problem is NP-hard to approximate within 1.4568. The previous best bound was  $11/8 - \epsilon$  [1].

##### References

- 1 J. Håstad, S. Huang, R. Manokaran, R. O'Donnell, and J. Wright. Improved NP-Inapproximability for 2-Variable Linear Equations. *Theory of Computing*, 13(1):1–51, 2017.

#### 3.2 Cyclic operations in promise constraint satisfaction problems

*Libor Barto (Charles University – Prague, CZ)*

**License** © Creative Commons BY 3.0 Unported license  
© Libor Barto

The promise constraint satisfaction problem (PCSP) is a generalization of the constraint satisfaction problem (CSP), where the aim is to distinguish between instances satisfiable over a fixed constraint language and instances that are not satisfiable over another fixed, weaker, constraint language. I will talk about two modest applications of cyclic operations (which proved useful in CSP) in PCSP. The first one is a negative result saying that the tractability of a PCSP cannot be always explained by a tractable finite-domain CSP, in a certain precise sense. The second one shows that monotone Boolean PCPs with enough cyclic polymorphisms are tractable.

#### 3.3 On algebraic and semi-algebraic relaxations for Boolean CSPs

*Christoph Berkholz (HU Berlin, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Christoph Berkholz

**Main reference** Christoph Berkholz: “The Relation between Polynomial Calculus, Sherali-Adams, and Sum-of-Squares Proofs”, in Proc. of the 35th Symposium on Theoretical Aspects of Computer Science, STACS 2018, February 28 to March 3, 2018, Caen, France, LIPIcs, Vol. 96, pp. 11:1–11:14, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2018.

**URL** <http://dx.doi.org/10.4230/LIPIcs.STACS.2018.11>

In this talk I will give an overview on relaxations for Boolean CSPs based on algebraic methods (such as Hilbert’s Nullstellensatz and the Gröbner basis Algorithm) as well as relaxations based on linear and semi-definite programming (such as the Sherali-Adams and the sum-of-squares hierarchy). A particular focus will be on comparing these methods with respect to their complexity, including recent findings from the main reference.

### 3.4 Reducing Infinite-Domain CSPs to Finite-Domain CSPs

*Manuel Bodirsky (TU Dresden, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Manuel Bodirsky

In this talk I present two fundamentally different techniques to reduce infinite-domain CSPs to finite-domain CSPs, and give many examples of CSPs that can be solved in polynomial time using these methods.

### 3.5 Promise Constraint Satisfaction

*Joshua Brakensiek (Carnegie Mellon University – Pittsburgh, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Joshua Brakensiek

**Joint work of** Joshua Brakensiek, Venkatesan Guruswami

**Main reference** Joshua Brakensiek, Venkatesan Guruswami: “Promise Constraint Satisfaction: Algebraic Structure and a Symmetric Boolean Dichotomy”, CoRR, Vol. abs/1704.01937, 2017.

**URL** <http://arxiv.org/abs/1704.01937>

With the recent resolution of the dichotomy conjecture, much focus is now on understanding generalizations of CSPs. We propose a generalization of CSPs called Promise CSPs. In a Promise CSP, constraints come in pairs: “stricter” and “weaker” versions (possibly over different domains). The main computational question is: given that there is an assignment satisfying the strict constraints, find an assignment to the weaker versions. Besides capturing ordinary CSPs, Promise CSPs also capture other problems in the literature, such as approximate graph coloring and  $(2 + \epsilon)$ -SAT [1].

Promise CSPs can still be studied with polymorphic techniques, but they differ substantially from their CSP cousins. First, due to the gap between “strict” and “weak” versions, polymorphisms are no longer closed under composition (in particular they do not form a clone). Thus, tractability can no longer be explained by a single polymorphism but requires understanding an infinite sequence of polymorphisms simultaneously. In particular, it seems like topological methods rather than algebraic methods may be of greater use in understanding Promise CSPs.

Besides this general overview of Promise CSPs, some new results are discussed. On the algorithmic side, families of polymorphisms known as “threshold-periodic polymorphisms” are discussed whose corresponding co-clones seem to necessarily require both bounded-width and linear equation methods in order to solve. On the hardness side, the standard paradigm in hardness of approximation (long code testing + a suitable label cover variant) is sufficient to give a complete classification of Boolean Promise CSPs with symmetric, folded predicates.

#### References

- 1 P. Austrin, V. Guruswami, and J. Håstad.  $(2 + \epsilon)$ -Sat Is NP-hard. *SIAM Journal on Computing*, 46(5):1554–1573, 2017.



### 3.6 CSP Dichotomy

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

**License** © Creative Commons BY 3.0 Unported license  
© Andrei A. Bulatov

**Main reference** Andrei A. Bulatov: “A Dichotomy Theorem for Nonuniform CSPs”, in Proc. of the 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15–17, 2017, pp. 319–330, IEEE Computer Society, 2017.

**URL** <http://dx.doi.org/10.1109/FOCS.2017.37>

**Main reference** Andrei A. Bulatov: “A dichotomy theorem for nonuniform CSPs”, CoRR, Vol. abs/1703.03021, 2017.

**URL** <http://arxiv.org/abs/1703.03021>

In 1993 Feder and Vardi posed a conjecture that suggests that every Constraint Satisfaction problem (CSP) parametrized by its target structure is either solvable in polynomial time or is NP-complete [2]. Later this conjecture was made more precise by delineating the exact condition that separates the polynomial case from the NP-complete one [1]. The hardness part of the conjecture has been known for long time. In this work we present a polynomial time algorithm that solves all CSPs satisfying the condition, thus, confirming Feder-Vardi conjecture.

#### References

- 1 A. A. Bulatov, P. Jeavons, and A. A. Krokhin. Classifying the complexity of constraints using finite algebras. *SIAM Journal on Computing*, 34(3):720–742, 2005.
- 2 T. Feder and M. Y. Vardi. Monotone monadic SNP and constraint satisfaction. *STOC*, pages 612–622, 1993.

### 3.7 Boolean approximate counting CSPs with weak conservativity

Leslie Ann Goldberg (*University of Oxford, GB*)

**License** © Creative Commons BY 3.0 Unported license  
© Leslie Ann Goldberg

**Joint work of** Miriam Backens, Andrei A. Bulatov, Leslie Ann Goldberg, Colin McQuillan, Stanislav Živný

**Main reference** Miriam Backens, Andrei Bulatov, Leslie Ann Goldberg, Colin McQuillan, Stanislav Živný: “Boolean approximate counting CSPs with weak conservativity, and implications for ferromagnetic two-spin”, CoRR, Vol. abs/1804.04993, 2018.

**URL** <http://arxiv.org/abs/1804.04993>

We analyse the complexity of approximate counting constraint satisfactions problems  $\#CSP(F)$ , where  $F$  is a set of nonnegative rational-valued functions of Boolean variables. A complete classification is known in the conservative case, where  $F$  is assumed to contain arbitrary unary functions. We strengthen this result by fixing any permissive strictly increasing unary function and any permissive strictly decreasing unary function, and adding only those to  $F$ : this is weak conservativity. The resulting classification is employed to characterise the complexity of a wide range of two-spin problems, fully classifying the ferromagnetic case.

### 3.8 Tight Bounds for the Graph Homomorphism Problem

*Alexander Golovnev (Yahoo! Research – New York, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Alexander Golovnev

**Joint work of** Marek Cygan, Fedor V. Fomin, Alexander Golovnev, Alexander S. Kulikov, Ivan Mihajlin, Jakub Pachocki, Arkadiusz Socala

**Main reference** Marek Cygan, Fedor V. Fomin, Alexander Golovnev, Alexander S. Kulikov, Ivan Mihajlin, Jakub Pachocki, Arkadiusz Socala: “Tight Lower Bounds on Graph Embedding Problems”, J. ACM, Vol. 64(3), pp. 18:1–18:22, 2017.

**URL** <http://dx.doi.org/10.1145/3051094>

We prove that unless the Exponential Time Hypothesis (ETH) fails, deciding if there is a homomorphism from graph  $G$  to graph  $H$  cannot be done in time  $|V(H)|^{o(|V(G)|)}$ . We also show an exponential-time reduction from Graph Homomorphism to Subgraph Isomorphism. This rules out (subject to ETH) a possibility of  $|V(H)|^{o(|V(H)|)}$ -time algorithm deciding if graph  $G$  is a subgraph of  $H$ . For both problems our lower bounds asymptotically match the running time of brute-force algorithms trying all possible mappings of one graph into another. Thus, our work closes the gap in the known complexity of these problems.

Moreover, as a consequence of our reductions, conditional lower bounds follow for other related problems such as Locally Injective Homomorphism, Graph Minors, Topological Graph Minors, Minimum Distortion Embedding and Quadratic Assignment Problem.

### 3.9 Approximately counting list $H$ -colourings: a complexity classification

*Mark R. Jerrum (Queen Mary University of London, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Mark R. Jerrum

**Joint work of** Andreas Galanis, Leslie Ann Goldberg, Mark R. Jerrum

**Main reference** Andreas Galanis, Leslie Ann Goldberg, Mark Jerrum: “A Complexity Trichotomy for Approximately Counting List  $H$ -Colorings”, TOCT, Vol. 9(2), pp. 9:1–9:22, 2017.

**URL** <http://dx.doi.org/10.1145/3037381>

Suppose  $H$  is a fixed graph.  $H$ -colourings of a graph  $G$  (a.k.a. homomorphisms from  $G$  to  $H$ ) generalise familiar proper vertex colourings of  $G$ . More than 15 years ago, Dyer and Greenhill considered the computational complexity of counting  $H$ -colourings, and demonstrated a dichotomy, in terms of the graph  $H$ , between polynomial time and  $\#P$ -complete.

That result was for exact counting, and, even now, there is only a partial complexity classification for approximate counting. However, the classification problem becomes tractable if we look instead at *list*  $H$ -colourings. In this talk, I’ll present a classification (in fact a trichotomy) for the problem of approximately counting list  $H$ -colourings of a graph. It turns out that some interesting hereditary graph classes come into play in describing and proving the trichotomy result.

### 3.10 Reflections for VCSP

Alexandr Kazda (*Charles University – Prague, CZ*)

**License** © Creative Commons BY 3.0 Unported license  
© Alexandr Kazda

Take a weighted constraint language  $\Gamma$  on a finite domain. The complexity of  $\text{VCSP}(\Gamma)$  depends only on the clone of weighted polymorphisms of  $\Gamma$ . Kolmogorov, Rolínek and Krokhin gave a complete classification of this complexity in 2015 (for the journal version, see [1]); however there is still room for improvement of our understanding of weighted clones.

In our talk, we gave an overview of VCSP and conjectured that the theory of weighted clones could be made more elegant if one could make the idea of a reflection from [2] work for weighted clones and VCSPs.

#### References

- 1 V. Kolmogorov, A. Krokhin, and M. Rolínek. The Complexity of General-Valued CSPs. *SIAM Journal on Computing*, 46(3):1087–1110, 2017.
- 2 L. Barto, J. Opršal, and M. Pinsker. The wonderland of reflections. *Israel Journal of Mathematics*, 223(1):363–398, 2018.

### 3.11 Even Delta-Matroids and the Complexity of Planar Boolean CSPs

Vladimir Kolmogorov (*IST Austria – Klosterneuburg, AT*)

**License** © Creative Commons BY 3.0 Unported license  
© Vladimir Kolmogorov

**Joint work of** Alexandr Kazda, Vladimir Kolmogorov, Michal Rolínek

**Main reference** Alexandr Kazda, Vladimir Kolmogorov, Michal Rolínek: “Even Delta-Matroids and the Complexity of Planar Boolean CSPs”, CoRR, Vol. abs/1602.03124, 2016.

**URL** <http://arxiv.org/abs/1602.03124>

We study Boolean CSPs where each variable appears in exactly two constraints and all constraints come from some language  $\Gamma$  (we call it “EdgeCSP( $\Gamma$ )”). If  $\Gamma$  contains the two constant unary relations then the only interesting case is when all constraints in  $\Gamma$  are “ $\Delta$ -matroids” (otherwise EdgeCSP( $\Gamma$ ) and CSP( $\Gamma$ ) have the same complexities, as shown by Feder). I will present a polynomial-time algorithm for “Even  $\Delta$ -matroids”, as well as an extension to some non-even  $\Delta$ -matroids. One consequence of this result is settling the complexity classification of planar Boolean CSPs started by Dvořák and Kupec.

### 3.12 The Equivalence of Two Dichotomy Conjectures for Infinite Domain CSPs

*Michael Kompatscher (Charles University – Prague, CZ)*

**License** © Creative Commons BY 3.0 Unported license

© Michael Kompatscher

**Joint work of** Libor Barto, Michael Kompatscher, Miroslav Olsák, Trung Van Pham, Michael Pinsker

**Main reference** Libor Barto, Michael Kompatscher, Miroslav Olsák, Michael Pinsker, Van Trung Pham:

“Equations in oligomorphic clones and the Constraint Satisfaction Problem for  $\omega$ -categorical structures”, CoRR, Vol. abs/1612.07551, 2016.

**URL** <http://arxiv.org/abs/1612.07551>

CSPs of reducts of finitely bounded homogeneous structures form a natural generalization of finite CSPs and also allow us to also apply the universal algebraic approach. As for finite structures, non-triviality of the equational structure of the corresponding polymorphism clone is conjectured to be a/the source of tractability. In my talk I would like to discuss to which extend it is enough to consider only non-trivial equations of height 1.

### 3.13 Finite pseudoloops

*Marcin Kozik (Jagiellonian University – Kraków, PL)*

**License** © Creative Commons BY 3.0 Unported license

© Marcin Kozik

The loop lemma states: if a finite digraph has no sources and no sinks, contains a closed walk with one more forward than backward edge and has a Taylor polymorphism then it contains a loop. This result is closely connected to algebraic investigations of CSPs with finite templates.

When investigating a structure of CSPs with omega-categorical templates the question becomes: take a graph with no sources and no sinks, and an oligomorphic subgroup of its automorphism group: if the quotient of the graph modulo the orbit equivalence contains a walk as in the loop lemma and the graph does not pp-interpret 3-clique with parameters does the quotient necessarily contain a loop? We prove this statement true in the finite case.

### 3.14 Integrality Gap For Minimum Multiway Cut

*Yury Makarychev (TTIC – Chicago, US)*

**License** © Creative Commons BY 3.0 Unported license

© Yury Makarychev

**Joint work of** Haris Angelidakis, Yury Makarychev, Pasin Manurangsi

**Main reference** Haris Angelidakis, Yury Makarychev, Pasin Manurangsi: “An Improved Integrality Gap for the Călinescu-Karloff-Rabani Relaxation for Multiway Cut”, in Proc. of the Integer Programming and Combinatorial Optimization - 19th International Conference, IPCO 2017, Waterloo, ON, Canada, June 26-28, 2017, Proceedings, Lecture Notes in Computer Science, Vol. 10328, pp. 39–50, Springer, 2017.

**URL** [http://dx.doi.org/10.1007/978-3-319-59250-3\\_4](http://dx.doi.org/10.1007/978-3-319-59250-3_4)

We consider the Călinescu–Karloff–Rabani linear programming relaxation for Minimum Multiway Cut. We prove that its integrality gap is at least  $\alpha_k = 6/(5 + 1/(k - 1))$ , where  $k$  is the number of terminals. Our result improves the previous best known gap lower bound of  $8/(7 + 1/(k - 1))$ , which was established by Freund and Karloff in 2000. Our result implies that it is NP-hard to get an  $\alpha_k - \epsilon$  approximation for the problem (for every  $\epsilon > 0$ ), if the Unique Games Conjecture is true.

### 3.15 Classification transfer for qualitative reasoning problems

Barnaby Martin (*Durham University, GB*)

**License** © Creative Commons BY 3.0 Unported license  
© Barnaby Martin

**Joint work of** Manuel Bodirsky, Peter Jonsson, Barnaby Martin, Antoine Mottet

**Main reference** Barnaby Martin, Peter Jonsson, Manuel Bodirsky, Antoine Mottet: “Classification transfer for qualitative reasoning problems”, CoRR, Vol. abs/1805.02038, 2018.

**URL** <http://arxiv.org/abs/1805.02038>

We study formalisms for temporal and spatial reasoning in the modern, algebraic and model-theoretic, context of infinite-domain Constraint Satisfaction Problems (CSPs). We show how questions on the complexity of their subclasses can be solved using existing results via the powerful use of primitive positive (pp) interpretations and pp-homotopy.

We demonstrate the methodology by giving a full complexity classification of all constraint languages that are first-order definable in Allen’s Interval Algebra and contain the basic relations (s) and (f). In the case of the Rectangle Algebra we answer in the affirmative the old open question as to whether ORD-Horn is a maximally tractable subset among the (disjunctive, binary) relations. We then generalise our results for the Rectangle Algebra to the  $r$ -dimensional Block Algebra.

### 3.16 Small Set Expansion in The Johnson Graph

Dana Moshkovitz (*University of Texas – Austin, US*)

**License** © Creative Commons BY 3.0 Unported license  
© Dana Moshkovitz

**Joint work of** Subhash Khot, Dor Minzer, Dana Moshkovitz, Shmuel Safra

**Main reference** S. Khot, D. Minzer, D. Moshkovitz, and M. Safra: “Small Set Expansion in The Johnson Graph”. *Electronic Colloquium on Computational Complexity (ECCC)*, 25:78, 2018.

**URL** <https://eccc.weizmann.ac.il/report/2018/078>

We study expansion properties of the (generalized) Johnson Graph. For natural numbers  $t < l < k$ , the nodes of the graph are sets of size  $l$  in a universe of size  $k$ . Two sets are connected if their intersection is of size  $t$ . The Johnson graph arises often in combinatorics and theoretical computer science: it represents a “slice” of the noisy hypercube, and it is the graph that underlies direct product tests as well as a candidate hard unique game.

We prove that any small set of vertices in the graph either has near perfect edge expansion or is not pseudorandom. Here “not pseudorandom” means that the set becomes denser when conditioning on containing a small set of elements. In other words, we show that slices of the noisy hypercube – while not small set expanders like the noisy hypercube – only have non-expanding small sets of a certain simple structure.

This paper is related to a recent line of work establishing the 2-to-2 Theorem in PCP. The result was motivated, in part, by [1] which hypothesized and made partial progress on similar result for the Grassmann graph. In turn, our result for the Johnson graphs served as a crucial step towards the full result for the Grassmann graphs completed subsequently in [2].

#### References

- 1 I. Dinur, S. Khot, G. Kindler, D. Minzer, and M. Safra. On Non-Optimally Expanding Sets in Grassmann Graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:94, 2017.
- 2 S. Khot, D. Minzer, and M. Safra. On Independent Sets, 2-to-2 Games and Grassmann Graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:124, 2016.

### 3.17 A universal-algebraic proof of the dichotomy for MMSNP

*Antoine Mottet (TU Dresden, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Antoine Mottet

**Joint work of** Manuel Bodirsky, Florent Madelaine, Antoine Mottet  
**Main reference** Manuel Bodirsky, Florent R. Madelaine, Antoine Mottet: “A universal-algebraic proof of the complexity dichotomy for Monotone Monadic SNP”, in Proc. of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018, Oxford, UK, July 09-12, 2018, pp. 105–114, ACM, 2018.  
**URL** <http://dx.doi.org/10.1145/3209108.3209156>

The logic MMSNP is a restricted fragment of existential second-order logic that was discovered by Feder and Vardi, who showed that every MMSNP sentence is computationally equivalent to a finite-domain CSP; the involved probabilistic reductions were derandomized by Kun using explicit constructions of expander structures. I will present a new proof based on the universal-algebraic method and recent Ramsey-theoretic results by Hubička and Nešetřil. This new proof allows us to verify the infinite-domain dichotomy conjecture by Bodirsky and Pinsker for infinite-domain CSPs in MMSNP.

### 3.18 Algebraic view on PCSP and hardness of coloring a $d$ -colorable graph with $2d - 1$ colors

*Jakub Opršal (TU Dresden, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Jakub Opršal

**Joint work of** Jakub Bulín, Andrei Krokhin, Jakub Opršal

The talk focuses on fixed template promise constraint satisfaction problem (PCSP). A template of PCSP is a pair of finite relational structures **A** and **B** with a homomorphism between them. The goal is to decide, given third structure **C**, whether **C** maps homomorphically to **A**, or it does not even map to **B**. Recently, polymorphisms have been successfully used by Austrin, Håstad, Guruswami and Brakiensiek to provide several results on the complexity of PCSPs.

We show that the complexity depends only on minor (height 1) identities satisfied by the polymorphisms. Further, we use this result to give a reduction from promise 3-uniform hypergraph coloring to promise graph coloring providing that deciding whether a graph is  $d$ -colorable or not even  $(2d - 1)$ -colorable is NP-hard for any  $d > 2$ .

### 3.19 Algebraic structure of polymorphism clones of infinite CSP templates

*Michael Pinsker (TU Wien, AT)*

**License**  Creative Commons BY 3.0 Unported license  
© Michael Pinsker

**Joint work of** Pierre Gillibert, Julius Jonušas, Michael Pinsker

We survey recent results on the algebraic structure of polymorphism clones of infinite CSP templates. Satisfaction of identities in such clones is reflected by the satisfaction of so-called loop conditions, which are in some sense fixed point properties of actions of the clone on graphs. We compare the relative strength of such conditions.

### 3.20 A Dichotomy Theorem for the Inverse Satisfiability Problem

Biman Roy (Linköping University, SE)

**License** © Creative Commons BY 3.0 Unported license  
© Biman Roy

**Joint work of** Victor Lagerkvist, Biman Roy

**Main reference** Victor Lagerkvist, Biman Roy: “A Dichotomy Theorem for the Inverse Satisfiability Problem”, in Proc. of the 37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2017, December 11-15, 2017, Kanpur, India, LIPIcs, Vol. 93, pp. 39:39–39:14, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2017.

**URL** <http://dx.doi.org/10.4230/LIPIcs.FSTTCS.2017.39>

The inverse satisfiability problem over a set of Boolean relations  $\Gamma$  ( $\text{Inv-SAT}(\Gamma)$ ) is the computational decision problem of, given a relation  $R$ , deciding whether there exists a  $\text{SAT}(\Gamma)$  instance with  $R$  as its set of models. This problem is co-NP-complete in general and a dichotomy theorem for finite  $\Gamma$  containing the constant Boolean relations was obtained by Kavvadias and Sideri. In this paper we remove the latter condition and prove that  $\text{Inv-SAT}(\Gamma)$  is always either tractable or co-NP-complete for all finite sets of relations  $\Gamma$ , thus solving a problem open since 1998. Very few of the techniques used by Kavvadias and Sideri are applicable and we have to turn to more recently developed algebraic approaches based on partial polymorphisms. We also consider the case when  $\Gamma$  is infinite, where the situation differs markedly from the case of SAT. More precisely, we show that there exists infinite  $\Gamma$  such that  $\text{Inv-SAT}(\Gamma)$  is tractable even though there exists finite  $\Delta \subset \Gamma$  such that  $\text{Inv-SAT}(\Delta)$  is co-NP-complete.

#### References

- 1 P. Jonsson, V. Lagerkvist, and G. Nordh. Constructing NP-intermediate problems by blowing holes with parameters of various properties. *Theoretical Computer Science*, 581:67–82, 2015.
- 2 P. Jonsson, V. Lagerkvist, G. Nordh, and B. Zanuttini. Strong partial clones and the time complexity of SAT problems. *Journal of Computer and System Sciences*, 84:52–78, 2017.
- 3 D. Kavvadias and M. Sideri. The inverse satisfiability problem. *SIAM Journal on Computing*, 28:152–163, 1998.

### 3.21 Refuting Random CSPs: Survey + NAE-3SAT

Tselil Schramm (University of California – Berkeley, US)

**License** © Creative Commons BY 3.0 Unported license  
© Tselil Schramm

**Joint work of** Yash Deshpande, Andrea Montanari, Ryan O’Donnell, Tselil Schramm, Subhabrata Sen

**Main reference** Yash Deshpande, Andrea Montanari, Ryan O’Donnell, Tselil Schramm, Subhabrata Sen: “The threshold for SDP-refutation of random regular NAE-3SAT”, CoRR, Vol. abs/1804.05230, 2018.

**URL** <http://arxiv.org/abs/1804.05230>

In this talk I’ll give a survey of recent advances in our understanding of the power of Semidefinite Programs for refuting random CSPs, then describe a recent result in which we establish the exact basic SDP threshold for refuting random regular NAE-3SAT.

### 3.22 From Weak to Strong LP Gaps for all CSPs

Madhur Tulsiani (TTIC – Chicago, US)

**License** © Creative Commons BY 3.0 Unported license  
© Madhur Tulsiani

**Joint work of** Mrinalkanti Ghosh, Madhur Tulsiani

**Main reference** M. Ghosh and M. Tulsiani: “From Weak to Strong Linear Programming Gaps for All Constraint Satisfaction Problems” *Computational Complexity Conference 2017*, Vol. 14, pages 1–33, Theory of Computing.

**URL** <http://dx.doi.org/10.4086/toc.2018.v014a010>

We study the approximability of constraint satisfaction problems (CSPs) by linear programming (LP) relaxations. We show that for every CSP, the approximation obtained by a basic LP relaxation, is no weaker than the approximation obtained using relaxations given by  $\Omega(\log n / (\log \log n))$  levels of the of the Sherali-Adams hierarchy on instances of size  $n$ .

It was proved by Chan et al. [1] that any polynomial size LP extended formulation is no stronger than relaxations obtained by a super-constant levels of the Sherali-Adams hierarchy. Combining this with our result also implies that any polynomial size LP extended formulation is no stronger than the basic LP, which can be thought of as the base level of the Sherali-Adams hierarchy. This essentially gives a dichotomy result for approximation of CSPs by polynomial size LP extended formulations.

Using our techniques, we also simplify and strengthen the result by Khot et al. [2] on (strong) approximation resistance for LPs. They provided a necessary and sufficient condition under which  $\Omega(\log \log n)$  levels of the Sherali-Adams hierarchy cannot achieve an approximation better than a random assignment. We simplify their proof and strengthen the bound to  $\Omega(\log n / (\log \log n))$  levels.

#### References

- 1 S. O. Chan, J. R. Lee, P. Raghavendra, and D. Steurer. Approximate Constraint Satisfaction Requires Large LP Relaxations. *FOCS 2013*, pages 350–359.
- 2 S. Khot, M. Tulsiani, and P. Worah. A characterization of strong approximation resistance. *STOC 2014*, pages 634–643.

### 3.23 Submodular cost functions and valued constraint satisfaction problems over infinite domains

Caterina Viola (TU Dresden, DE)

**License** © Creative Commons BY 3.0 Unported license  
© Caterina Viola

**Joint work of** Manuel Bodirsky, Marcello Mamino, Caterina Viola

**Main reference** Manuel Bodirsky, Marcello Mamino, Caterina Viola: “Submodular Functions and Valued Constraint Satisfaction Problems over Infinite Domains”, *CoRR*, Vol. abs/1804.01710, 2018.

**URL** <http://arxiv.org/abs/1804.01710>

Valued constraint satisfaction problems (VCSPs) are a generalisation of constraint satisfaction problems (CSPs) that capture combinatorial optimisation problems. Recently, the computational complexity of all VCSPs for finite sets of cost functions over finite domains has been classified completely. Many natural optimisation problems, however, cannot be formulated as VCSPs over a finite domain (e.g., the linear programming problem), but can be modelled as a VCSP over the domain of rational numbers,  $\mathbb{Q}$ .

I will focus a special class of VCSPs over  $\mathbb{Q}$ , namely the class of VCSPs for finite piecewise linear homogeneous (PLH) valued languages. A PLH valued language is a set of cost



functions that admit a first-order definition over  $\mathbb{Q}$ , using the order relation, the element 1, and the scalar multiplication by rationals. For these languages the VCSP is solvable in polynomial time when the cost functions are additionally submodular. Moreover, the submodularity is a condition of maximal tractability for PLH valued languages.

### 3.24 An algorithm for Constraint Satisfaction Problem on finite set

*Dmitriy Zhuk (Moscow State University, RU)*

**License** © Creative Commons BY 3.0 Unported license  
© Dmitriy Zhuk

**Main reference** Dmitriy Zhuk: “A Proof of CSP Dichotomy Conjecture”, in Proc. of the 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017, pp. 331–342, IEEE Computer Society, 2017.

**URL** <http://dx.doi.org/10.1109/FOCS.2017.38>

**Main reference** Dmitriy Zhuk: “The Proof of CSP Dichotomy Conjecture”, CoRR, Vol. abs/1704.01914, 2017.

**URL** <http://arxiv.org/abs/1704.01914>

In 2017 two proofs of CSP Dichotomy Conjecture and correspondingly two polynomial algorithms for constraint languages admitting a weak near-unanimity polymorphism were developed. In the talk one of the algorithms will be discussed in detail. I will try to minimize the amount of algebraic notions and focus on the new methods and ideas.

## Participants

- Isolde Adler  
University of Leeds, GB
- Per Austrin  
KTH Royal Institute of  
Technology – Stockholm, SE
- Libor Barto  
Charles University – Prague, CZ
- Christoph Berkholz  
HU Berlin, DE
- Manuel Bodirsky  
TU Dresden, DE
- Joshua Brakensiek  
Carnegie Mellon University –  
Pittsburgh, US
- Andrei A. Bulatov  
Simon Fraser University –  
Burnaby, CA
- Clement Carbonnel  
University of Oxford, GB
- Hubie Chen  
Universidad del País Vasco –  
Donostia, ES
- Victor Dalmau  
UPF – Barcelona, ES
- Peter Fulla  
University of Oxford, GB
- Leslie Ann Goldberg  
University of Oxford, GB
- Alexander Golovnev  
Yahoo! Research – New York, US
- Martin Grohe  
RWTH Aachen, DE
- Venkatesan Guruswami  
Carnegie Mellon University –  
Pittsburgh, US
- Johan Hastad  
KTH Royal Institute of  
Technology – Stockholm, SE
- Mark R. Jerrum  
Queen Mary University of  
London, GB
- Peter Jonsson  
Linköping University, SE
- Alexandr Kazda  
Charles University – Prague, CZ
- Eun Jung Kim  
University Paris-Dauphine, FR
- Vladimir Kolmogorov  
IST Austria –  
Klosterneuburg, AT
- Michael Kompatscher  
Charles University – Prague, CZ
- Marcin Kozik  
Jagiellonian University –  
Kraków, PL
- Andrei Krokhn  
Durham University, GB
- Victor Lagerqvist  
TU Dresden, DE
- Euiwoong Lee  
New York University, US
- Yury Makarychev  
TTIC – Chicago, US
- Barnaby Martin  
Durham University, GB
- Dana Moshkovitz  
University of Texas – Austin, US
- Antoine Mottet  
TU Dresden, DE
- Ryan O'Donnell  
Carnegie Mellon University –  
Pittsburgh, US
- Miroslav Olsak  
Charles University – Prague, CZ
- Jakub Opršal  
TU Dresden, DE
- Marcin Pilipczuk  
University of Warsaw, PL
- Michael Pinski  
TU Wien, AT
- Akbar Rafiey  
Simon Fraser University –  
Burnaby, CA
- Miguel Romero  
University of Oxford, GB
- Biman Roy  
Linköping University, SE
- Tselil Schramm  
University of California –  
Berkeley, US
- Stefan Szeider  
TU Wien, AT
- Johan Thapper  
University Paris-Est –  
Marne-la-Vallée, FR
- Madhur Tulsiani  
TTIC – Chicago, US
- Caterina Viola  
TU Dresden, DE
- Ross Willard  
University of Waterloo, CA
- Yuichi Yoshida  
National Institute of Informatics –  
Tokyo, JP
- Dmitriy Zhuk  
Moscow State University, RU
- Stanislav Živný  
University of Oxford, GB



# High-Performance Graph Algorithms

Edited by

Henning Meyerhenke<sup>1</sup>, Richard Peng<sup>2</sup>, and Ilya Safro<sup>3</sup>

<sup>1</sup> HU Berlin, DE, [meyerhenke@hu-berlin.de](mailto:meyerhenke@hu-berlin.de)

<sup>2</sup> Georgia Institute of Technology – Atlanta, US, [rpeng@cc.gatech.edu](mailto:rpeng@cc.gatech.edu)

<sup>3</sup> Clemson University, US, [isafro@g.clemson.edu](mailto:isafro@g.clemson.edu)

---

## Abstract

This report documents the program and outcomes of Dagstuhl Seminar 18241 “High-performance Graph Algorithms”. The seminar reflected the ongoing qualitative change how graph algorithms are used in practice due to (i) the complex structure of graphs in new and emerging applications, (ii) the size of typical inputs, and (iii) the computer systems with which graph problems are solved. This change is having a tremendous impact on the field of graph algorithms in terms of algorithm theory and implementation as well as hardware requirements and application areas.

The seminar covered recent advances in all these aspects, trying to balance and mediate between theory and practice. The abstracts included in this report contain and survey recent state-of-the-art results, but also point to promising new directions for high-performance graph algorithms and their applications, both from a theoretical and a practical point of view.

**Seminar** June 10–15, 2018 – <http://www.dagstuhl.de/18241>

**2012 ACM Subject Classification** Mathematics of computing → Discrete mathematicsMathematics of computing → Graph theory, Theory of computation → Design and analysis of algorithms, Theory of computation → Models of computation, Theory of computation → Randomness, geometry and discrete structures

**Keywords and phrases** algorithm engineering, combinatorial scientific computing, graph algorithms, high-performance computing, theoretical computer science

**Digital Object Identifier** 10.4230/DagRep.8.6.19


**Edited in cooperation with** Manuel Penschuck

## 1 Executive Summary

*Henning Meyerhenke*

*Richard Peng*

*Ilya Safro*

**License**  Creative Commons BY 3.0 Unported license  
© Henning Meyerhenke, Richard Peng, and Ilya Safro

Many presentations in this Dagstuhl seminar emphasized recent trends regarding typical inputs and their effect on graph algorithm development. From a high-level perspective, one can divide the presentations into two categories: either more focused on algorithm theory or more focused on practical algorithmic results. Many talks considered both theoretical and practical aspects. Furthermore, attention was given to intermix talks with theoretical and practically motivated starting points in order to encourage discussions among attendees. We were happy to see such discussions, as well as synergy of both aspects, carrying over to working groups on open problems.



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

High-Performance Graph Algorithms, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 19–39

Editors: Henning Meyerhenke, Richard Peng, and Ilya Safro



Dagstuhl Reports

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

Theory-focused talks were given by Sachdeva, Nanongkai, Jacob, Mouatadid, Kyng, Tsourakakis, and Litvak. They considered numerous topics such as Laplacian solvers and related optimization techniques, dynamic graph algorithms, external-memory graph algorithms, graph decompositions, and generative models.

The talks with emphasis on practical performance can be further subdivided into three subclasses: (i) graph mining, network analysis and optimization, (ii) parallel, distributed and streaming graph algorithms and (iii) graph generation. The talks given by Koutra, Ahmed, Klymko, Angriman, Gleich and Schulz fall into the first subclass, with a wide variation of algorithmic problems under consideration. Likewise, it was interesting to see the variety in computing platforms and tools (for example shared memory, message passing, distributed systems, streaming from databases, GraphBLAS) used in the eight talks of subclass two, presented by Besta, Shun, Predari, Ramachandran, Pothén, Bader, Finocchi and Davis. Finally, the talks by Phillips, Sanders and Penschuck as well as Crescenzi dealt with generating very large graphs with properties also found in real-world graphs – which is important, among others, for convincing scaling studies in algorithm engineering.

## 2 Table of Contents

### Executive Summary

<i>Henning Meyerhenke, Richard Peng, and Ilya Safro</i> . . . . .	19
---	----

### Overview of Talks

Sampling from Massive Graph Streams: A Unifying Framework <i>Nesreen K. Ahmed</i> . . . . .	23
Centrality Computation in Real-World Networks: New Challenges <i>Eugenio Angriman</i> . . . . .	23
Massive-scale Graph Analytics: A New Algorithmic Model <i>David A. Bader</i> . . . . .	24
Core-periphery clustering and complex networks <i>Pierluigi Crescenzi</i> . . . . .	24
Scaling problems with metric constraints <i>David F. Gleich</i> . . . . .	25
SuiteSparse:GraphBLAS: graph algorithms in the language of linear algebra <i>Timothy Alden Davis</i> . . . . .	25
Beyond triangles <i>Irene Finocchi</i> . . . . .	25
The Parallel External Memory Model: Sparse Matrix Multiply and List Ranking <i>Riko Jacob</i> . . . . .	26
Theoretically Efficient Parallel Graph Algorithms Can Be Fast and Scalable <i>Julian Shun</i> . . . . .	26
An Ensemble Framework for Detecting Community Changes in Dynamic Networks <i>Christine Klymko</i> . . . . .	27
Scalable Inference and Summarization of Multi-source Network Data <i>Danaï Koutra</i> . . . . .	28
Optimization on Graphs <i>Rasmus Kyng</i> . . . . .	28
Power-law hypothesis for PageRank <i>Nelly Litvak</i> . . . . .	29
To Push or To Pull: On Reducing Communication and Synchronization in Graph Computations <i>Maciej Besta</i> . . . . .	29
A Toolbox to Extract Structure From Graphs <i>Lalla Mouatadid</i> . . . . .	30
Dynamic Algorithms and Complexity, A Short Survey <i>Danupon Nanongkai</i> . . . . .	30
Generating random massive graphs that mimic real data <i>Cynthia A. Phillips</i> . . . . .	30
When Exact Fails to be Parallel: Parallel Algorithms through Approximation <i>Alex Pothen</i> . . . . .	31

Topology-induced Enhancement of Mappings <i>Maria Predari</i> . . . . .	31
A Round-Efficient and Communication-Efficient Algorithm for Distributed Betweenness Centrality <i>Vijaya Ramachandran</i> . . . . .	32
Fast Approximate Gaussian Elimination for Laplacians <i>Sushant Sachdeva</i> . . . . .	33
Massively parallel communication-free graph generators <i>Peter Sanders and Manuel Penschuck</i> . . . . .	33
Practical Kernelization <i>Christian Schulz</i> . . . . .	34
Clustering with a faulty oracle <i>Charalampos E. Tsourakakis</i> . . . . .	34
<b>Working groups</b>	
Time segmentation working group <i>David A. Bader, Irene Finocchi, George Karypis, Michel A. Kinsy, Christine Klymko, Danaï Koutra, Cynthia A. Phillips, and Ilya Safro</i> . . . . .	35
Dynamic Graphs Working Group <i>Richard Peng, Nesreen K. Ahmed, Rob Bisseling, George Karypis, Danupon Nanongkai, Manuel Penschuck, Alex Pothén, Vijaya Ramachandran, and Julian Shun</i> .	36
k-GRIP: Combinatorial Approaches <i>Blair D. Sullivan</i> . . . . .	36
<b>Panel discussions</b>	
Applications of the theory of random graphs in algorithmic analysis of large networks <i>Nelly Litvak</i> . . . . .	37
<b>Participants</b> . . . . .	39

### 3 Overview of Talks

#### 3.1 Sampling from Massive Graph Streams: A Unifying Framework

*Nesreen K. Ahmed (Intel Labs – Santa Clara, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Nesreen K. Ahmed

**Joint work of** Nesreen K. Ahmed, Nick G. Duffield, Theodore L. Willke, Ryan A. Rossi

**Main reference** Nesreen K. Ahmed, Nick G. Duffield, Theodore L. Willke, Ryan A. Rossi: “On Sampling from Massive Graph Streams”, PVLDB, Vol. 10(11), pp. 1430–1441, 2017.

**URL** <http://dx.doi.org/10.14778/3137628.3137651>

The rapid growth of the Internet and the explosion in online social media has led to a data deluge. A growing set of online applications are continuously generating data at unprecedented rates, from the Internet of things (e.g., connected devices, routers), electronic communication (e.g., email, IMs), social media (e.g., blogs), to the vast collection of online social networks and content sharing applications (e.g., Facebook, Twitter). Graphs are a natural data representation in many of these application domains, where nodes represent individuals/entities and edges represent the interaction, communication, or connectivity among them. Consider interaction and activity networks formed from electronic communication between online users. These resulting interaction and activity networks manifest as a stream of edges, where edges (i.e., interactions) occur one at a time, carrying a wealth of behavioral, community, and relationship information. Many of these networks are massive in size, due to the prolific amount of activity data. To keep up with the growing pace of this data, we need efficient methods to analyze dynamic interaction networks as the data arrives in streams, rather than static snapshots of graphs. Sampling provides an attractive approach to quickly and efficiently find an approximate answer to a query, or more generally, any analysis objective. In this talk, I will discuss a novel adaptive general-purpose framework for sampling from graph streams, called graph priority sampling (GPS). From a high-volume stream of edges, our proposed framework maintains a generic sample of limited size that can be used at any time to accurately estimate the total weight of arbitrary graph subsets (i.e., triangles, cliques). To obtain accurate estimates of various graph properties, our proposed framework maintains a weight sensitive sample that devotes sampling resources to edges that are informative for those properties. Unbiasedness of subgraph estimators is established through a new Martingale formulation of graph stream sampling, in which subgraph estimators are unbiased, even when computed at different points in the stream. I will summarize the results of our large-scale experimental study on real-world graphs from various domains.

#### 3.2 Centrality Computation in Real-World Networks: New Challenges

*Eugenio Angriman (Universität Köln, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Eugenio Angriman

**Joint work of** Patrick Bisenius, Elisabetta Bergamini, Eugenio Angriman, Henning Meyerhenke

**Main reference** Patrick Bisenius, Elisabetta Bergamini, Eugenio Angriman, Henning Meyerhenke: “Computing Top- $k$  Closeness Centrality in Fully-dynamic Graphs”, in Proc. of the Twentieth Workshop on Algorithm Engineering and Experiments, ALENEX 2018, New Orleans, LA, USA, January 7-8, 2018., pp. 21–35, SIAM, 2018.

**URL** <http://dx.doi.org/10.1137/1.9781611975055.3>

The efficient computation of the top- $k$  most central nodes of a graph has been widely studied, especially for well-known centrality metrics such as Closeness, Katz, or Betweenness. We present several techniques that allow us to efficiently recompute the top- $k$  nodes with highest Closeness Centrality after the insertion or removal of an edge in a network without requiring

asymptotically more memory than the static algorithms. We also give a brief introduction to NetworKit, an open-source toolkit for large-scale network analysis with shared-memory systems. NetworKit includes a wide range of state-of-the-art graph algorithms implemented in C++, which can be easily used from a Python frontend. Finally, we introduce some open problems related to the extension of single-node centrality metrics to groups of nodes.

### 3.3 Massive-scale Graph Analytics: A New Algorithmic Model

*David A. Bader (Georgia Institute of Technology – Atlanta, US)*

**License** © Creative Commons BY 3.0 Unported license

© David A. Bader

**Joint work of** David A. Bader, E. Jason Riedy, Chunxing Yin

Emerging real-world graph problems include: detecting and preventing disease in human populations; revealing community structure in large social networks; and improving the resilience of the electric power grid. Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new challenges because of the sparsity and lack of locality in the data, the need for research on scalable algorithms and development of frameworks for solving these real-world problems on high performance computers, and for improved models that capture the noise and bias inherent in the torrential data streams.

Focusing on parallel algorithm design and implementation, Bader formalizes a practical model for graph analysis on streaming data. In this model, a massive graph undergoes changes from an input stream of edge insertions and removals. The model supports concurrent updating of the graph while algorithms execute concurrently on the dynamic data structure. The talk introduces a concept of validity: an algorithm is valid if the output is correct for a graph consisting of the initial graph with some subset of concurrent changes. Practical examples of this model are given for valid implementations of breadth first search, connected components, PageRank, and triangle counting, all useful graph kernels in real-world applications.

### 3.4 Core-periphery clustering and complex networks

*Pierluigi Crescenzi (University of Florence, IT)*

**License** © Creative Commons BY 3.0 Unported license

© Pierluigi Crescenzi

**Joint work of** Pierluigi Crescenzi, Pierre Fraigniaud, Zvi Lotker, Paolo Penna

**Main reference** Pierluigi Crescenzi, Pierre Fraigniaud, Zvi Lotker, Paolo Penna: “Core-periphery clustering and collaboration networks”, in Proc. of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, pp. 525–528, IEEE Computer Society, 2016.

**URL** <http://dx.doi.org/10.1109/ASONAM.2016.7752285>

In this talk we analyse the core-periphery clustering properties of complex networks, where the core of a network is formed by the nodes with highest degree. In particular, we first observe that, even for random graph models aiming at matching the degree-distribution and/or the clustering coefficient of real networks, these models produce synthetic graphs which have a spatial distribution of the triangles with respect to the core and to the periphery which does not match the spatial distribution of the triangles in the real networks. We



therefore propose a new model, called CPCL, whose aim is to distribute the triangles in a way fitting with their real core-periphery distribution, and thus producing graphs matching the core-periphery clustering of real networks.

### 3.5 Scaling problems with metric constraints

*David F. Gleich (Purdue University – West Lafayette, US)*

**License** © Creative Commons BY 3.0 Unported license  
© David F. Gleich

**Joint work of** David F. Gleich, Nate Veldt, Anthony Wirth, James Saunderson

**Main reference** Nate Veldt, David F. Gleich, Anthony Wirth, James Saunderson: “A Projection Method for Metric-Constrained Optimization”, CoRR, Vol. abs/1806.01678, 2018.

**URL** <http://arxiv.org/abs/1806.01678>

We evaluate techniques to solve LP formulations with metric constraints (triangle inequality constraints). We are able to solve these with a-posteriori approximation guarantees for problems with 10000 datapoints (or roughly 700 billion constraints) on standard desktop hardware through the careful use of projection methods.

### 3.6 SuiteSparse:GraphBLAS: graph algorithms in the language of linear algebra

*Timothy Alden Davis (Texas A&M University – College Station, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Timothy Alden Davis

**Main reference** Timothy A. Davis: “Algorithm 9xx: SuiteSparse:GraphBLAS: graph algorithms in the language of sparse linear algebra”. Submitted to ACM Transactions on Mathematical Software, June 2018

**URL** <http://suitsparse.com>

SuiteSparse:GraphBLAS is a full implementation of the GraphBLAS standard, which defines a set of sparse matrix operations on an extended algebra of semirings using an almost unlimited variety of operators and types. When applied to sparse adjacency matrices, these algebraic operations are equivalent to computations on graphs. GraphBLAS provides a powerful and expressive framework for creating graph algorithms based on the elegant mathematics of sparse matrix operations on a semiring. Key features and performance of the SuiteSparse implementation of GraphBLAS package are described.

### 3.7 Beyond triangles

*Irene Finocchi (Sapienza University of Rome, IT)*

**License** © Creative Commons BY 3.0 Unported license  
© Irene Finocchi

**Joint work of** Irene Finocchi, Marco Finocchi, Emanuele Guido Fusco

**Main reference** Irene Finocchi, Marco Finocchi, Emanuele G. Fusco: “Clique Counting in MapReduce: Algorithms and Experiments”, ACM Journal of Experimental Algorithmics, Vol. 20, pp. 1.7:1–1.7:20, 2015.

**URL** <http://dx.doi.org/10.1145/2794080>

The talk addressed the problem of counting the number  $q_k$  of  $k$ -cliques in large-scale graphs, for any constant  $k \geq 3$ . This is essential in a variety of applications, including social network analysis. Due to the computationally intensive nature of the clique counting problem, we

settle for exact/approximate parallel solutions in the MapReduce framework, discussing both theoretical and experimental contributions. In particular, our algorithms make it possible to compute  $q_k$  for several real-world graphs and shed light on its growth rate as a function of  $k$ .

### 3.8 The Parallel External Memory Model: Sparse Matrix Multiply and List Ranking

*Riko Jacob (IT University of Copenhagen, DK)*

**License** © Creative Commons BY 3.0 Unported license  
© Riko Jacob

**Joint work of** Michael Bender, Gerth Brodal, Rolf Fagerberg, Riko Jacob, Elias Vicari, Tobias Lieber, Nodari Sitchinava

The talk surveyed the parallel external memory model (PEM) and some fundamental results on permuting/sparse matrix multiply. It also gave an impression of a very peculiar lower bound for the list ranking problem, meant as an invitation to work on the open problem of generalizing the lower bound to a more standard setting.

#### References

- 1 Michael A. Bender, Gerth Stølting Brodal, Rolf Fagerberg, Riko Jacob, and Elias Vicari. Optimal sparse matrix dense vector multiplication in the I/O-model. *Theoretical Computer Science*, 47(4):934–962, 2010.
- 2 Riko Jacob, Tobias Lieber, and Nodari Sitchinava. On the complexity of list ranking in the parallel external memory model. In *Proceedings 39th International Symposium on Mathematical Foundations of Computer Science (MFCS'14)*, volume 8635 of *Lecture Notes in Computer Science*, pages 384–395. Springer, 2014.

### 3.9 Theoretically Efficient Parallel Graph Algorithms Can Be Fast and Scalable

*Julian Shun (MIT – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Julian Shun

**Joint work of** Laxman Dhulipala, Guy E. Blelloch, Julian Shun

**Main reference** Laxman Dhulipala, Guy E. Blelloch, Julian Shun: “Theoretically Efficient Parallel Graph Algorithms Can Be Fast and Scalable”, CoRR, Vol. abs/1805.05208, 2018.

**URL** <http://arxiv.org/abs/1805.05208>

There has been significant interest in parallel graph processing recently due to the need to quickly analyze the large graphs available today. Many graph codes have been designed for distributed memory or external memory. However, today even the largest publicly-available real-world graph (the Hyperlink Web graph with over 3.5 billion vertices and 128 billion edges) can fit in the memory of a single commodity multicore server. Nevertheless, most experimental work in the literature report results on much smaller graphs, and the ones that use the Hyperlink graph are done in distributed or external memory. Therefore it is natural to ask whether we can efficiently solve a broad class of graph problems on this graph in memory.

With a graph of this size it is important to use theoretically-efficient parallel algorithms as even minor inefficiencies in the work or parallelism of an algorithm can lead to a significant

increase in running time. This talk shows that theoretically-efficient parallel graph algorithms can scale to the largest publicly-available graphs using a single machine with a terabyte of RAM, processing them in minutes. We give implementations of theoretically-efficient parallel algorithms for 13 important graph problems. We also present the optimizations and techniques that we used in our implementations, which were crucial in enabling us to process these large graphs quickly. We show that the running times of our implementations outperform existing state-of-the-art implementations on the largest real-world graphs. For many of the problems that we consider, this is the first time they have been solved on graphs at this scale.

### 3.10 An Ensemble Framework for Detecting Community Changes in Dynamic Networks

*Christine Klymko (LLNL – Livermore, US)*

**License** © Creative Commons BY 3.0 Unported license

© Christine Klymko

**Joint work of** Timothy La Fond, Geoffrey Sanders, Christine Klymko, Van Emden Henson

**Main reference** Timothy La Fond, Geoffrey Sanders, Christine Klymko, Van Emden Henson: “An ensemble framework for detecting community changes in dynamic networks”, in Proc. of the 2017 IEEE High Performance Extreme Computing Conference, HPEC 2017, Waltham, MA, USA, September 12-14, 2017, pp. 1–6, IEEE, 2017.

**URL** <http://dx.doi.org/10.1109/HPEC.2017.8091035>

Dynamic networks, especially those representing social networks, undergo constant evolution of their community structure over time. Nodes can migrate between different communities, communities can split into multiple new communities, communities can merge together, etc. In order to represent dynamic networks with evolving communities it is essential to use a dynamic model rather than a static one. Here we use a dynamic stochastic block model where the underlying block model is different at different times. In order to represent the structural changes expressed by this dynamic model the network will be split into discrete time segments and a clustering algorithm will assign block memberships for each segment. We show that using an ensemble of clustering assignments accommodates for the variance in scalable clustering algorithms and produces superior results in terms of pairwise-precision and pairwise-recall. We also demonstrate that the dynamic clustering produced by the ensemble can be visualized as a flowchart which encapsulates the community evolution succinctly.

### 3.11 Scalable Inference and Summarization of Multi-source Network Data

*Danai Koutra (University of Michigan – Ann Arbor, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Danai Koutra

- Joint work of** Danai Koutra, Tara Safavi, Chandra Sripada, U Kang, Jilles Vreeken, Neil Shah, Christos Faloutsos, Yujun Yan, Di Jin, Mark Heimann
- Main reference** Tara Safavi, Chandra Sripada, Danai Koutra: “Scalable Hashing-Based Network Discovery”, in Proc. of the 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017, pp. 405–414, IEEE Computer Society, 2017.
- URL** <http://dx.doi.org/10.1109/ICDM.2017.50>
- Main reference** Danai Koutra, U. Kang, Jilles Vreeken, Christos Faloutsos: “VOG: Summarizing and Understanding Large Graphs”, in Proc. of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014, pp. 91–99, SIAM, 2014.
- URL** <http://dx.doi.org/10.1137/1.9781611973440.11>
- Main reference** Neil Shah, Danai Koutra, Tianmin Zou, Brian Gallagher, Christos Faloutsos: “TimeCrunch: Interpretable Dynamic Graph Summarization”, in Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pp. 1055–1064, ACM, 2015.
- URL** <http://dx.doi.org/10.1145/2783258.2783321>
- Main reference** Yujun Yan, Mark Heimann, Di Jin, Danai Koutra: “Fast Flow-based Random Walk with Restart in a Multi-query Setting”, in Proc. of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA., pp. 342–350, SIAM, 2018.
- URL** <http://dx.doi.org/10.1137/1.9781611975321.39>

Networks naturally capture a host of real-world interactions, from social interactions and email communication to brain activity. However, graphs are not always directly observed, especially in scientific domains, such as neuroscience, where monitored brain activity is often captured as time series. How can we efficiently infer networks from time series data (e.g., model the functional organization of brain activity as a network) and speed up the network construction process to scale up to millions of nodes and thousands of graphs? Further, what can be learned about the structure of the graph data? How can we automatically summarize a network ‘conditionally’ to its domain, i.e., summarize its most important properties by taking into account the properties of other graphs in that domain (e.g., neuroscience, social science)? In this talk I will present our recent work on scalable algorithms for inferring, summarizing and mining large collections of graph data coming from different sources. I will also discuss applications in various domains, including connectomics and social science.

### 3.12 Optimization on Graphs

*Rasmus Kyng (Harvard University – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Rasmus Kyng

Many of our favorite questions about graphs are answered by solving optimization problems. For example, we can analyze social networks using clustering and regression, and this leads to problems that are solved using optimization techniques. Similarly, planning flows of goods or data in transportation networks boils down to solving optimization problems. Second order methods are a powerful tool in optimization, but they require solving linear equations, which can be prohibitively expensive. However, when the optimization problem comes from a graph, this adds structure to the linear equations. We can leverage this structure to solve the equations quickly, making second order methods tractable.

### 3.13 Power-law hypothesis for PageRank

*Nelly Litvak (University of Twente, NL)*

- License** © Creative Commons BY 3.0 Unported license  
© Nelly Litvak
- Joint work of** Nelly Litvak, Remco van der Hofstad, Alessandro Garavaglia, Ningyuan Chen, Mariana Olvera-Cravioto
- Main reference** Ningyuan Chen, Nelly Litvak, Mariana Olvera-Cravioto: “Generalized PageRank on directed configuration networks”, *Random Struct. Algorithms*, Vol. 51(2), pp. 237–274, 2017.
- URL** <http://dx.doi.org/10.1002/rsa.20700>
- Main reference** Garavaglia, A., van der Hofstad, R., & Litvak: “Local weak convergence for PageRank”. *arXiv preprint arXiv:1803.06146*.
- URL** <https://arxiv.org/abs/1803.06146>

PageRank is a well-known algorithm for measuring centrality in networks, originally proposed by Google for ranking pages in the World-Wide Web. One of the intriguing properties of PageRank is the so-called “power-law hypothesis”: in a scale-free network the PageRank follows a power law with the same exponent as (in-)degrees. Up to date, this hypothesis has been confirmed many times empirically. However, formalizing and proving this result mathematically turns out to be challenging. In this talk I will discuss recent progress on limiting behavior of PageRank in random graphs. I will first present results on the power-law behavior for PageRank was obtained in the directed configuration model with independent in- and out-degrees. This result follows essentially from the coupling of the graph with a branching tree. I will continue with recent work, where we take a different approach. Instead of focusing on a particular random graph model, we investigate the asymptotic PageRank distribution when the graph size goes to infinity, using the notion of local weak convergence recently developed in the theory of random graphs. To this end, we define an exploration process in the directed setting that keeps track of in- and out-degrees of vertices. Then we use this to prove the existence of the asymptotic PageRank distribution. As a result, the limiting distribution of PageRank can be computed directly as a function of the limiting object. As examples, we apply our results in the directed configuration model with dependent in- and out-degrees, continuous-time branching processes trees, and preferential attachment models.

### 3.14 To Push or To Pull: On Reducing Communication and Synchronization in Graph Computations

*Maciej Besta*

- License** © Creative Commons BY 3.0 Unported license  
© Maciej Besta
- Joint work of** Maciej Besta, Michal Podstawski, Linus Groner, Edgar Solomonik, Torsten Hoefer
- Main reference** Maciej Besta, Michal Podstawski, Linus Groner, Edgar Solomonik, Torsten Hoefer: “To Push or To Pull: On Reducing Communication and Synchronization in Graph Computations”, in *Proc. of the 26th International Symposium on High-Performance Parallel and Distributed Computing, HPDC 2017*, Washington, DC, USA, June 26-30, 2017, pp. 93–104, ACM, 2017.
- URL** <http://dx.doi.org/10.1145/3078597.3078616>

We reduce the cost of communication and synchronization in graph processing by analyzing the fastest way to process graphs: pushing the updates to a shared state or pulling the updates to a private state. We investigate the applicability of this push-pull dichotomy to various algorithms and its impact on complexity, performance, and the amount of used locks, atomics, and reads/writes. We consider 11 graph algorithms, 3 programming models, 2 graph abstractions, and various families of graphs. The conducted analysis illustrates

surprising differences between push and pull variants of different algorithms in performance, speed of convergence, and code complexity; the insights are backed up by performance data from hardware counters. We use these findings to illustrate which variant is faster for each algorithm and to develop generic strategies that enable even higher speedups. Our insights can be used to accelerate graph processing engines or libraries on both massively-parallel shared-memory machines as well as distributed-memory systems

### 3.15 A Toolbox to Extract Structure From Graphs

*Lalla Mouatadid (University of Toronto, CA)*

**License**  Creative Commons BY 3.0 Unported license  
© Lalla Mouatadid

**Joint work of** Michel Habib, Lalla Mouatadid

We survey a few techniques to extract structural properties on certain graph classes. We focus in particular on two techniques: Graph searching and modular decomposition, and illustrate these ideas on certain graph classes. We then introduce a relaxation of modular decomposition ( $\epsilon$ -modules) to try to lift these techniques to less structured graphs.

### 3.16 Dynamic Algorithms and Complexity, A Short Survey

*Danupon Nanongkai (KTH Royal Institute of Technology – Stockholm, SE)*

**License**  Creative Commons BY 3.0 Unported license  
© Danupon Nanongkai

The purpose of this talk is to give people in other fields a taste of the dynamic graph algorithms research in the theory (FOCS/STOC) community. The talk is a compression of a survey given last week at HALG 2018. The original abstract is below.

In this talk I will attempt to answer the following questions I have been asked quite often: What are the current challenges in dynamic graph algorithms? What are good starting points for PhD students and experienced researchers from other fields, who want to try working in this field? The talk will focus on challenges for basic graph problems (e.g. connectivity, shortest paths, maximum matching), and will survey some existing upper and lower bound results and techniques.

### 3.17 Generating random massive graphs that mimic real data

*Cynthia A. Phillips (Sandia National Labs – Albuquerque, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Cynthia A. Phillips

**Joint work of** Cynthia A. Phillips, George Slota, Jonathan Berry, Siva Rajamanickam

We introduce wrapped BTER, a way to take a single input to the Block Two-Level Erdős-Renyi (BTER) scalable graph generator and produce a family of related graphs which vary by tightness of connection within BTER affinity blocks. This gives an alternative to the Lancichinetti-Fortunato-Radicchi (LFR) community detection benchmark capable of

generating billion-edge graphs in less than a minute. We are currently extending this to generate graphs that have the same density, but are scaled up or down in size. We will present the latest results on scaling generation or describe the currently blocking challenges. We feel this technique will provide an important new way to generate realistic data sets for experimental analysis on huge graphs when there is not a lot of real data, or it is millions of nodes instead of billions.

### 3.18 When Exact Fails to be Parallel: Designing Parallel Algorithms through Approximation

Alex Pothen (Purdue University – West Lafayette, US)

License © Creative Commons BY 3.0 Unported license

© Alex Pothen

Joint work of Arif Khan, S M Ferdous, Alex Pothen

We describe a paradigm for designing parallel algorithms that employs approximation techniques. Instead of solving a problem exactly, for which efficient parallel algorithms might not exist, we seek a solution with provable approximation guarantees via approximation algorithms. Furthermore, we design approximation algorithms with high degrees of concurrency. We show  $b$ -edge covers and  $b$ -matchings in graphs as examples of this paradigm.

For  $b$ -edge covers, we describe four techniques for designing approximation algorithms that are concurrent. The first is the Greedy algorithm, the second is the use of a primal-dual formalism, and the third and fourth reduce edge cover computations to computing matchings. We show that some of these algorithms obtain high performance on both shared memory and distributed memory computers. We also describe an application of matchings and edge covers to a problem in data privacy.

#### References

- 1 Arif Khan, Alex Pothen and S M Ferdous, *Designing Parallel Algorithms via Approximation:  $b$ -Edge Cover*, Proceedings of IPDPS, 12 pp., 2018.
- 2 S M Ferdous, Alex Pothen and Arif Khan, *New Approximation Algorithms for Minimum Weighted Edge Cover*, Proceedings of SIAM Workshop on Combinatorial Scientific Computing, 12 pp., 2018.
- 3 Arif Khan, Alex Pothen, Mostofa Patwary, Nadathur Satish, Narayanan Sunderam, Fredrik Manne, Mahantesh Halappanavar and Pradeep Dubey, *Efficient approximation algorithms for weighted  $b$ -Matching*, SIAM J. Scientific Computing, 38(5), S593-S619, 2016.

### 3.19 Topology-induced Enhancement of Mappings

Maria Predari (Universität Köln, DE)

License © Creative Commons BY 3.0 Unported license

© Maria Predari

Joint work of Maria Predari, Henning Meyerhenke, Roland Glantz

Main reference Roland Glantz, Maria Predari, Henning Meyerhenke: “Topology-induced Enhancement of Mappings”, CoRR, Vol. abs/1804.07131, 2018.

URL <http://arxiv.org/abs/1804.07131>

We propose a new method to enhance a mapping  $\mu(\cdot)$  of a parallel application’s computational tasks to the processing elements (PEs) of a parallel computer. The idea behind our method TIMER is to enhance such a mapping by drawing on the observation that many topologies

take the form of a partial cube. This class of graphs includes all rectangular and cubic meshes, any such torus with even extensions in each dimension, all hypercubes, and all trees.

Following previous work, we represent the parallel application and the parallel computer by graphs  $G_a$  and  $G_p$ , respectively.  $G_p$  being a partial cube allows us to label its vertices, the PEs, by bitvectors such that the cost of exchanging one unit of information between any two vertices of  $G_p$  amounts to the Hamming distance between their labels. By transferring these bitvectors to the vertex set of  $G_a$  via  $\mu^{-1}(\cdot)$  and extending them to be unique on  $G_a$ , we can enhance  $\mu(\cdot)$  by swapping labels on  $G_a$  in a new way. Pairs of swapped labels are local w.r.t. the PEs, but not w.r.t.  $G_a$ . Moreover, permutations of the bitvectors' entries give rise to a plethora of hierarchies on the PEs. Through these hierarchies we turn TiMER into a hierarchical method for improving  $\mu(\cdot)$  that is complementary to state-of-the-art methods for computing  $\mu(\cdot)$  in the first place.

In our experiments we use TiMER to enhance mappings of complex networks onto rectangular meshes and tori with 256 and 512 nodes, as well as hypercubes with 256 nodes. It turns out that common quality measure of mappings derived from state-of-the-art tools, such as scotch and KAHIP, can be improved up to 34 %.

### 3.20 A Round-Efficient and Communication-Efficient Algorithm for Distributed Betweenness Centrality

*Vijaya Ramachandran (University of Texas – Austin, US)*

**License** © Creative Commons BY 3.0 Unported license

© Vijaya Ramachandran

**Joint work of** Roshan Dathathri, Gurbinder Gill, Loc Hoang, Keshav Pingali, Matteo Pontecorvi, Vijaya Ramachandran

**Main reference** Matteo Pontecorvi, Vijaya Ramachandran: “Distributed Algorithms for Directed Betweenness Centrality and All Pairs Shortest Paths”, CoRR, Vol. abs/1805.08124, 2018.

**URL** <http://arxiv.org/abs/1805.08124>

Betweenness centrality is a graph computation that has many uses in the analysis of networks. Network graphs today consist of billions of nodes and trillions of edges and do not fit in the memory of a single machine, so distributed algorithms must be used to compute betweenness centrality. However, distributed algorithms are more difficult to implement efficiently than their shared-memory counterparts because the cost of synchronization is much higher.

In this talk, we present a round-efficient and communication-efficient algorithm for betweenness centrality and describe its implementation in D-Galois, a state-of-the-art distributed graph analytics framework. Our efficient distributed betweenness centrality algorithm for unweighted directed graphs runs in  $2n + \mathcal{O}(\mathcal{D})$  rounds in the CONGEST model, where  $n$  is the number of nodes and  $\mathcal{D}$  is the diameter of the graph. We adapt and implement this algorithm in D-Galois by further optimizing our CONGEST distributed algorithm to substantially reduce communication and synchronization costs in the D-Galois execution model. Preliminary experiments with large graphs on big clusters show that this new distributed-memory algorithm is substantially faster than several prior implementations including a state-of-the-art combinatorial BLAS implementation.



### 3.21 Fast Approximate Gaussian Elimination for Laplacians

*Sushant Sachdeva (University of Toronto, CA)*

- License** © Creative Commons BY 3.0 Unported license  
© Sushant Sachdeva
- Joint work of** Rasmus Kyng, Sushant Sachdeva
- Main reference** Rasmus Kyng, Sushant Sachdeva: “Approximate Gaussian Elimination for Laplacians – Fast, Sparse, and Simple”, in Proc. of the IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9–11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA, pp. 573–582, IEEE Computer Society, 2016.
- URL** <http://dx.doi.org/10.1109/FOCS.2016.68>

Solving systems of linear equations in graph Laplacians is a fundamental primitive in scientific computing and optimization. Starting with the seminal work of Spielman-Teng that gave the first nearly-linear time algorithm for solving Laplacian systems, there has been a long line of work giving faster Laplacian solvers. These solvers have had a large impact on the design of fast graph algorithms.

I’ll present a very simple, nearly-linear time Laplacian solver that is based purely on random sampling, and does not use any graph theoretic constructions such as low-stretch trees, sparsifiers, or expanders. Our solver builds a sparse Cholesky factorization for Laplacians – the symmetric version of Gaussian elimination. More precisely, it approximates a Laplacian  $L$  as  $U'U$ , where  $U$  is a sparse upper triangular matrix. Since triangular matrices are easy to invert, this immediately implies a fast Laplacian solver via iterative refinement.

### 3.22 Massively parallel communication-free graph generators

*Peter Sanders (KIT – Karlsruher Institut für Technologie, DE) and  
Manuel Penschuck (Goethe-Universität – Frankfurt, DE)*

- License** © Creative Commons BY 3.0 Unported license  
© Peter Sanders and Manuel Penschuck
- Joint work of** Daniel Funke, Sebastian Lamm, Ulrich Meyer, Manuel Penschuck, Peter Sanders, Christian Schulz, Darren Strash, Moritz von Looz
- Main reference** Daniel Funke, Sebastian Lamm, Peter Sanders, Christian Schulz, Darren Strash, Moritz von Looz: “Communication-Free Massively Distributed Graph Generation”, in Proc. of the 2018 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2018, Vancouver, BC, Canada, May 21–25, 2018, pp. 336–347, IEEE Computer Society, 2018.
- URL** <http://dx.doi.org/10.1109/IPDPS.2018.00043>
- Main reference** Manuel Penschuck: “Generating Practical Random Hyperbolic Graphs in Near-Linear Time and with Sub-Linear Memory”, in Proc. of the 16th International Symposium on Experimental Algorithms, SEA 2017, June 21–23, 2017, London, UK, LIPIcs, Vol. 75, pp. 26:1–26:21, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2017.
- URL** <http://dx.doi.org/10.4230/LIPIcs.SEA.2017.26>

Analyzing massive complex networks yields promising insights about our everyday lives. Building scalable algorithms to do that is a challenging task that requires a careful analysis and extensive evaluation. However, engineering such algorithms is often hindered by the scarcity of publicly available datasets. Network generators serve as a tool to alleviate this problem by providing synthetic instances with controllable parameters.

We present efficient distributed algorithms for generating massive graphs for several popular models of random graphs. By making use of pseudorandomization and divide-and-conquer schemes, our generators follow a communication-free paradigm in the sense that the amount of communication necessary between the processors does not depend on the output size. The resulting generators are often embarrassingly parallel and have a near optimal scaling behavior.

Currently supported models include Erdős-Reny (directed/undirected), Barabasi-Albers scale-free, random geometric (2D/3D), Delaunay triangulations of random point sets (2D/3D), and random hyperbolic.

### 3.23 Practical Kernelization

*Christian Schulz (Universität Wien, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Christian Schulz

**Joint work of** Christian Schulz, Sebastian Lamm, Alexander Noe, Peter Sanders, Darren Strash

Many NP-hard graph problems have been shown to be fixed-parameter tractable (FPT): large inputs can be solved efficiently and provably optimally, as long as some problem parameter is small. Here the parameter measures the ‘difficulty’ of the input in some mathematically well-defined way, for example using the treewidth of the underlying graph. Over the last two decades, significant advances have been made in the design and analysis of FPT algorithms for a wide variety of graph problems. This has resulted in a rich algorithmic toolbox that are by now well-established and are described in several textbooks and surveys. They lead to algorithms that are theoretically efficient: they allow problems of size  $n$  with a parameter value of  $k$  to be solved in time  $f(k)n^c$  for some (exponential) function  $f$  and constant  $c$ , thereby restricting the exponential dependence of the running time to the parameter  $k$  only. However, these theoretical algorithmic ideas have received very little attention from the practical perspective. Few of the new techniques are implemented and tested on real datasets, and their practical potential is far from understood. The rich toolbox of parameterized algorithm theory offers a rich set of algorithmic ideas that are challenging to implement and engineer in practical settings. In this talk, we survey our recent progress in applying reductions/kernelization routines to graph problems in order to obtain algorithms that run fast in practice. Specifically, we look at recent results for the independent set and the minimum cut problem.

### 3.24 Clustering with a faulty oracle

*Charalampos E. Tsourakakis (Harvard University – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Charalampos E. Tsourakakis

**Joint work of** Charalampos E. Tsourakakis, Michael Mitzenmacher, Kasper Green Larsen, Jaroslaw Blasiok, Preetum Nakkiran, Ben Lawson, Vasilis Nakos

**Main reference** Charalampos E. Tsourakakis, Michael Mitzenmacher, Jaroslaw Blasiok, Ben Lawson, Preetum Nakkiran, Vasilis Nakos: “Predicting Positive and Negative Links with Noisy Queries: Theory & Practice”, CoRR, Vol. abs/1709.07308, 2017.

**URL** <http://arxiv.org/abs/1709.07308>

Social networks and interactions in social media involve both positive and negative relationships. Signed graphs capture both types of relationships: positive edges correspond to pairs of “friends”, and negative edges to pairs of “foes”. The *edge sign prediction problem* aims to predict whether an interaction between a pair of nodes will be positive or negative. We provide theoretical results for this problem that motivates natural improvements to recent heuristics for the problem on practical networks.

On the theoretical side, we model the edge sign prediction problem as follows: we are allowed to query any pair of nodes whether they belong to the same cluster or not, but the answer to the query is corrupted with some probability  $0 < q < \frac{1}{2}$ . Let  $\delta = 1 - 2q$  be the bias. We provide an algorithm that recovers all signs correctly with high probability in the presence of noise with  $O(\frac{n \log n}{\delta^2} + \frac{\log^2 n}{\delta^6})$  queries. This is the best known result for this problem for all but tiny  $\delta$ , and improves the recent work of Mazumdar and Saha. Our result naturally generalizes to the case of  $k$  clusters as well. We also provide an algorithm that performs  $O(\frac{n \log n}{\delta^4})$  queries, and uses breadth first search as its main algorithmic primitive. While both the running time and the number of queries are sub-optimal, our result relies on novel theoretical techniques, and naturally suggests the use of edge-disjoint paths as a feature for predicting signs in online social networks. Specifically, we use edge disjoint  $s$ - $t$  paths of short length as a feature for predicting the sign of edge  $(s, t)$  in real-world signed networks. Empirical findings suggest that the use of such paths improves the classification accuracy, especially for pairs of nodes with no common neighbors.

## 4 Working groups

### 4.1 Time segmentation working group

*David A. Bader (Georgia Institute of Technology – Atlanta, US), Irene Finocchi (Sapienza University of Rome, IT), George Karypis (University of Minnesota – Minneapolis, US), Michel A. Kinsy (Boston University, US), Christine Klymko (LLNL – Livermore, US), Danaï Koutra (University of Michigan – Ann Arbor, US), Cynthia A. Phillips (Sandia National Labs – Albuquerque, US), and Ilya Safro (Clemson University, US)*

**License** © Creative Commons BY 3.0 Unported license

© David A. Bader, Irene Finocchi, George Karypis, Michel A. Kinsy, Christine Klymko, Danaï Koutra, Cynthia A. Phillips, and Ilya Safro

This working group explored methods for detecting changing community structure within graph data presented as an edge-stream where edges have time stamps. Within this framework, we divided the problem into two different but related subproblems and discussed each of them, as described below:

1. Burn-in: given the start of a data stream of time-stamped edges, how can one determine at what point they have observed enough data to make running a community detection algorithm reasonable (or any other algorithm concerning higher-order graph properties). Are there other, easier/more concrete metrics that could be used as a proxy?
2. Change detection: assuming one has a given community structure at the start of a data stream, when should one re-partition the network? Due to the fact that many community detection algorithms are computationally expensive, the goal would be to minimize the number of times this repartitioning is done without missing key changes in network structure.
  - a. Could a reconstruction-error based approach which projects the network onto a low-rank model and determines the probability of incoming edges fitting the model work?
  - b. Could we develop a function  $f(E)$  which triggers a repartitioning when it reaches a certain threshold?

## 4.2 Dynamic Graphs Working Group

*Richard Peng (Georgia Institute of Technology – Atlanta, US), Nesreen K. Ahmed (Intel Labs – Santa Clara, US), Rob Bisseling (Utrecht University, NL), George Karypis (University of Minnesota – Minneapolis, US), Danupon Nanongkai (KTH Royal Institute of Technology – Stockholm, SE), Manuel Penschuck (Goethe-Universität – Frankfurt, DE), Alex Pothén (Purdue University – West Lafayette, US), Vijaya Ramachandran (University of Texas – Austin, US), and Julian Shun (MIT – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license

© Richard Peng, Nesreen K. Ahmed, Rob Bisseling, George Karypis, Danupon Nanongkai, Manuel Penschuck, Alex Pothén, Vijaya Ramachandran, and Julian Shun

The discussion largely focused on two ways of modeling dynamically changing graphs stored in distributed settings.

The first is a streaming model where both the computation and the stream of updates happen on the same clock. That is, the arrival of updates (from the stream) continues irrespective of how long it took to process the previous one, in contrast with more standard streaming models where each computation (no matter how expensive) is completed before the arrival of the next update. This model can reflect the notion of computing an answer that is correct w.r.t. some snapshot of the graph. Furthermore, the accuracy of the algorithm can be measured by the difference in time between the query and the snapshot where the answer came from. Also, the difficulty of these problems can be quantified by the rate of arrival of the updates.

For the graph connectivity problem, we showed that under a constant arrival rate, a delay of  $O(n)$  can be achieved. Then considerable discussion went into the possible tradeoffs between arrival rates and delays in query answers.

We then discussed a parallel setting that focuses on the total communication, specifically ways of extending the bulk synchronous parallel (BSP) model to analyze the communications of dynamic graph algorithms. We discussed problems such as deleting from linked lists, hypergraph partitioning, and list ranking under this communication-centric view. Most of the technical discussions focused on the list ranking problem, which seeks to compute for each element in a linked list its distance to the head of the list. Here rearranging the storage so that consecutive portions of the list are on the same machine leads to gains in communication proportional to the storage associated with each processor. These discussions led us to believe that such partitions / rearrangements of data on machines can lead to provable gains in dynamic updates to the data.

## 4.3 k-GRIP: Combinatorial Approaches

*Blair D. Sullivan (North Carolina State University – Raleigh, US)*

**License** © Creative Commons BY 3.0 Unported license

© Blair D. Sullivan

**Joint work of** Blair D. Sullivan, Tim Davis, John Gilbert, Riko Jacob, Fredrik Manne, Lalla Mouatadid, Maria Predari, Christian Schultz, Rasmus Kyng, Alex Pothén, Sushant Sachdeva, Henning Meyerhenke

This group primarily focused on understanding the hardness of  $k$ -GRIP, the  $k$ -edge Graph Robustness Improvement Problem, where  $R$  is an arbitrary measure of “robustness”. This is a slight generalization of the  $k$ -GRIP problem introduced by Henning Meyerhenke, which sets

$R$  to be the total effective Resistance of the graph. Specifically, the problem asks for the set of  $k$  edges whose addition to a graph will minimize the resistance (maximize the robustness).

We considered three notions of robustness: diameter  $R_{\text{diam}}$ , “natural connectivity”  $R_{\text{nc}}$  (the sum of the subgraph centralities), and effective resistance  $R_{\text{eff}}$ . Starting with the diameter, we designed a reduction from  $k$ -SETCOVER to show that this problem is  $\mathcal{NP}$ -hard. We confirmed that this was previously known (Demaine et al, 2014), and the problem has a  $(4 + \varepsilon)$ -approximation. The measure  $R_{\text{nc}}$  was previously shown to be  $\mathcal{NP}$ -hard in the node/edge removal variant of the problem (Chan et al 2014), but the complexity of edge addition was left open. The group has a tentative  $\mathcal{NP}$ -hardness reduction for this problem (using  $k$ -SETCOVER again), but the issue of approximation remains open.

Finally, we returned to the original question of  $R_{\text{eff}}$ , which remains resistant to gadgeteering. Attempts were made in finding faster/better approaches than the naive greedy algorithm – with no immediate conclusive results. Moreover, we established one hurdle in reducing from problems like  $k$ -SETCOVER and  $k$ -CLIQUE: in effective resistance, several paths of length  $x$  can actually be better than a direct connection, thwarting our initial attempts to force the location of edge additions. In later sessions, we considered the problem of Outlier Detection, which is also an  $L_2$  objective, and was established to be  $\mathcal{NP}$ -hard by reduction from minimum bisection, but this remains work in progress. The same is true for establishing measures of robustness that are monotone and sub- or supermodular. These properties would admit an immediate greedy approximation.

We also implemented a heuristic method, and tested its performance on many matrices in the SuiteSparse Matrix Collection.

## 5 Panel discussions

### 5.1 Applications of the theory of random graphs in algorithmic analysis of large networks

*Nelly Litvak (University of Twente, NL)*

License © Creative Commons BY 3.0 Unported license  
© Nelly Litvak

Random graphs have been introduced by Paul Erdős and Alfréd Rényi at the end of 1950s. Initially, they were invented and used to solve difficult combinatorial problems in graph theory. Notice that this was a very different purpose than modeling real-life networks, such as World Wide Web or Twitter, which did not even exist at that time!

The relevance to the network data motivated enormous developments in the theory of random graphs. More and more mathematicians get involved in this research, their background varying from theoretical probability and statistical mechanics to combinatorics and operations research. This has resulted in spectacular recent developments in the theory of random graphs and their applications to networks.

For example, by now we understand very well how the inhomogeneity in the degrees of nodes and presence of hubs affects graph distances (small-world phenomena) and network algorithms such as PageRank. Many new insightful results have been obtained on spreading phenomenon including competitive spreading of information or products. Many approaches, which did not exist before 2000, have by now become standard, such as applications of martingales and continuous time branching processes in the analysis of preferential attachment model – the famous model popularized by Albert and Barabasi, where new nodes arrive

to a network and have preference to connect to nodes with high degrees. Recently several beautiful theoretical concepts have been introduced to formalize a limit of sparse graphs when graph size goes to infinity.

In my own research I analyze algorithms for large networks in the framework of the theory of random graphs. As an outcome of this discussion, I hope to learn about classes of problems where breakthrough in algorithmic analysis of real-life networks can be achieved by building on modern developments in the theory of random graphs.

## Participants

- Nesreen K. Ahmed  
Intel Labs – Santa Clara, US
- Eugenio Angriman  
Universität Köln, DE
- David A. Bader  
Georgia Institute of Technology – Atlanta, US
- Maciej Besta  
ETH Zürich, CH
- Rob Bisseling  
Utrecht University, NL
- Timothy Chu  
Carnegie Mellon University – Pittsburgh, US
- Pierluigi Crescenzi  
University of Florence, IT
- Timothy Alden Davis  
Texas A&M University – College Station, US
- Irene Finocchi  
Sapienza University of Rome, IT
- John Gilbert  
University of California – Santa Barbara, US
- David F. Gleich  
Purdue University – West Lafayette, US
- Riko Jacob  
IT University of Copenhagen, DK
- George Karypis  
University of Minnesota – Minneapolis, US
- Michel A. Kinsy  
Boston University, US
- Marsha Kleinbauer  
TU Kaiserslautern, DE
- Christine Klymko  
LLNL – Livermore, US
- Yiannis Koutis  
NJIT – Newark, US
- Danai Koutra  
University of Michigan – Ann Arbor, US
- Rasmus Kyng  
Harvard University – Cambridge, US
- Nelly Litvak  
University of Twente, NL
- Fredrik Manne  
University of Bergen, NO
- Henning Meyerhenke  
HU Berlin, DE
- Marco Minutoli  
Pacific Northwest National Lab. – Richland, US
- Lalla Mouatadid  
University of Toronto, CA
- Danupon Nanongkai  
KTH Royal Institute of Technology – Stockholm, SE
- Lorenzo Orecchia  
Boston University, US
- Richard Peng  
Georgia Institute of Technology – Atlanta, US
- Manuel Penschuck  
Goethe-Universität – Frankfurt a. M., DE
- Cynthia A. Phillips  
Sandia National Labs – Albuquerque, US
- Alex Pothén  
Purdue University – West Lafayette, US
- Maria Predari  
Universität Köln, DE
- Vijaya Ramachandran  
University of Texas – Austin, US
- Sushant Sachdeva  
University of Toronto, CA
- Ilya Safro  
Clemson University, US
- Peter Sanders  
KIT – Karlsruher Institut für Technologie, DE
- Christian Schulz  
Universität Wien, AT
- Julian Shun  
MIT – Cambridge, US
- Blair D. Sullivan  
North Carolina State University – Raleigh, US
- Charalampos E. Tsourakakis  
Harvard University – Cambridge, US





# Secure Routing for the Internet

Edited by

Phillipa Gill<sup>1</sup>, Adrian Perrig<sup>2</sup>, and Matthias Wählisch<sup>3</sup>

<sup>1</sup> University of Massachusetts – Amherst, US, [phillipa@cs.umass.edu](mailto:phillipa@cs.umass.edu)

<sup>2</sup> ETH Zürich, CH, [adrian.perrig@inf.ethz.ch](mailto:adrian.perrig@inf.ethz.ch)

<sup>3</sup> FU Berlin, DE, [m.waehlich@fu-berlin.de](mailto:m.waehlich@fu-berlin.de)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18242 “Secure Routing for the Internet”, which ran from Monday 11/6 (morning) to Wednesday 13/6 (noon), and employed 27 participants in total (including 3 network operators).

**Seminar** June 10–13, 2018 – <http://www.dagstuhl.de/18242>

**2012 ACM Subject Classification** Computer systems organization → Dependable and fault-tolerant systems and networks, Networks → Network architectures, Networks → Network components, Networks → Network properties, Networks → Network protocols, Networks → Network security, Networks → Network services

**Keywords and phrases** Anonymity, BGP, BGP Prefix Hijack, Denial of Service, Deployment Incentives, Detection, Monitoring, Network Operations, Privacy, Protocol, Public Key Infrastructure, Routing, Routing Policy, Routing Security, Testbed

**Digital Object Identifier** 10.4230/DagRep.8.6.40

**Edited in cooperation with** Vasileios Kotronis

## 1 Executive Summary

*Vasileios Kotronis (FORTH – Heraklion, GR)*

**License**  Creative Commons BY 3.0 Unported license  
© Vasileios Kotronis

The seminar was focused on the following aspects of routing security, mostly in the context of traditional inter-domain routing security: (i) Protocol design vs tooling, (ii) sources of relevant routing data and their accuracy/collection challenges, including policy databases, (iii) the need for metadata and dataset “labelling”, (iv) monitoring and detection of routing attacks and anomalous incidents, such as BGP hijacks and route leaks, incentives for network operators to adopt routing security protocols, (v) testbeds for routing experiments, (vi) hijacks as enabling attacks against ToR and Bitcoin, on the application level, (vii) prevention of routing attacks, (viii) anonymity, privacy and (anti-)censorship. Moreover, we discussed in depth about (ix) PKI and cryptographic verification and protection mechanisms, and their use in securing routing infrastructures, such as the RPKI and BGPsec protocols. Finally, we (x) approached BGP flowspecs, DDoS attacks and QoS in the Internet as separate topics of interest in the field. Another goal of the seminar was to touch upon (xi) future network routing architectures which offer routing security “by design”, especially in light of demanding upcoming applications such as IoT, car-to-car communications, sensor swarms, and wireless routing at scale, and identify related security and privacy concerns and objectives.



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

Secure Routing for the Internet, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 40–62

Editors: Phillipa Gill, Adrian Perrig, and Matthias Wählisch



Dagstuhl Reports

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



Besides the specific goals of the seminar, it is also worth noting some interesting aspects of Dagstuhl seminars in general, that played a critical role in fueling the related talks, discussions and reports. In summary, the 3-day seminar in which we participated, focused not solely on the presentation of established results but also on ideas, sketches, and open (research and operations) problems. The pace and program was guided by topics and presentations that evolved through discussions. This report contains an executive summary of the material that was transcribed during the entire seminar.

Overall, some participants of the seminar seem to be more “pessimistic” about routing security. Both the research and operator communities need to consolidate more data sources to facilitate progress. Any deployment progress is only possible if operator incentives are improved, however, it remains an open problem on how to provide strong incentives. In practice, a good technical solution is insufficient without first tackling the “politics”. We discussed about routing/network testbeds and the role they can play in emulating and verifying many of the discussed concepts. However, in the wild (or the “real world”), it is surprisingly hard to implement something like RPKI; even more so for BGPsec. We all need a better understanding of the problem space; formal taxonomies of routing attacks, such as hijacks, would be of great help on this front. Regarding improving BGP itself, we have seen many prevention mechanisms, whose deployment is the end-goal for the Internet. However, as we have to live with BGP at least in the intermediate term, we can also explore research on overlay solutions to achieve the properties that we need, at least for the time being. These solutions need to support incremental deployment for obvious reasons.

In general, deployment progress has been slow which is feared not change in the near future. It is reassuring to see that a lot of work is being done in the measurement area; we were also reminded how hard is it to get the ground truth, labelled with useful metadata. Some fundamentally new and secure approaches were discussed, for instance the SCION secure Internet architecture, however, the deployment of new inter-domain routing protocols is very challenging. To improve the deployment incentives of secure routing protocols for operators, the creation of a catalog of routing incidents could be beneficial.

Moreover, it seems that the community may have underestimated the importance of monitoring tools and their utility in the wild. We have learned about new data sets, as well as interesting insights on the Impact of prefix hijacks on the application layer. In general though, we were hoping to see more enthusiasm for new solutions.

Finally, it is worth noting that having a mixed group of researchers and operators is very important to exchange information and discuss potential approaches, which made the seminar an interesting and worthwhile experience.

## 2 Table of Contents

### Executive Summary

<i>Vasileios Kotronis</i> . . . . .	40
-------------------------------------	----

### Overview of Talks

PEERING: An AS for Us <i>Ítalo Cunha</i> . . . . .	44
ARTEMIS: Neutralizing BGP Hijacking within a Minute <i>Vasileios Kotronis</i> . . . . .	44
Next-Generation Public-Key Infrastructures <i>Adrian Perrig</i> . . . . .	45
SCIONLab: A Next-Generation Internet Architecture Testbed you can use Today <i>Adrian Perrig</i> . . . . .	45
Hijacking Bitcoin: Routing Attacks on Cryptocurrencies <i>Laurent Vanbever</i> . . . . .	46
Hijacks: myth or reality? <i>Pierre-Antoine Vervier and Marc C. Dacier</i> . . . . .	46
An RPKI Primer <i>Matthias Wählisch</i> . . . . .	47

### Working groups

Anonymity <i>Nikita Borisov</i> . . . . .	47
Operations of (R)PKI <i>Georg Carle</i> . . . . .	48
Hijack Detection <i>Ítalo Cunha</i> . . . . .	49
Policy Databases <i>Ítalo Cunha</i> . . . . .	51
Data Accuracy Breakout Group <i>Victoria Manfredi</i> . . . . .	51
Routing Security Incentives <i>Adrian Perrig</i> . . . . .	53
Metadata <i>Pierre-Antoine Vervier</i> . . . . .	54

### Panel discussions

BGP Flowspecs <i>Vasileios Kotronis</i> . . . . .	56
Data Sources <i>Vasileios Kotronis</i> . . . . .	56
DDoS / Routing Security Attacks <i>Adrian Perrig</i> . . . . .	60

Quality of Service (QoS)  
    *Adrian Perrig* . . . . . 61

**Participants** . . . . . 62

### 3 Overview of Talks

#### 3.1 PEERING: An AS for Us

*Ítalo Cunha (Federal University of Minas Gerais-Belo Horizonte, BR)*

**License** © Creative Commons BY 3.0 Unported license

© Ítalo Cunha

**Joint work of** Brandon Schlinder, Kyriakos Zarifis, Ítalo S. Cunha, Nick Feamster, Ethan Katz-Bassett

**Main reference** Brandon Schlinder, Kyriakos Zarifis, Ítalo S. Cunha, Nick Feamster, Ethan Katz-Bassett: “PEERING: An AS for Us”, in Proc. of the 13th ACM Workshop on Hot Topics in Networks, HotNets-XIII, Los Angeles, CA, USA, October 27-28, 2014, pp. 18:1–18:7, ACM, 2014.

**URL** <https://doi.org/10.1145/2670518.2673887>

Internet routing suffers from persistent and transient failures, circuitous routes, oscillations, and prefix hijacks. A major impediment to progress is the lack of ways to conduct impactful inter-domain research. Most research is based either on passive observation of existing routes, keeping researchers from assessing how the Internet will respond to route or policy changes; or simulations, which are restricted by limitations in our understanding of topology and policy. We propose a new class of inter-domain research: researchers can instantiate an AS of their choice, including its intra-domain topology and inter-domain interconnectivity, and connect it with the “live” Internet to exchange routes and traffic with real inter-domain neighbors. Instead of being observers of the Internet ecosystem, researchers become members. Towards this end, we present the Peering testbed. In its nascent stage, the testbed has proven extremely useful, resulting in a series of studies that were nearly impossible for researchers to conduct in the past. In this paper, we present a vision of what the testbed can provide. We sketch how to extend the testbed to enable future innovation, taking advantage of the rise of IXPs to expand our testbed.

#### 3.2 ARTEMIS: Neutralizing BGP Hijacking within a Minute

*Vasileios Kotronis (FORTH – Heraklion, GR)*

**License** © Creative Commons BY 3.0 Unported license

© Vasileios Kotronis

**Joint work of** Pavlos Sermpezis, Vasileios Kotronis, Petros Gigis, Xenofontas A. Dimitropoulos, Danilo Cicalese, Alistair King, Alberto Dainotti

**Main reference** Pavlos Sermpezis, Vasileios Kotronis, Petros Gigis, Xenofontas A. Dimitropoulos, Danilo Cicalese, Alistair King, Alberto Dainotti: “ARTEMIS: Neutralizing BGP Hijacking within a Minute”, CoRR, Vol. abs/1801.01085, 2018.

**URL** <https://arxiv.org/abs/1801.01085>

ARTEMIS (Automatic and Real-Time dEtection and Mitigation System), is a research effort between the INSPIRE group, FORTH, Greece ([www.inspire.edu.gr](http://www.inspire.edu.gr)) and the Center for Applied Internet Data Analysis (CAIDA), University of California San Diego, USA ([www.caida.org](http://www.caida.org)). ARTEMIS is a defense approach versus BGP prefix hijacking attacks (a) based on accurate and fast detection operated by the AS itself, leveraging the pervasiveness of publicly available BGP monitoring services and their recent shift towards real-time streaming, thus (b) enabling flexible and fast mitigation of hijacking events. Compared to existing approaches/tools, ARTEMIS combines characteristics desirable to network operators such as comprehensiveness, accuracy, speed, privacy, and flexibility. With the ARTEMIS approach, prefix hijacking can be neutralized within a minute.

### 3.3 Next-Generation Public-Key Infrastructures

*Adrian Perrig (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Adrian Perrig

**Joint work of** all seminar participants

Public-key infrastructures form the core of authentication systems that are in use in today's Internet. Unfortunately, the inadequacies of the design of currently used PKIs are emerging with the constant evolution of the Internet and its uses.

In this talk, we discuss the different types of PKIs that are needed to secure Internet communication, and show how we can design next-generation PKIs to achieve better scalability, security, trust agility, and usability.

In particular, we address the following challenges. How can we design a highly available PKI system to support a routing infrastructure? Can we design a PKI that allows to control/limit the power of authorities (e.g., no kill switch possibilities)? How can we securely, scalably, and efficiently update compromised root keys? What considerations do we have for the design a DNS PKI? Should we base the TLS PKI on the DNS PKI as proposed in DANE? Or should we design a TLS PKI that is independent of a secure DNS system? What are the human aspects of running a PKI of an ISP?

In terms of “kill-switches” [1], nation-state adversaries could potentially “turn off” communication for entire regions; could their power be limited through the user of isolation domains? Another major challenge in PKI is key management and tooling (hosted vs non-hosted model, management of private keys, BBN, key revocation). Monitoring is of paramount importance; heuristics used for tracking suspicious information/changes and for taking action are still open research questions. Open-source libraries to ease developer's processes for secure applications should also be considered and developed. Distributed ledgers (e.g., blockchain) cannot currently serve as a solution, due to the critical cyclic effect of communication between the participating nodes (routing – routing verification dependency).

#### References

- 1 Cooper, Danny, et al. “On the risk of misbehaving RPKI authorities.” Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks. ACM, 2013.

### 3.4 SCIONLab: A Next-Generation Internet Architecture Testbed you can use Today

*Adrian Perrig (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Adrian Perrig

**Joint work of** Adrian Perrig, Pawel Szalachowski, Raphael M. Reischuk, Laurent Chuat

**Main reference** Adrian Perrig, Pawel Szalachowski, Raphael M. Reischuk, Laurent Chuat: “SCION: A Secure Internet Architecture”, Springer, 2017.

**URL** <http://dx.doi.org/10.1007/978-3-319-67080-5>

The Internet has not been designed for high availability in the face of malicious actions by adversaries. Recent patches improving security and availability are constrained by the current Internet architecture, business, and legal aspects.

To address these issues, we propose SCION [1], a next-generation Internet architecture that is secure, available, offers privacy, and considers economic and policy issues at the design stage.

We have implemented SCION and deployed it worldwide as a global testbed called SCIONLab, which consists of more than 20 collaborators including research institutions, companies and ISPs. With SCIONLab, researchers can explore today the desirable properties that a next-generation secure Internet architecture can provide.

#### References

- 1 Perrig, Adrian, et al. SCION: a secure Internet architecture. Springer International Publishing, 2017.

### 3.5 Hijacking Bitcoin: Routing Attacks on Cryptocurrencies

*Laurent Vanbever (ETH Zürich, CH)*

**License**  Creative Commons BY 3.0 Unported license  
© Laurent Vanbever

We study the impact that Internet routing attacks (such as BGP hijacks) and malicious Internet Service Providers (ISP) can have on the Bitcoin cryptocurrency. Because of the extreme efficiency of Internet routing attacks and the centralization of the Bitcoin network in few networks worldwide, we show that the following two attacks are practically possible today:

- Partition attack: Any ISP can partition the Bitcoin network by hijacking few IP prefixes.
- Delay attack: Any ISP carrying traffic from and/or to a Bitcoin node can delay its block propagation by 20 minutes while staying completely under the radar.

The potential damage to Bitcoin is worrying. Among others, these attacks could reduce miner's revenue and render the network much more susceptible to double spending. These attacks could also prevent merchants, exchanges and other large entities that hold bitcoins from performing transactions.

#### References

- 1 H Maria Apostolaki, Aviv Zohar, Laurent Vanbever, *Hijacking Bitcoin: Routing Attacks on Cryptocurrencies* IEEE Symposium on Security and Privacy 2017. San Jose, CA , USA (May 2017).

### 3.6 Hijacks: myth or reality?

*Pierre-Antoine Vervier (Symantec Research Labs – Sophia Antipolis, FR) and Marc C. Dacier (EURECOM – Sophia Antipolis, FR)*

**License**  Creative Commons BY 3.0 Unported license  
© Pierre-Antoine Vervier and Marc C. Dacier

Some recent research presented evidence of blocks of IP addresses being stolen by BGP hijacks to launch spam campaigns. This was the first time BGP hijacking was seen in the wild. Since then, only a very few anecdotal cases have been reported as if hackers were not interested in running these attacks. However, it is a common belief among network operators and ISPs that these attacks could be taking place but, so far, no one produced evidence to back up that claim.

In this talk, we report on the analysis of 4 years of data collected by an infrastructure specifically designed to answer that question: are intentional stealthy BGP hijacks routinely taking place in the Internet. The identification of what we believe of more than 5,000 malicious hijacks leads to a positive answer. The lack of ground truth is, of course, a problem. We managed to get confirmation of some of our findings, thanks to an ISP unwittingly involved.

The talk aims of being an eye-opener for the community by shedding some light on this undocumented threat. Depending on BGP attacks that are carried out, they can be very disruptive for the whole Internet and should be looked at very carefully.

### 3.7 An RPKI Primer

*Matthias Wählisch (FU Berlin, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Matthias Wählisch

A fundamental part for securing BGP is the Resource Public Key Infrastructure (RPKI), which consists of a distributed public key infrastructure responsible for Internet resources, i.e., AS numbers and IP prefixes. An RPKI repository stores certificates and Route Origin Authorization (ROAs) objects. A ROA provides a secure binding between one or multiple IP prefixes and an AS that is allowed to originate that prefix.

Using ROA data, an RPKI-enabled router is able to verify the BGP updates it receives. The prefix information within the BGP update might be valid (i.e., the origin AS is allowed to announce this prefix), invalid (i.e., the origin AS is incorrect or the announced prefix is too specific), or not found (i.e., the announced prefix is not covered by the RPKI). Rejecting an invalid route helps to successfully suppress an incorrectly announced prefix, which finally secures network layer reachability of services assigned with an IP address of this prefix.

In this talk, we will give a brief overview about the design, implementation, and deployment of the RPKI.

## 4 Working groups

### 4.1 Anonymity

*Nikita Borisov (University of Illinois – Urbana Champaign, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Nikita Borisov

**Joint work of** anonymity breakout group participants

We first investigated the relationship between censorship and hijacking. For example, we ask the following questions:

- Could BGP hijacking be used for anti-censorship?
- In case hijacking is actually used for censorship (e.g., China Telecom case, Pakistan/Youtube hijack), then how could we do an associated analysis of the event? Could it be circumvented, e.g., via deaggregation? Moreover, it would be useful to have mechanisms to detect when hijacks leak out and prevent collateral damage.

Moreover, we discussed how one could prevent on-path and hijack attacks on ToR, by considering the following possibilities:

- Investigate strategic placement of ToR nodes, e.g., close to destinations to prevent destination-based attacks.
- Use feeds of information on BGP changes, by e.g., “subscribing” to the hijack alert feed of one or multiple hijack detection systems.
- Take into account that unlike everyday routing, a false-positive in ToR is far less costly.
- Investigate different optimal placement/interconnection strategies for different Internet applications (ToR, BTC, email, etc.).

As a thought experiment, we consider the requirements of a secure routing protocol in the context of ToR. Impossibility of performing hijacks is good, but maybe a weaker requirement would be more than enough for ToR, such as limited hijacking capability. Moreover, the protocol/mechanism should ideally provide the capability of notifications of potential hijacking events, as well as potential control and transparency of end-to-end paths. However, all this seems to run counter to ISP culture of “secrecy”.

We further identify the following benefits that RPKI would bring for protecting ToR against hijacks:

- The attacker would need to hijack potentially longer paths to complete a successful attack.
- The attacker cannot announce more specifics if ROA limits the maximum prefix length, thus preventing many sub-prefix hijacks.
- When combined with BGPsec, the authenticity of the control plane would lead to much better robustness of the routing infrastructure.

However, the data plane (on which e.g., ToR packets flow) could still differ for malicious or benign reasons. We need to investigate the frequency with which inconsistent setups occur in the wild.

Finally, we ask whether there are benefits of partial deployment of routing security protocols. For example, we could locate relays which are hosted in networks covered by ROAs and matching valid announcements. We could also prefer paths to/via relays that have BGPsec validation. However, how much deployment is eventually useful? This remains an open question.

## 4.2 Operations of (R)PKI

*Georg Carle (TU München, DE)*

License © Creative Commons BY 3.0 Unported license

© Georg Carle

Joint work of operations of (R)PKI breakout group participants

The goal we target is to design a PKI system with high availability in the presence of human errors. For this we need formal verification of PKI (including actions done by human entities). However, a formal model would also possess certain limitations; for example, the perceived (by humans) system state might differ from the actual system state. Moreover, some artifacts may be missing, while others may be incorrect. Focusing on the requirement of formally specifying and verifying RPKI, we ask the following question: “Which are the right consistency models for (R)PKI?”



While trying to answer that question, operational aspects of RPKI may be the most difficult aspect to account for. For example, during its early operation stages, blackouts from the main databases occurred, but problems were not sufficiently analyzed. Regarding the structuring of operation, we note that one part is comparable to operation of a X.509-based PKI, while another part is routing-specific. Finally, there is the issue of management of private keys on routers, which could be dealt with e.g., HSM, Intel SGX, ASM trustzone, etc. This is critical: what would happen if an adversary who compromised the router accesses the private key?

In order to formally assess RPKI, one can take a leaf out of the web security book. For example, the following work looks at the actual deployment of HTTPS, also assessing deployment effort and availability risks, which are critical metrics also for (R)PKI [1].

## References

- 1 Amann, Johanna, et al. “Mission accomplished?: HTTPS security after DigiNotar.” Proceedings of the 2017 Internet Measurement Conference. ACM, 2017.

## 4.3 Hijack Detection

Ítalo Cunha (Federal University of Minas Gerais-Belo Horizonte, BR)

License © Creative Commons BY 3.0 Unported license

© Ítalo Cunha

Joint work of hijack detection breakout group participants

We focused on the detection of BGP prefix hijacks. First, we classify them as follows:

- Malicious vs. accidental. A catalog could be useful for tagging accidental hijacks. An open question is: do these events differ w.r.t. their signature on the control/data plane? Knowledge of policies might be useful to identify the type of hijack.
- Prefix manipulation. This can be further divided into the following categories, depending on the kind of prefix advertised fraudulently:
  - Squatting, i.e., advertisement of unused prefixes that are not owned by the attacker AS.
  - Same-granularity (exact prefix).
  - More specifics (sub-prefix).
- Manipulation of the AS-PATH attribute. This can be divided into the following categories, depending on the rightmost location of a fake AS being present on the path:
  - Type-0 (fake origin)
  - Type-1 (fake first hop)
  - Type-2, ..., Type-N (fake N-hop)
  - Type-U (no path manipulation)
- How the packets are handled on the data plane. An attacker could drop (blackholing) or detour packets (man-in-the-middle), or terminate connections and use legitimate IPs to perform impersonation attacks.

An example of such a detection system is ARTEMIS, which for example detects (among others) AS-adjacencies that do not make sense, using previous AS-path information, link information, ground truth and other real-time or offline data sources.

Coming back to the basics, we ask the critical question of who wants to detect hijacks. We identify the following entities:

- The owner of the prefix. He/she requires visibility of all the routes towards their prefix. However, stealthy man-in-the-middle detours are hard to detect, even for owners.
- Third-parties. They are accompanied with a lot of ambiguity. For example, BGP “optimizers” may change routes without notice; on the control-plane, they might look very similar to hijacks.

Moreover, the power of the hijackers themselves depends on different factors, such as (i) the number of peers/customers, (ii) the provider cone, (iii) the respective locations and connectivity of the target/victim and attacker networks, as well as (iv) the stealthiness of the hijacker, e.g., in case of man-in-the-middle attacks. Note also that for example, shorter paths are generally harder to hijack, since they are more preferred by networks.

We further identified useful data sources for hijack detection:

- BGP itself (prefixes, AS-paths, topology). This can also be used to estimate AS sizes (e.g., depending on their connectivity and number of prefixes they advertise), as well as the most known/used peering links.
- BGP communities. However, these might get mangled, while knowledge about policies might be required to understand and properly use communities.
- Routing policies. They are private and quite dynamic.
- Data plane measurements (e.g., latency, bandwidth, traceroutes). They could be triggered via control-plane signaling, in order to detect more stealthy attacks. While they are very sensitive to factors not related to actual attacks (e.g., congestion), there is a lot of historical information available; moreover, they could possibly be very useful to the prefix owner, who knows the data-plane behavior of his/her own network.

Finally, we identified some open issues pertaining to hijack detection:

- The BGPv4 RFC does not mandate the following: if AS1 and AS2 neighbors, then a route to a prefix advertised by AS1 to AS2, is valid even if AS1 is not present on-path. Checking that one has received a route from its peer, and that the peer is present as the last-hop AS on the associated path, is not mandatory leading to all kinds of possible attacks.
- BGP optimizers monitor communications (and their quality) on critical paths, and “play” with BGP advertisements and traffic exit points in case, e.g., the latency increases. This is an attempt to avoid congestion events. However, such advertisements should be flagged by their originator to avoid false positives; for example consider the case where a customer has a private peering (with private ASNs) with its providers, and “optimized” advertisements leak to the outside world.
- It is quite hard to filter prefixes at the border of the Internet (i.e., close to the stubs), despite the availability of guides and best filtering practices, such as BCP38. In general, there are no incentives for operators to do this aggressively; they filter only what their customers tell them. Closer to the core of the Internet, performing filtering for transit is technically infeasible.

## 4.4 Policy Databases

*Ítalo Cunha (Federal University of Minas Gerais-Belo Horizonte, BR)*

**License** © Creative Commons BY 3.0 Unported license  
© Ítalo Cunha

**Joint work of** policy databases breakout group participants

The state of the art w.r.t. routing policy databases consists of Internet Routing Registries (IRRs) and the Routing Policy Specification Language (RPSL). Route objects within IRRs have some kind of authorization, but there is no information about the length of the IP prefix; some providers do not add more specifics to the DB. However, maximum length, e.g., in RPKI is required. It is worth noting that ISPs such as AT&T and Level3 have and maintain their own databases. AS-SET/RS-SET objects lack authorization function (“no man’s land”). Moreover, there is no support for all possible real-world policies.

In terms of route authorization, since RPSL databases are not suitable for this task, RPKI could be employed. The generated route objects could then be safely inserted to an RPSL DB; essentially using RPKI for bootstrapping RPSL authorization. Mechanisms on securing RPSL objects with RPKI signatures can be found at RFC7909.

Other ideas to make progress in this regard are the following:

- Track route propagation authorization.
- Make documentation and authorization workable processes for operations.
- Prefer operational relevance over provisioning of data for research.
- Use BGP monitoring tools to check alarms (e.g., about policy violations, route leaks).
- Identify and model what operators need to have in the DB, subject to RPSL limitations.

Moreover, we should note that tooling is necessary, but current tools have their own limitations. E.g., the `IRRtool` that came out with RPSL is not accompanied by useful documentation, while `rtconfig`, which is useful for generating configurations for major routers, might be tricky to apply in the wild. In general, there is no tool that is commonly agreed upon by the community. Moreover, continuous learning and formation of incentives is critical; operations, as a set of processes itself, is not rocker science.

## 4.5 Data Accuracy Breakout Group

*Victoria Manfredi (Wesleyan University – Middletown, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Victoria Manfredi

**Joint work of** data accuracy breakout group participants

The focus of the breakout group was put on traceroute, and in particular on its following aspects:

- How can we collect the needed Internet path data (active traceroute, using network monitoring points such as RIPE Atlas probes, information in databases, etc.)?
- Is traceroute accurate? Does data (i.e., actual user traffic) follow the same route that traceroute uses?
- How can we infer non-responding intermediate hops?
- Application of Paris traceroute in combination with sampling to infer TE changes.
- Limitations of traceroute, and relevant open questions:

- Showing only a single route per run. How can we access information on alternate (e.g., backup) routes?
- One way is to sample from multiple sources to the same destination. However, which paths should we sample? Could architectures like SCION be of help to specify what paths we want to use for sampling?
- What is the utility of information about multiple routes within an AS, and from an AS to other ASes? Can other ASes leverage this information in some way, even if it is not available in BGP? What is the utility of this for research?
- If one makes statistical assumptions about actual network topology, combined with traceroute information, can the accuracy gap be evaluated, e.g., in terms of variance? How can this be done? One idea in this context is to combine this process with timing information (to improve accuracy) and network tomography mechanisms (to infer connectivity from timing measurements).
- There are differences depending on whether you're an sampling end-user or on-path (AS) sources. What if we combine samples from both sources? W.r.t. the end-user, one samples close to the "edge"; on-path ASes would probably sample on their border gateway routers. What if no cooperation between different ASes exists though? Would this sampling be of use? Moreover, the ASes themselves can make routing decisions that implement the decisions of others, causing different routes to be used (and thus sampled).
- Meta-sampling:
  - Accuracy of information.
  - Completeness of information. For example, are all feasible routes required to be known?
  - Per-route costs.
  - Can this information be aggregated?
  - Relationships between ASes (a "social" graph of ASes). Is this level of sampling abstract enough? If yes, how can this information be leveraged for bootstrapping?
  - Relationship of sampling to link bandwidth. For example, a low-bandwidth edge may be harder to sample, but maybe it is less important anyway.
- How dynamic is the collected information (route, BGP advertisement, connectivity)? Depending on the dynamics and importance of the available information, one can choose to use it for different reasons: e.g., an ISP to detect a live attack, or an academic user to verify information.
- Information in databases:
  - Can it be automatically verified?
  - What are the most prevalent reasons for inaccuracies? Examples include: out-of-date information, information hiding, human error or malicious intent.
  - Can one bootstrap queries from trusted data, to prove or disprove untrusted data?
  - An important motivation for keeping accurate information in the database, is that it helps detect different kinds of attacks, such as BGP hijacks.

Potentially useful related work in the context of this working group includes network discovery mechanisms from ad hoc networks, such as from non-cooperating nodes, as well as sampling of large graphs, such as the Facebook social network.

## 4.6 Routing Security Incentives

*Adrian Perrig (ETH Zürich, CH)*

License © Creative Commons BY 3.0 Unported license  
© Adrian Perrig

Joint work of routing security incentives breakout group participants

First of all, there are incentives both for “good” and “bad” guys. The joint use of RPKI and BGPsec to ensure some kind of routing security is probably not going to happen any time soon; it is too expensive and a partial deployment would be completely useless. Even RPKI alone, as long as its deployment is not complete, is not very useful/impactful.

A trade-off solution would be to create communities of ISPs and other organizations deploying, for instance, RPKI. If the whole community deploys RPKI then they’re in a position to enforce it (e.g., drop traffic violating ROAs). Ecuador is a particular example case where all ISPs are connected to a single IXP; the IXP enforces participating ISPs to use RPKI.

Another idea is to deploy routing security protocols in a large (e.g., research) sub-network like Internet2, where one can also apply and enforce policies they wouldn’t be able to apply on the whole Internet (e.g., deploy RPKI, BGPsec).

It is important to encourage people to certify their prefixes, by taking advantages of large network operator meetings such as IETF and NANOG. To achieve that, one needs to educate people to convince them that this process useful and important and make it easy for ISPs to certify prefixes and create ROAs (for example, by using a nice and usable portal). Ease of deployment and maintenance is of paramount importance. Moreover, note that there is a big difference in adoption between EU and US, as the influence from government interest, research funding, customer requirements/procurement are different. A critical challenge is that, unlike RPKI, partial/incremental deployment of BGPsec would not bring any added value; forcing adoption of RPKI/BGPsec requires that partial/incremental adoption will create incremental added value to the organization doing it.

Other potential incentives for ISPs to deploy routing security protocols are the following:

- Defense against attacks targeting their reputation or aiming at decreasing their income (inducing loss).
- Non-deployment would decrease the revenue of the non-deployer.
- Attacks could be turned into revenue-affecting events by getting attention on the victim (e.g., Youtube hijack)
- In the past, high-profile attacks lead to big leaps forward in security.

A “sad” conclusion is that ISPs and Internet organizations in general will always be more prone to doing detection rather than prevention. However, detection will never provide a perfect protection. In theory, RPKI+BGPsec should supposedly avoid high FP (False Positives) rates coming with detection techniques.

Next, we are looking at incentives for bad guys to abuse the routing infrastructure. The vulnerability of the infrastructure has been known for years, so how is this abuse not more common today? A potential response is that maybe there are much more hijacks than we observe/hear about. A lot of networks could be hijacked and don’t want to disclose it. In the early 2000’s, around 1/week, 2/month hijacks would be reported (Cisco), in parallel with the telephony network that also suffered from hijacks. Moreover, w.r.t. DDoS attacks, is it cheaper/easier for cybercriminals to just rely on booter services rather than entering the BGP hijack game. However, in the context of man-in-the-middle attacks, BGP hijacks would be a big enabler, for instance, to perform delay attacks.

It is also worth noting that even “bad guys” need the Internet to work to run their business. That being said, all techniques to perform targeted BGP hijacks are well known and can be used by them. An automated BGP hijacking tool, similar to booter services for DDoS attacks, could provide cybercriminals enough incentives to start leveraging that kind of attacks.

As a conclusion, we may not yet be at a stage where BGP hijacking is easy enough to see a lot of bad guys do it at scale.

## 4.7 Metadata

*Pierre-Antoine Vervier (Symantec Research Labs – Sophia Antipolis, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Pierre-Antoine Vervier

**Joint work of** metadata breakout group participants

We first discussed some examples of what metadata might be needed; for example, information on peering routers as well as the routing policies of the monitors would be useful metadata to annotate RouteViews datasets.

The most profound questions related to metadata are the following:

- Who needs metadata?
- For what purpose are metadata required?
- Depending on the data user, do the metadata requirements differ?
- Who is going to collect the metadata?

As a use case, we dig in example datasets and see what metadata could improve them. An example would be the labeling of data from bgpstream.com with information from network operators, as follows:

```
[AS1, AS2, AS3, Prefix1] : hijack by AS2 against Prefix1, or:
[AS4, AS5, AS6, Prefix2] : route leak from AS5 which acted as upstream of AS6.
```

However, this generates some challenges:

- What would be the incentives for operators to do that? Note that labeling large datasets requires joint effort from a lot of people.
- One motivation is to improve the quality of the feed by providing it with feedback about the incidents. However, ISPs are often reluctant to confessing they made a mistake.
- On the other hand, the victim itself might be ready to stand up; in the end, this approach could be turned into a reputation system.

We proceed with another example dataset which would benefit from metadata: jupyter notebooks, for which we envision the following uses:

- Publish along with research papers to encourage other researchers to reproduce their experiments.
- This would encourage researchers to be more transparent w.r.t. their work.
- This would also deal with the issue of reproducibility; reproducibility is often hindered by the unavailability of datasets, which are not necessarily open.
- This would require the whole data and code to be packaged together. Frameworks, such as BGPstream from CAIDA, could make this reproducibility easier by providing a uniform access to data.

Another example of useful metadata is the inter-AS business relationship dataset inferred from inter-AS links. While an initial, large-scale effort is done by CAIDA, such datasets are often kept private to avoid disclosing business relationships. The consequence is that everyone kind of redoes this mapping with their own approach; there is no incentive from ISPs' perspectives to make this information public as metadata. The same applies in general to network policies.

An interesting question is whether labeling e.g., BGP routing announcements would facilitate/enable using clustering/ML techniques to detect abnormal events (route leaks, policy violations, BGP hijacks, etc.). We are not so certain about the answer. The core problems are feature engineering and labeling. We need a ground truth. We also need to factor in the fact that attackers will try to fool the system, simply by changing their behavior. Results should also be finally explainable.

Moreover, we should also investigate how we should annotate data so as to maximize the added value of the metadata to improve the detection/information learned from the labeled data. Another way to look at the metadata problem is to try to label events by learning the processes/practices used by network operators that lead to the observed events (instead of trying to have the network operators label the data themselves).

On the hijack detection front, additional metadata can be collected and added to the raw data by using the output from:

- Hijack detection systems (e.g., presence of SSL/TLS-enabled server, the output from ARTEMIS, the output from bgpstream.com, etc.).
- IP-based reputation (spam senders, hosting malicious websites, etc.).
- AS-based reputation (see [circl.lu](http://circl.lu))

However, we need to find if there are other datasets different than “hijack candidates/abnormal events” that could benefit from metadata (e.g., data-plane data, IRR).

An interesting research direction is to investigate how testbeds (e.g., PEERING) could be used to enrich current routing-related datasets. For example, they could be used for testing how malicious announcements are propagated, and test patterns w.r.t. different attack/anomaly categories/classes. However, the value of these data might be limited due to inherent testbed limitations; for example, one can only hijack his/her own prefixes.

A suitable partner to initiate the routing data labeling would be Internet2 and/or Geant, which would probably be more open to help research. However, we ask how to aggregate metadata obtained from data enrichment/labeling, assuming that this effort turns large-scale.

Finally, an approach of automatically generating metadata in case of need, is performing active measurements in case an anomaly is detected. One example of this approach is setting up a TLS connection to an https enabled web server located in a prefix that may be subject to a hijack [1].

## References


- 1 Schlamp, Johann, et al. “Investigating the nature of routing anomalies: Closing in on subprefix hijacking attacks.” International Workshop on Traffic Monitoring and Analysis. Springer, Cham, 2015.



## 5 Panel discussions

### 5.1 BGP Flowspecs

*Vasileios Kotronis (FORTH – Heraklion, GR)*

License  Creative Commons BY 3.0 Unported license  
© Vasileios Kotronis

Joint work of all seminar participants

We started a discussion on BGP Flowspecs, a trigger of a more general discussion on inter-domain routing security. BGP Flowspecs can be used to provide instructions on traffic redirection/blackholing, as long as it matches certain attribute filters. What would happen if Flowspecs start crossing domain borders, used between ASes for exchange of routing and traffic management instructions (i.e., blackholing/redirection/polishing, on a flow level)? This would potentially open new security holes enabling traffic manipulation, based on the exposure of network control to external entities. This can probably be done also today, but it is much harder. In general, there is no notion of accountability/verification/authorization w.r.t. BGP Flowspecs; some knobs allow to accept but not apply them, defeating their own purpose.

The main issue with such proposals, is that we are trying to overload BGP with all kinds of functionality, while “bolting” security on top. BGP Flowspecs are not an exception; we need to take a long hard look at new vulnerabilities of such proposals.

Moreover, we should also reconsider the BGP “trust” model. In inter-domain routing, “trust” stems from business relationships, at least currently. However, this model is currently disputed. In principle no e.g., BGP updates are/should be trusted without a proper infrastructure in place; however, we do not have a practical way to verify the information that the protocol provides. This is not the same as stating that “we assume trust between the domains”; we simply do not have a better alternative to remain inter-operable with the rest of the Internet. Therefore, current routing is essentially a compromise between mutually distrusting domains. Network operators cannot extend their non-transitive “fragile” trust towards even more BGP features.

BGPsec might help to restore some trust, making sure that at least the received BGP update includes an untampered, trusted path. However, it is not useful in the event of partial deployment; it is an open issue how we can achieve total deployment.

### 5.2 Data Sources

*Vasileios Kotronis (FORTH – Heraklion, GR)*

License  Creative Commons BY 3.0 Unported license  
© Vasileios Kotronis

Joint work of all seminar participants

#### 5.2.1 Existing Control-plane Data Sources

We discussed the following *existing* control-plane (CP) data sources for BGP and inter-domain routing.

- RouteViews [1], RIPE [2], PCH [3]. BGPStream [4] from CAIDA can be used as a software framework and API for accessing the feed of BGP route monitors.



- Looking Glasses. A unified interface (such as Periscope from CAIDA) can be used to access these Looking Glass servers.
- RPKI databases. Can be accessed with mechanisms such as RPKI MIRO [6].
- IRR (RSPL databases [7])
  - Documentation of routing policies.
  - Routing data from the entire global registry may be obtained by entering 'whois' commands such as:
 

```
whois -h whois.radb.net <network_IP>
```

 or:
 

```
whois -h whois.radb.net AS<Autonomous_System_Number>
```
  - One can obtain extensive IRR data through FTP [9] or access it indirectly through the use of free user resources.
  - A list of routing registries can be found at [10].
- BGPmon [8]. Detection of BGP hijacks, route leaks and Internet outages.
- CAIDA AS-level topologies [11].
- PeeringDB [12].
- BMP:
  - Protocol to share control-plane data.
  - In contrast to route monitors, one can acquire data before best path selection takes place.
  - Will be soon supported by BGPstream.
  - Supported in OpenBMP [13]; mostly for "local" ("my AS") analysis.
- Mapping of ASN to names:
  - <http://stat.ripe.net>
  - <http://as-rank.caida.org/>
  - <http://www.cidr-report.org/as2.0/autnums.html>
  - <http://bgp.he.net/irr/as-set/AS-RR-Res>
  - <http://irrexplorer.nlnog.net/search/AS-RR-Res>
- Mapping of IP to ASN:
  - CAIDA's pfx2as [15].
  - Team Cymru [14].

### 5.2.2 Existing Data-plane Data Sources

We discussed the following *existing* data-plane (DP) data sources for BGP and inter-domain routing.

- RIPE Atlas [16].
- PlanetLab [17].
- Ark [18].

### 5.2.3 Ideal Data Sources

We discussed the following *ideal* CP and DP data sources for BGP and inter-domain routing.

- BMP feed from all existing BGP vantage points and route collectors.
- More vantage points.
- Better metadata; for example, the peering policy of the vantage point AS.

- Labelled data about hijacks versus misconfiguration. Ideally, we would also like to have information about policy violations. In practice, we would like to start with ground truth on BGP hijacks and DDoS attacks. Of course, an open challenge here is how we verify the ground truth, as well as keeping track of who owns the data.
- Mechanism to more easily query the RIPE database. For example, “give me traces that go through AMS-IX”. A related work is the one on practical Internet route oracles [19].
- IP-to-AS path mapping. Useful works are MAP-IT [20], *bdrmap* [21] and the work of Kai *et al.* [22]. The work of Nomikos *et al.* can further detect IXP crossings in traceroutes [23].
- Catalog of BGP communities (both for public and non-public ones). A useful database can be found at [24].
- Log of curious observations that people can label. One can simplify this process by applying clustering techniques. An important challenge here is crowd sourcing, i.e., making operators aware of such mechanisms and need for data.
- Information from companies doing BGP “tricks” (e.g., DDoS mitigation, BGP “optimizers”).
- Reliable sibling AS detection.
- Dealing with ASes that cannot be modelled as a single router [25].

#### 5.2.4 Data Collection Challenges

We identified the following challenges related to the collection of data related to BGP and inter-domain routing.

- Accuracy.
- Where to find relevant data?
- Authentication/authorization. What should the policies for inserting data be?
- Incentives:
  - Collect data, which provide benefit to others, accompanied by clearly described metadata.
  - Keep the provided data accurate.
  - How to validate? What needs to be known?
  - Incentives needed both for collectors and consumers of data, such as operators and academics/researchers.
- Regional differences between RIPE and other regions. Amongst other reasons, these are due to different data quality.
- Some of the IRR issues are documented in rfc7682.
- Queriability of traceroute data/VPs (e.g., RIPE Atlas, paths via AS/IXP).
- Data availability over time
- IP-to-AS path conversion
- Coverage of collectors. Need increased to see local events, such as targeted BGP hijacks.
- Dealing with inaccuracy / incompleteness. If sampling is applied, the robustness of the results needs to be evaluated to know inaccuracies.
- Related data to see impact / goal of a BGP hijack. Examples are attacks against DNS, Bitcoin, or TOR (deanonymization), as well as Spam campaigns. The challenging question is: “How to get the data?”
- Defining/finding “weird” (i.e., suspicious) activities. Can we cluster/group related “weird” items/entities?
- Dealing with companies doing DDoS mitigation based on BGP.
- Reliable detection of sibling ASes.
- ASes are not atomar entities.

### 5.2.5 Consumers of Data

We identified the following consumers of data related to BGP and inter-domain routing.

- Network operators.
- Academics / researchers.

#### References

- 1 University of Oregon. “Route Views Project.” <http://www.routeviews.org/routeviews/>.
- 2 RIPE NCC. “Routing Information Service (RIS).” <https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris>.
- 3 Packet Clearing House (PCH). <https://www.pch.net/>.
- 4 CAIDA. “BGPStream: An open-source software framework for live and historical BGP data analysis.” <https://bgpstream.caida.org/>.
- 5 CAIDA. “Periscope Looking Glass API.” <https://www.caida.org/tools/utilities/looking-glass-api/>.
- 6 Andreas Reuter, Matthias Wählisch, Thomas C. Schmidt. “RPKI MIRO: Monitoring and Inspection of RPKI Objects.” In Proc. of ACM SIGCOMM, pp. 107–108, New York:ACM, August 2015.
- 7 Merit. “IRR Internet Routing Registry.” <http://www.irr.net>.
- 8 BGPmon. “BGPStream: a free resource for receiving alerts about hijacks, leaks, and outages in the Border Gateway Protocol.” <https://bgpstream.com/>.
- 9 RADB. “RADB FTP database.” <ftp://ftp.radb.net/radb/dbase>.
- 10 IRR. “List of Routing Registries.” <http://www.irr.net/docs/list.html#RADB>.
- 11 CAIDA. “AS relationships.” <http://www.caida.org/data/as-relationships/>.
- 12 PeeringDB. “Information related to Peering.” <https://www.peeringdb.com/>.
- 13 Cisco. “Open BGP Monitoring Protocol (OpenBMP) Collection Framework.” <https://github.com/OpenBMP/openbmp>.
- 14 Team Cymru. “IP to ASN Mapping.” <http://www.team-cymru.com/IP-ASN-mapping.html>.
- 15 CAIDA. “Routeviews Prefix to AS mappings Dataset (pfx2as) for IPv4 and IPv6.” <https://www.caida.org/data/routing/routeviews-prefix2as.xml>.
- 16 RIPE NCC. “RIPE Atlas.” <https://atlas.ripe.net/>.
- 17 PlanetLab. “An open platform for developing, deploying, and accessing planetary-scale services.” <https://www.planet-lab.org/>.
- 18 CAIDA. “Archipelago (Ark) Measurement Infrastructure.” <http://www.caida.org/projects/ark/>.
- 19 Cunha, Italo, et al. “Sibyl: a practical Internet route oracle.” (2016): 325-344.
- 20 Marder, Alexander, and Jonathan M. Smith. “MAP-IT: Multipass accurate passive inferences from traceroute.” Proceedings of the 2016 Internet Measurement Conference. ACM, 2016.
- 21 Luckie, Matthew, et al. “bdrmap: inference of borders between IP networks.” Proceedings of the 2016 Internet Measurement Conference. ACM, 2016.
- 22 Chen, Kai, et al. “Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users.” Proceedings of the 5th international conference on Emerging networking experiments and technologies. ACM, 2009.
- 23 Nomikos, George, et al. “traIXroute: Detecting IXPs in traceroute paths.” in Passive and Active Measurement (PAM), 2016.
- 24 One Step. “BGP Community Guides.” <https://onestep.net/communities/>.
- 25 Mühlbauer, Wolfgang, et al. “Building an AS-topology model that captures route diversity.” ACM SIGCOMM Computer Communication Review 36.4 (2006): 195-206.

### 5.3 DDoS / Routing Security Attacks

*Adrian Perrig (ETH Zürich, CH)*

License  Creative Commons BY 3.0 Unported license  
© Adrian Perrig

Joint work of all seminar participants

We discussed DDoS attacks, as well as attacks against routing security in general.

We begin with algorithmic complexity attacks against routers; these can be used to exploit the “slow path” of a router, and, using a small number of packets (such as routing updates) crash a router. This attack is quite hard for well-protected IGP setups, where filtering is applied on routing updates on ingress (depending on the source of the update). However, this is a general problem when one has the capability to inject crafted updates in a network (IGP/BGP), e.g., via a compromised router. One possible solution to defend against such attacks is to rate-limit traffic from the data plane towards the slow path (i.e., the control plane of the router), effectively limiting the rate of packets that need to be processed by the router’s CPU. However, this does not solve the problem of carefully crafted packets that aim to overload the router not through a volumetric attack, but through an algorithmic attack.

Another defense approach is to prevent control-plane traffic from being crippled by volumetric DDoS attacks on links carrying BGP traffic. In fact, this defense can be applied in practice by prioritizing control-plane traffic above all others. The latter action requires of course configuration and tuning of some knobs, and we do not know the extent at which this is applied in the wild.

With respect to identifying the “shape” of DDoS attacks, one could employ compromised honeypots. An example is the AMPPot Honeypot, used to detect/measure spoofed amplification attacks.

We also discuss other interesting types of attack in gaming setups; an attacker could force a player lag in e.g., “shoot-and-kill” games, by performing a short-lived amplification attack using spoofing against the opponent’s IP address. Note that the gaming industry is quite large; DDoS can be used to “win” in different ways for financial or other gain (even for retaliation). One possible incentive is user dominance, leading to “gang wars” between private server hosts. This leads to the natural consequence of “DDoS-as-a-service” for gamers? Moreover, one needs to consider attacks aiming to disrupt microtransactions which are heavily used in several online games.

An important attack vector usually employed to launch large-scale DDoS attacks and resurfaces every now and then, is IP spoofing. The question is: “will spoofing be eventually reduced?” According to network operators, today it is stricter than in the past, but this is still a problem. The main counter-argument that some providers put forth is that they want to allow their customers to do anything (including legitimate spoofing). This is rarer for customers with static routing, but providers will not invest effort in filtering by hand for BGP-based customers. So in fact, allowing anything is the practical choice currently, with the repercussions that this is associated with.

## 5.4 Quality of Service (QoS)

*Adrian Perrig (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Adrian Perrig

**Joint work of** all seminar participants

We further discuss Quality of Service (QoS), w.r.t. routing. Fine-grained QoS (e.g., on the flow level) has been proven to be counterproductive, since it is associated with huge complexity to run in the core of the network due to practical constraints. QoS capabilities in the OSPF routing protocol mostly remain unused.

However, QoS is highly related to product management. A provider can sell a new business model to a customer. Currently, we best understand coarse-grained QoS. As an operational practice, the rule of thumb is the application of fine-grained QoS at the edge of the Internet (e.g., rate-limiting flows according to their traffic class), and coarse-grained **DiffServ** in the core. We note that there maybe a new need for QoS in upcoming applications; for scalability though, its delivery should ideally not keep any state in the network (**DiffServ** vs **IntServ**). Moreover, there should be (e.g., business-side) accountability in multi-hop scenarios; for example, if traffic related to an ongoing telesurgery is going through 5 providers, and failure happens, how does one know and resolve it, or know who to blame if it all goes wrong?

In general, QoS is mainly driven by business solutions rather than technical ones. Desired protocol properties, include but are not limited to, the following:

- Express policies in multi-hop scenarios, including business layer.
- Employ stateless protocols/mechanisms (at least in the core network).
- Report reliable data.

QoS itself may depend on physical deployment. For example, real-time translation requires low latency (and thus “close-by” servers). The main challenge that we face in the context of QoS is to come up with an approach that provides enough incentives for a small group to deploy (and then scale it up).

## Participants

- Mai Ben-Adar Bessos  
Bar-Ilan University –  
Ramat Gan, IL
- Nikita Borisov  
University of Illinois –  
Urbana Champaign, US
- Georg Carle  
TU München, DE
- Shinyoung Cho  
Stony Brook University, US
- Ítalo Cunha  
Federal University of Minas  
Gerais-Belo Horizonte, BR
- Marc C. Dacier  
EURECOM –  
Sophia Antipolis, FR
- Phillipa Gill  
University of Massachusetts –  
Amherst, US
- Joel M. Halpern  
Ericsson – Leesburg, US
- Raphael Hiesgen  
HAW – Hamburg, DE
- Carlee Joe-Wong  
Carnegie Mellon University –  
Pittsburgh, US
- Mattijs Jonker  
University of Twente, NL
- Vasileios Kotronis  
FORTH – Heraklion, GR
- Taeho Lee  
ETH Zürich, CH
- Hemi Leibowitz  
Bar-Ilan University –  
Ramat Gan, IL
- Victoria Manfredi  
Wesleyan University –  
Middletown, US
- Marcin Nawrocki  
FU Berlin, DE
- Christos Pappas  
ETH Zürich, CH
- Adrian Perrig  
ETH Zürich, CH
- Alvaro Retana  
Huawei Technologies –  
Santa Clara, US
- Andreas Reuter  
FU Berlin, DE
- Thomas C. Schmidt  
HAW – Hamburg, DE
- Laurent Vanbever  
ETH Zürich, CH
- Pierre-Antoine Vervier  
Symantec Research Labs –  
Sophia Antipolis, FR
- Stefano Vissicchio  
University College London, GB
- Rüdiger Volk  
Deutsche Telekom – Münster, DE
- Matthias Wählisch  
FU Berlin, DE
- Bing Wang  
University of Connecticut –  
Storrs, US



# Database Architectures for Modern Hardware

Edited by

Peter A. Boncz<sup>1</sup>, Goetz Graefe<sup>2</sup>, Bingsheng He<sup>3</sup>, and  
Kai-Uwe Sattler<sup>4</sup>

1 CWI – Amsterdam, NL, [p.boncz@cwi.nl](mailto:p.boncz@cwi.nl)

2 Google – Madison, US, [goetzg@google.com](mailto:goetzg@google.com)

3 National University of Singapore, SG, [he.bingsheng@gmail.com](mailto:he.bingsheng@gmail.com)

4 TU Ilmenau, DE, [kus@tu-ilmenau.de](mailto:kus@tu-ilmenau.de)

---

## Abstract

The requirements of emerging applications on the one hand and the trends in computing hardware and systems on the other hand demand a fundamental rethinking of current data management architectures. Based on the broad consensus that this rethinking requires expertise from different research disciplines, the goal of this seminar was to bring together researchers and practitioners from these areas representing both the software and hardware sides and to foster cross-cutting architectural discussions. The outcome of this seminar was not only an identification of promising hardware technologies and their exploitation in data management systems but also a set of use cases, studies, and experiments for new architectural concepts.

**Seminar** June 17–22, 2018 – <http://www.dagstuhl.de/18251>

**2012 ACM Subject Classification** Information systems → Database management system engines, Computer systems organization → Architectures

**Keywords and phrases** co-processors, computer architecture, database systems, hardware support for databases, non-volatile memory

**Digital Object Identifier** 10.4230/DagRep.8.6.63

## 1 Executive Summary

*Peter A. Boncz (CWI – Amsterdam, NL)*

*Goetz Graefe (Google – Madison, US)*

*Bingsheng He (National University of Singapore, SG)*

*Kai-Uwe Sattler (TU Ilmenau, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Peter A. Boncz, Goetz Graefe, Bingsheng He, and Kai-Uwe Sattler

Over the last years, the social and commercial relevance of efficient data management has led to the development of database systems as foundation of almost all complex software systems. Hence there is a wide acceptance of architectural patterns for database systems which are based on assumptions on classic hardware setups. However, the currently used database concepts and systems are not well prepared to support emerging application domains such as eSciences, Internet of Things or Digital Humanities. From a user’s perspective, flexible domain-specific query languages or at least access interfaces are required, novel data models for these application domains have to be integrated, and consistency guarantees which reduce flexibility and performance should be adaptable according to the requirements. Finally, volume, variety, veracity as well as velocity of data caused by ubiquitous sensors have to be mastered by massive scalability and online processing by providing traditional qualities of



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

Database Architectures for Modern Hardware, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 63–76

Editors: Peter A. Boncz, Goetz Graefe, Bingsheng He, and Kai-Uwe Sattler



Dagstuhl Reports

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



database systems like consistency, isolation and descriptive query languages. At the same time, current and future hardware trends provide new opportunities such as:

- many-core CPUs: Next-generation CPUs will provide hundreds of compute cores already in the commodity range. In order to allow high degrees of parallelism some architectures already provide hardware support for the necessary synchronization, e.g. transactional memory. However, it is not clear yet how to fully utilize these degrees of parallelism and synchronization mechanism for database processing.
- co-processors like GPU and FPGA: Special-purpose computing units such as GPUs and FPGAs allow for parallelism at much higher degrees accelerating compute-intensive tasks significantly. Moreover, heterogeneous hardware designs such as coupled CPU-FPGA and CPU-GPU architectures represent a trend of close integration between classic hardware and emerging hardware. However, such designs require new architectural concepts for data management.
- novel storage technologies like NVRAM and SSD: Even modern in-memory database system solutions rely mostly on block-based media (e.g. SSD and HDD) for ensuring persistence of data. Emerging memory technologies such as non-volatile memory (NVRAM) promise byte-addressable persistence with latencies close to DRAM. Currently, the usage of this technology is discussed for instant failure recovery of databases, but the role of NVRAM in future data management system architectures is still open.
- high-speed networks: Both in scale-up and scale-out scenarios efficient interconnects play a crucial role. Today, high-speed networks based on 10 Gbit/s Ethernet or InfiniBand support already Remote DMA, i.e. direct access to memory of a remote node. However, this requires to deal with distributed systems properties (unreliability, locality) and it is still unclear how database systems can utilize this mechanism.

In order to open up the exemplarily mentioned application domains together with exploiting the potential of future hardware generations it becomes necessary now to fundamentally rethink current database architectures.

One of the main challenges of this rethinking is that it requires expertise from different research disciplines: hardware design, computer architectures, networking, operating systems, distributed systems, software engineering, and database systems.

Thus, the goal of this Dagstuhl Seminar was to bring together researchers and practitioners from these areas representing both the software and hardware sides and therefore different disciplines to foster cross-cutting architectural discussions. In this way, the seminar extended the series of previous Dagstuhl seminars on database systems aspects, such as “Robust Query Processing” (10381, 12321, 17222) as well as “Databases on Future Hardware” (17101).

The seminar was organized into six working groups where the participants discussed opportunities and challenges in order to exploit different features of modern hardware and operating system primitives for data processing:

- Database accelerators: Based on an analysis of use cases for database accelerators from the level of individual operators and algorithms up to the level of complex database tasks, the group discussed ways of exploiting and evaluating accelerator technologies as well as future research directions with respect to hardware acceleration in databases.
- Memory hierarchies: The group discussed design recipes for database nodes with non-trivial memory hierarchies containing not only disk and RAM but also non-volatile memory. Within such a hierarchy different caching strategies are employed: exclusive caching for functionally equivalent levels and inclusive caching for levels with different functionality.
- Remote direct memory access: The group discussed ways of exploiting RDMA in data-intensive applications. Particularly, an interface providing a set of useful abstractions for



network-aware data-intensive processing called DPI was proposed. Similar to MPI, DPI is designed as an interface that can have multiple implementations for different networking technologies to enable the exploitation of RDMA and in-network processing.

- Heterogeneous database architectures: This topic was addressed by two working groups. Both groups discussed a database software architecture that is capable of making use of multiple hardware devices (GPU, TPU, FPGA, ASICs), in addition to the CPU for handling database workloads. The principle goal was an architecture that would never be worse than a state-of-the-art CPU-centered database architecture, but would get significant benefit on those workloads where the heterogeneous devices can exploit their strengths. The first group developed a morsel-driven architecture, where pipelines are broken up into sub-pipelines and adaptive execution strategies are exploited. The second group discussed operating system support and primitives for heterogeneous architectures.
- Machine learning in database systems: The goal of this working group was to investigate the application of machine learning methods for estimating operator selectivities as part of query optimization. Such an approach could overcome the inaccuracies of traditional cost estimation techniques especially for queries comprised of complex predicates and multiple joins.

The progress and outcome of the individual working groups was presented in a daily plenary session, details of the results are given below.

## References

- 1 Gustavo Alonso, Michaela Blott, Jens Teubner: *Databases on Future Hardware* (Dagstuhl Seminar 17101). Dagstuhl Reports 7(3):1–18 (2017)
- 2 Renata Borovica-Gajic, Goetz Graefe, Allison Lee: *Robust Performance in Database Query Processing* (Dagstuhl Seminar 17222). Dagstuhl Reports 7(5):169–180 (2017)
- 3 Goetz Graefe, Wey Guy, Harumi A. Kuno, Glenn N. Paulley: *Robust Query Processing* (Dagstuhl Seminar 12321). Dagstuhl Reports 2(8):1–15 (2012)
- 4 Goetz Graefe, Arnd Christian König, Harumi Anne Kuno, Volker Markl, Kai-Uwe Sattler: *Robust Query Processing*. Dagstuhl Seminar Proceedings 10381, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany 2010

## 2 Table of Contents

### Executive Summary

*Peter A. Boncz, Goetz Graefe, Bingsheng He, and Kai-Uwe Sattler . . . . .* 63

### Working groups

#### Database Accelerators

*Gustavo Alonso, Witold Andrzejewski, Bingsheng He, Holger Fröning, Kai-Uwe Sattler, Bernhard Seeger, Evangelia Sitaridi, Jürgen Teich, and Marcin Zukowski . . . . .* 67

#### Memory Hierarchies

*Philippe Bonnet, Goetz Graefe, Alfons Kemper, Viktor Leis, Justin Levandoski, Stefan Manegold, Danica Porobic, and Caetano Sauer . . . . .* 68

#### Remote Direct Memory Access

*Gustavo Alonso, Carsten Binnig, Ippokratis Pandis, Ken Salem, Jan Skrzypczak, Ryan Stutsman, Tianzheng Wang, and Zeke Wang . . . . .* 69

#### Heterogeneous Database Architectures I

*Peter A. Boncz, Sebastian Breß, Thomas Neumann, and Holger Pirk . . . . .* 70

#### Heterogeneous Database Architectures II

*Thomas Leich, Thilo Pionteck, Gunter Saake, and Olaf Spinczyk . . . . .* 71

#### Machine Learning in Database Systems

*Daniel Lemire, Klaus Meyer-Wegener, Anisoara Nica, and Andrew Pavlo . . . . .* 72

**Open problems . . . . .** 73

**Participants . . . . .** 76

### 3 Working groups

#### 3.1 Database Accelerators

*Gustavo Alonso (ETH Zürich, CH), Witold Andrzejewski (Poznan University of Technology, PL), Bingsheng He (National University of Singapore, SG), Holger Fröning (Universität Heidelberg – Mannheim, DE), Kai-Uwe Sattler (TU Ilmenau, DE), Bernhard Seeger (Universität Marburg, DE), Evangelia Sitaridi (Amazon.com, Inc. – Palo Alto, US), Jürgen Teich (Universität Erlangen-Nürnberg, DE), and Marcin Zukowski (Snowflake Computing Inc. – San Mateo, US)*

**License** © Creative Commons BY 3.0 Unported license

© Gustavo Alonso, Witold Andrzejewski, Bingsheng He, Holger Fröning, Kai-Uwe Sattler, Bernhard Seeger, Evangelia Sitaridi, Jürgen Teich, and Marcin Zukowski

Hardware-based acceleration technologies provide great opportunities for speeding up database processing. GPUs (optionally with Tensor Cores), iGPUs, FPGA, TPUs, intelligent network devices, memory and disks are only some examples of suitable approaches. Based on an analysis of available technologies we discussed in the working group use cases for database accelerators from the level of individual operators and algorithms up to the level of complex tasks such as query planning and optimization. Particularly, we investigated the following three questions: How to exploit tensor cores for query operators, e.g. for joins? How to speed up (by batch processing) scalar functions, parsing/deserialization of strings/CSV/JSON, as well as the transposition of batches of records? How to exploit accelerators for cardinality estimation and query planning? How to abstract the execution on heterogeneous resources with the help of a task dependency model?

There are a few interesting points to note in the discussion. First, it is still quite difficult to obtain the most efficient implementation for a given problem on a target architecture, although the problem has been studied to some extent in the form of paper publication or open source. Second, hardware architectures are evolving, and even the state-of-the-art implementations can become inefficient in future architectures. Third, one of the consequences from the first and second points is that, it is rather challenging and tedious to have a fair and complete benchmark on different implementations for a given problem across different architectures.

As results of the group's discussion we propose a public repository for implementations of database tasks using different accelerator technologies which forms the basis for programming contests and at the same time allows for a performance comparison of different implementations. As a second result we discussed a survey to be prepared that covers the state of the art of implementing fundamental database operations such as joins, aggregations, sorting, and advanced scans for different accelerators, so that the community can be aware of the state-of-the-art work that has been done, and identify the challenges and opportunities for improving the performance of those operations.

Finally, we discussed future research directions with respect to hardware acceleration for database tasks both on premise (intelligent memory and storage controllers, memory filters, as well as gather operations with static and dynamic strides) as well as for cloud environments (intelligent storage, virtualization of accelerators). The work items are to be defined, since the scope spans across many relevant areas. It can be the topic of a future Dagstuhl seminar.

As a side project, we had quite intensive discussions on how to exploit the Tensor cores for different data processing operations beyond deep learning. The recent and rapid development of deep learning systems and applications have driven tremendous efforts in

tensor accelerator units. One example is from NVIDIA’s tensor core and the other example is Google’s TPU. Those tensor hardware units can typically demonstrate superb tensor computation performance. In this study, we show how common database operations can be implemented from those tensor operations. We will implement our proposal on NVIDIA Volta architecture, and demonstrate its performance and tradeoff. We expect that there will be some tradeoff in such mappings since they may be too restrictive in implementing with tensor operations.

## References

- 1 Bingsheng He. *Data Management Systems on Future Hardware: Challenges and Opportunities*. Proc. 33rd Int. Conference on Data Engineering (ICDE), p. 1609, 2017.
- 2 Sebastian Breß, Max Heimes, Norbert Siegmund, Ladjed Bellatreche, and Gunter Saake. *GPU-Accelerated Database Systems: Survey and Open Challenges*, Trans. Large-Scale Data and Knowledge-Centered Systems, Vol. 15, pp. 1–35, 2014.

## 3.2 Memory Hierarchies

*Philippe Bonnet (IT University of Copenhagen, DK), Goetz Graefe (Google – Madison, US), Viktor Leis (TU München, DE), Justin Levandoski (Amazon Web Services – Seattle, US), Alfons Kemper (TU München, DE), Stefan Manegold (CWI – Amsterdam, NL), Danica Porobic (Oracle Labs – Redwood Shores, US), and Caetano Sauer (Tableau – München, DE)*

**License** © Creative Commons BY 3.0 Unported license  
 © Philippe Bonnet, Goetz Graefe, Alfons Kemper, Viktor Leis, Justin Levandoski, Stefan Manegold, Danica Porobic, and Caetano Sauer

One of our goals has been to draft a “recipe” for designing storage and compute nodes with non-trivial memory hierarchies, typically within a cluster. There are many such recipes for a two-level hierarchy of volatile memory and persistent disk storage, e.g., the five-minute rule in its various instantiations. Our particular interest was on non-volatile memory, which is likely to disrupt software and hardware architectures for data management. Hardware latency and bandwidth multiply to an approximation of the optimal page size, at least for hierarchical ordered search trees like b-tree indexes, see [1, 2]. Within a memory hierarchy, functionally equivalent levels (such as SSD and traditional HDD) may employ exclusive caching, but levels with different functionality (such as persistent storage vs volatile memory) should employ inclusive caching. Exclusive caching moves data such as pages between levels, whereas inclusive caching copies data pages. If there are two thresholds in storage reliability, then both require inclusive caching – all other levels can be exclusive. A workload plus a data structure (or storage structure) determine an access pattern. A logical access pattern (e.g., random key lookup and update) maps to a physical access pattern (e.g., log-structured merge forest). Full software control is required at the boundary of volatile memory and persistent storage, which implies that even NVM requires a buffer pool to implement read-ahead and write-ahead logging. When designing a system using a budget (e.g., purchase price, space, power, etc.), one should add or remove components by marginal gain (e.g., transaction processing bandwidth or latency), of course only within the feasible space (e.g., number of DIMM slots). This should work if the design space is convex. For example, our preliminary calculations using the five-minute rule calculations suggest for DRAM over NVM 64B cache lines lingering for 12 seconds and for NVM over SSD pages of 1 or 4KB lingering 30 seconds. While those calculations apply directly to random accesses, e.g., searching a hierarchical index such as a b-tree, they may or may not apply to access patterns that are principally

sequential, e.g., a file scan or (in the context of database query processing) a table scan, a merge sort, or a distribution sort. Note that a hash join spilling to overflow files on temporary storage is, in effect, a distribution sort. Another challenge the group grappled with, but did not resolve, is adding user time, e.g., query latency in a database context, to the five-minute rule calculations. With user time much more expensive than computers (when scaled to a minute or an hour, for example), the retention times recommended by the five-minute rules are likely to be dramatically longer. While this issue might seem straightforward in the context of a database query, it is less so in the context of a file system, a CPU cache, or an archival storage system. This is an opportunity for a future investigation and perhaps publication.

## References

- 1 Rudolf Bayer, Edward M. McCreight: *Organization and Maintenance of Large Ordered Indexes*. SIGFIDET Workshop 1970:107–141
- 2 Rudolf Bayer, Edward M. McCreight: *Organization and Maintenance of Large Ordered Indices*. Acta Inf. 1:173–189 (1972)

## 3.3 Remote Direct Memory Access

*Gustavo Alonso (ETH Zürich, CH), Carsten Binnig (TU Darmstadt, DE), Ippokratis Pandis (Amazon Web Services – Palo Alto, US), Ken Salem (University of Waterloo, CA), Jan Skrzypczak (Zuse Institute Berlin, DE), Ryan Stutsman (University of Utah – Salt Lake City, US), Tianzheng Wang (Simon Fraser University – Burnaby, CA), and Zeke Wang (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license

© Gustavo Alonso, Carsten Binnig, Ippokratis Pandis, Ken Salem, Jan Skrzypczak, Ryan Stutsman, Tianzheng Wang, and Zeke Wang

Traditional distributed database systems have been assigned under the assumption that the network is the bottleneck. With emerging network technologies such as RDMA this assumption no longer holds true: InfiniBand FDR allows a bandwidth close to a one memory channel [1]. Thus, these technologies will have a significant impact on data-intensive applications. For instance, data processing systems such as distributed database systems or analytics engines (Spark, Flink) can exploit these technologies, but doing this on a system by system basis demands repeated reinvention of the wheel. Thus, the question arises how will applications make best use of these network technologies?

InfiniBand supports two network communication stacks: IP over InfiniBand and Remote Direct Memory Access (RDMA). In the working group, particularly RDMA was discussed, which is already seeing significant adoption. RDMA provides a Verbs API which uses the capabilities of RDMA NICs for data transfer. In this way, most of the processing can be executed without OS involvement allowing to achieve low latencies. However, using RDMA is still complex due to missing higher-level abstractions. RDMA connections are implemented using pairs of send/receive queues. For communication, a client has to create a so-called Work Queue Element (WQE), put it into a send queue and inform the local NIC to process the element. Basically, communication and computation on the client can be efficiently overlapped without expensive synchronization. However, this low-level mechanism as well as other aspects such as cache coherence result in a complex programming model. This problem is even worse for emerging technologies, like smart NICs and switches for in-network processing.

After a discussion about APIs and programming models for RDMA as well as about experiences with existing techniques such as MPI, the group decided to propose a new programming interface for RDMA called DPI for Data Processing Interface. The aim of DPI is to provide simple yet powerful abstractions that are flexible enough to enable exploitation of RDMA and in-network processing. Like MPI, DPI is just an interface that can have multiple implementations for different networking technologies. To that end, a concrete DPI implementation can serve as a toolkit for implementing networked data-intensive applications, such as analytics engines or distributed database systems.

## References

- 1 Carsten Binnig, Andrew Crotty, Alex Galakatos, Tim Kraska, and Erfan Zamanian *The End of Slow Networks: It's Time for a Redesign*. PVLDB 9(7): 528–539, 2016.

## 3.4 Heterogeneous Database Architectures I

*Peter A. Boncz (CWI – Amsterdam, NL), Sebastian Breß (DFKI Berlin, DE), Thomas Neumann (TU München, DE), Holger Pirk (MIT – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Peter A. Boncz, Sebastian Breß, Thomas Neumann, and Holger Pirk

The working group asked the research question: what database software architecture would be capable of making use of multiple hardware devices (GPU, TPU, FPGA, ASICs), in addition to the CPU for handling data management workloads. The principle goal of this would be an architecture that would never be worse than a state-of-the-art CPU-centered database architecture, yet would get significant benefit on those workloads where the heterogeneous devices can exploit their strengths. The working group took a practical approach in tasking itself with actually designing and implementing such an architecture, in order to be confronted with the fundamental issues that arise when trying to combine heterogeneous hardware. The realized design builds on “morsel-driven” parallelism as it is known in CPU-centric database systems [1]. It focused on integrating CPU and GPU in a system that executes a particular just-in-time compiled query (a generic scan-select-join-aggregation task) across both devices. This work highlighted a number of open issues:

- how to deal with stateful data structures, such as hash-tables or indexes, given the fact that these must be spread over multiple device memories.
- how to do stateful operator pipeline scheduling, that e.g. take data locality into account.
- how to devise multiple compatible implementations of these query pipelines, and how to decide which to schedule when.
- how to deal with hardware-specific constraints and consequences of execution choices, e.g. possibly adverse down-clocking events due to concurrent usage of the devices.
- how to exploit hardware synchronization & communication features like unified memory and NVLink, when they are available, but still support devices on which these features are not implemented?

The working group developed a morsel-driven architecture, where pipelines can be broken up into sub-pipelines using the concept of “lolemps” (introduced long ago in IBM Starburst [2]) and adaptive execution strategies. This architecture is to be described in a vision paper and supported by experiments based on the code repository started in the Dagstuhl workshop.

## References

- 1 Viktor Leis, Peter A. Boncz, Alfons Kemper, Thomas Neumann: *Morsel-driven parallelism: a NUMA-aware query evaluation framework for the many-core age*. SIGMOD Conference 2014: 743–754
- 2 L.M. Haas, J.C. Freytag, G.M. Lohman, H. Pirahesh, *Extensible Query Processing in Starburst.*, Proceedings of ACM SIGMOD, Portland, Oregon, 1989.

## 3.5 Heterogeneous Database Architectures II

Thomas Leich (HS Harz – Wernigerode, DE), Thilo Pionteck (Universität Magdeburg, DE), Gunter Saake (Universität Magdeburg, DE), and Olaf Spinczyk (Uni Osnabrück, DE)

License © Creative Commons BY 3.0 Unported license  
© Thomas Leich, Thilo Pionteck, Gunter Saake, and Olaf Spinczyk

The original working group on Heterogeneous Database Architectures split after an intensive discussion about whether a comprehensive survey on heterogeneous database architectures is feasible during the Dagstuhl seminar or not. Whereas the first subgroup was engaged in a prototypical implementation on heterogeneous platforms (see above), the second subgroup focussed on the development of an abstract common framework for implementing parallel query processing in a heterogeneous hardware scenario. The members of the working group identified the following requirements for such an abstract processing framework:

- Hardware resources should be fairly assigned to isolated concurrent applications. Database processing is only one of these concurrent applications sharing the same hardware resources.
- There should be an abstraction from individual resource types without losing the ability to exploit a computing resource’s specific strengths.
- As a result, the basic building blocks of a parallel query processing are implemented by a pool of differently coded query processing operations.
- The abstracted hardware devices are modelled as containers which are elastic in nature, i.e., their capabilities and performance characteristics may change during runtime (because of resource needs of other concurrent application).

Aim of this work group was to develop a universal system architecture for integrating heterogeneous computing resources such as CPUs, GPUs and FPGAs into a database management system. Key challenges for such a system architecture are the different execution models of the underlying hardware, the exploration of the intra and inter-device parallelism and the system complexity (see, for example, the discussion in [1] for FPGAs). Therefore abstraction and encapsulation were identified as key design guidelines. After extensive discussion, the work group proposed a layered system architecture, consisting of three layers: resource partitioning (layer 0), task-based runtime system (layer 1) and data processing (layer 2). Layer 0 is responsible for the global resource management functions, such as partitioning resources for a number of concurrent queries or system-wide power management. Each concurrent query as well as global system software services run within isolated resource containers called “cells”. Cells <sup>1</sup> may be elastic in nature, as layer 0 might decide to add or withdraw computing or memory resources at runtime, e.g. when a new query starts.

<sup>1</sup> The Cell model has been inspired by the Tesselation manycore OS [2].



Layer 1 is a task-based runtime system, which is executed within each cell. It is responsible for exploiting the available resources in the most efficient way at any time. The provided API is intended to not only support data-intensive application cells but also arbitrary other applications that aim to exploit heterogeneous computing resources. Layer 2 provides generic reusable abstractions that are specialized for data processing. Its main purpose is to map the structure of data processing operation graphs to the task-based execution model provided by Layer 1.

## References

- 1 Andreas Becher, Lekshmi B.G., David Briones, Tobias Drewes, Bala Gurumurthy, Klaus Meyer-Wegener, Thilo Pionteck, Gunter Saake, Jürgen Teich, and Stefan Wildermann. *Integration of FPGAs in Database Management Systems: Challenges and Opportunities*. Datenbank-Spektrum, August 2018.
- 2 Juan A. Colmenares, Gage Eads, Steven Hofmeyr, Sarah Bird, Miquel Moretó, David Chou, Brian Gluzman, Eric Roman, Davide B. Bartolini, Nitesh Mor, Krste Asanović, and John D. Kubiatowicz. 2013. Tessellation: refactoring the OS around explicit resource containers with continuous adaptation. In Proceedings of the 50th Annual Design Automation Conference (DAC '13). ACM, New York, NY, USA, 2013.

## 3.6 Machine Learning in Database Systems

*Daniel Lemire (University of Québec – Montreal, CA), Klaus Meyer-Wegener (Universität Erlangen-Nürnberg, DE), Anisoara Nica (SAP SE – Waterloo, CA), and Andrew Pavlo (Carnegie Mellon University – Pittsburgh, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Daniel Lemire, Klaus Meyer-Wegener, Anisoara Nica, and Andrew Pavlo

The working group was interested in applying and exploiting Machine Learning techniques, particularly supported by specialized hardware processing units such as Tensor cores, for performance-critical tasks in database systems. Improving query optimization was identified as a promising area. The goal was to better estimate the logical properties of queries and the characteristics of their physical realizations (e.g., running time, memory).

Query optimization in database management systems (DBMS) relies on physical-cost models that are not always optimally tuned to the specific systems and computational units. An important component of the physical-cost estimation in the query optimizer is selectivity estimation. The traditional approach to computing the estimate of an operator is to use heuristics based on data statistics, such as samples and histograms, which the DBMS derives from the underlying base table. This approach, however, may be inaccurate, especially for queries comprised of complex predicates and multiple joins. This approach, however, may be inaccurate, especially for queries comprised of complex predicates and multiple joins. This is because one has to make assumptions about the data distributions and correlations, which are non-trivial to ascertain.

To address this problem, the group proposed and investigated approaches for two problems in query optimization:

- A first approach of using machine learning methods was proposed for estimating operator selectivities. For this purpose, conjuncts are encoded as a feature vector that captures the predicate expressions and their actual selectivities. To evaluate this approach, a single-layer quadratic regression model was trained from a sample corpus of 80,000 two-



predicate conjunction queries on the TPC-H database. The initial results show that this model allows to estimate selectivities with a mean absolute error (MAE) of 18%.

- In addition to predicting logical properties, ML models can also be directly used to predict runtimes and resource consumption given the logical properties of the data. To investigate this second possibility, a quadratic regression model was applied to the algorithm that computes the union of two non-uniform random arrays on both a standard Intel server and on an AMD server with an ARM processor. Results show that allows to predict the runtime with a relative MAE of less than 15%, using ML models trained on a specific hardware.

The group members plan to explore these problems further by investigating how to handle more complex query expressions for non-uniform and real-world data sets, or how to maintain trained models under data changes. Future work should address the case where we have variable numbers of parameters used for the feature vector.

## References

- 1 Victor A. E. de Farias, José G. R. Maia, Flávio R. C. Sousa, Leonardo O. Moreira, Gustavo A. C. Santos, and Javam C. Machado. *A machine learning approach for SQL queries response time estimation in the cloud*. In XXVIII Simpósio Brasileiro de Banco de Dados – Short Papers, Recife, Pernambuco, Brasil, September 30 – October 3, 2013., pages 23:1–23:6, 2013.
- 2 Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. *The case for learned index structures*. In Proc. Int. Conf. on Management of Data (SIGMOD, Houston, TX, USA, June 10-15), pages 489–504, 2018.
- 3 Lin Ma, Dana Van Aken, Ahmed Hefny, Gustavo Mezerhane, Andrew Pavlo, and Geoffrey J. Gordon. *Query-based workload forecasting for self-driving database management systems*. In Proc. Int. Conf. on Management of Data (SIGMOD, Houston, TX, USA, June 10-15), pages 631–645, 2018.
- 4 Ryan Marcus and Olga Papaemmanouil. *Deep reinforcement learning for join order enumeration*. CoRR, abs/1803.00055, 2018.

## 4 Open problems

During the seminar we identified several open problems and challenges which should be addressed in the future to make data management architectures ready for and aware of upcoming hardware technology.

### Open Problem: Specialized Hardware

Will specialized hardware for data processing units make sense? Database machines was a dream in database community, although they have not become popular in the market due to cost and delay of integrating new hardware into a software infrastructure. However, the computing landscape has changed, especially that the cost of building a specialized hardware architecture has dropped dramatically in the past decade. This enabled the recent trend of building specialized hardware architectures for deep learning applications.

- How does the wave of NPU affect the database? Will it be economic now to have specialized hardware designs for database machines?

- What are the right hardware/software interfaces in this specialized hardware?
- What are the workloads suitable for specialized hardware besides deep learning?

### Open Problem: Heterogeneous Hardware

Due to the hardware heterogeneity, database systems become more challenging to build, maintain and debug. The open problem is to investigate portable still efficient database designs on heterogeneous platforms.

- Hardware are becoming more diverse. How to balance portability and efficiency?
- What are the right hardware/software interfaces for database system designs?
- How can a database engine optimized for one platform be portable/auto tuned to another?
- How to exploit hardware synchronization and communication features like unified memory and NVLink when they are available but still support devices on which these features are not implemented.
- What new hardware would be desirable – what do the software people want the hardware people to invent?
- How can novel hardware contribute to new functionality, other than performance?
- How to invent a line of new hardware that ensures a sustainable performance advantage rather than a single-generation advantage?

### Open Problem: Highspeed Networking

Emerging network technologies open new opportunities for distributed data management both for data analytics but also distributed transactional databases. With technologies such as InfiniBand FDR the traditional assumption that the network is a bottleneck no longer holds. However, exploiting these technologies in database system design requires a rethinking of architectural concepts and algorithms:

- What are the right abstractions/communication primitives/interfaces for exploiting for instance RDMA?
- Which role plays remote memory in a memory hierarchy if access to remote data is no longer significantly slower than to local objects?
- Which impact has this to the architecture of distributed databases?
- How can we leverage for instance atomic primitives by RDMA for transactional data processing?
- How should we best leverage modern RDMA network cards, such as the Mellanox Innova-2 and Bluefield, that provide programmable devices (e.g., an FPGA or a many core ARM architecture) to extend the RDMA protocol?

### Open Problem: Memory Hierarchies

Storage and memory play an important role in database systems and the database community has a very good understanding about internal data structures, supporting different access patterns, and the role of the different storage technologies in the overall hierarchy. However, with emerging trends such as non-volatile memory (NVM) and programmable storage new opportunities arise.

- If new memory or storage hardware extends the memory and storage hierarchy, then what are the right policies and mechanisms for data placement and movement in this hierarchy? For example, what are the right “page” sizes, transactional semantics, read-ahead and write-behind, etc.
- Which role plays “byte addressable” NVM in the memory hierarchy of a data management system, also from a economic perspective?
- Which data structures are best suitable for this memory technology or even to cross multiple levels of the hierarchy?
- How can we utilize programmable memory and storage to offload functionalities such as scans or even predicate evaluation?
- The db “community” really only understands 2-level “hierarchies” of disk and memory – what about multiple volatile memory levels and multiple persistent storage levels?
- How do indexing, sorting (merge sort), and hashing (distribution sort) fit into and exploit a memory hierarchy?

## Participants

- Anastasia Ailamaki  
EPFL – Lausanne, CH
- Gustavo Alonso  
ETH Zürich, CH
- Witold Andrzejewski  
Poznan University of  
Technology, PL
- Carsten Binnig  
TU Darmstadt, DE
- Peter A. Boncz  
CWI – Amsterdam, NL
- Philippe Bonnet  
IT University of  
Copenhagen, DK
- Sebastian Breß  
DFKI – Berlin, DE
- Holger Fröning  
Universität Heidelberg –  
Mannheim, DE
- Goetz Graefe  
Google – Madison WI, US
- Bingsheng He  
National University of  
Singapore, SG
- Alfons Kemper  
TU München, DE
- Thomas Leich  
HS Harz – Wernigerode, DE
- Viktor Leis  
TU München, DE
- Daniel Lemire  
University of Québec –  
Montreal, CA
- Justin Levandoski  
Amazon Web Services –  
Seattle, US
- Stefan Manegold  
CWI – Amsterdam, NL
- Klaus Meyer-Wegener  
Universität Erlangen-Nürnberg,  
DE
- Onur Mutlu  
ETH Zürich, CH
- Thomas Neumann  
TU München, DE
- Anisoara Nica  
SAP SE – Waterloo, CA
- Ippokratis Pandis  
Amazon Web Services –  
Palo Alto, US
- Andrew Pavlo  
Carnegie Mellon University –  
Pittsburgh, US
- Thilo Pionteck  
Universität Magdeburg, DE
- Holger Pirk  
MIT – Cambridge, US
- Danica Porobic  
Oracle Labs –  
Redwood Shores, US
- Gunter Saake  
Universität Magdeburg, DE
- Ken Salem  
University of Waterloo, CA
- Kai-Uwe Sattler  
TU Ilmenau, DE
- Caetano Sauer  
Tableau – München, DE
- Bernhard Seeger  
Universität Marburg, DE
- Evangelia Sitaridi  
Amazon.com, Inc. –  
Palo Alto, US
- Jan Skrzypczak  
Zuse Institute Berlin, DE
- Olaf Spinczyk  
TU Dortmund, DE
- Ryan Stutsman  
University of Utah –  
Salt Lake City, US
- Jürgen Teich  
Universität Erlangen-Nürnberg,  
DE
- Tianzheng Wang  
Simon Fraser University –  
Burnaby, CA
- Zeke Wang  
ETH Zürich, CH
- Marcin Zukowski  
Snowflake Computing Inc. –  
San Mateo, US



# Ubiquitous Gaze Sensing and Interaction

Edited by

Lewis Chuang<sup>1</sup>, Andrew Duchowski<sup>2</sup>, Pernilla Qvarfordt<sup>3</sup>, and  
Daniel Weiskopf<sup>4</sup>

1 LMU München, DE, [lewis.chuang@um.ifi.lmu.de](mailto:lewis.chuang@um.ifi.lmu.de)

2 Clemson University, US, [duchowski@clemson.edu](mailto:duchowski@clemson.edu)

3 FX Palo Alto Laboratory, US, [pernilla.qvarfordt@gmail.com](mailto:pernilla.qvarfordt@gmail.com)

4 Universität Stuttgart, DE, [daniel.weiskopf@visus.uni-stuttgart.de](mailto:daniel.weiskopf@visus.uni-stuttgart.de)

---

## Abstract

This report documents the program and outcomes of the three day Dagstuhl Seminar 18252 “Ubiquitous Gaze Sensing and Interaction”. The miniaturization of optical devices and advances in computer vision, as well as a lower cost point, have led to an increased integration of gaze sensing capabilities in computing systems. Eye tracking is no longer restricted to a well controlled laboratory setting, but moving into everyday settings. Therefore, this Dagstuhl Seminar brought together experts in computer graphics, signal processing, visualization, human-computer interaction, data analytics, pattern analysis and classification along with researchers who employ eye tracking across a diverse set of disciplines: geo-information systems, medicine, aviation, psychology, and neuroscience, to explore future applications and to identify requirements for reliable gaze sensing technology. This fostered a dialog and allowed: (1) computing scientists to understand the problems that are faced in recording and interpreting gaze data; (2) gaze researchers to consider how modern computing techniques could potentially advance their research. Other issues concerning the ubiquitous deployment of gaze sensing and interaction were also discussed, such ethical and privacy concerns when deploying gaze monitoring devices in everyday settings.

**Seminar** June 18–21, 2018 – <http://www.dagstuhl.de/18252>

**2012 ACM Subject Classification** Human-centered computing → Ubiquitous and mobile computing, Human-centered computing → User models, Human-centered computing → Visualization, Applied computing → Law, social and behavioral sciences, Computing methodologies → Computer vision, Theory of computation → Pattern matching

**Keywords and phrases** eye tracking, computer vision, pattern analysis, ubiquitous computing, user modeling

**Digital Object Identifier** 10.4230/DagRep.8.6.77

**Edited in cooperation with** Tanja Blascheck



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

Ubiquitous Gaze Sensing and Interaction, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 77–148

Editors: Lewis Chuang, Andrew Duchowski, Pernilla Qvarfordt, and Daniel Weiskopf



Dagstuhl Reports

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

## 1 Executive Summary


*Tanja Blascheck (INRIA Saclay, FR)*

*Lewis Chuang (LMU München, DE)*

*Andrew Duchowski (Clemson University, US)*

*Pernilla Qvarfordt (FX Palo Alto Laboratory, US)*

*Daniel Weiskopf (Universität Stuttgart, DE)*

**License**  Creative Commons BY 3.0 Unported license

© Tanja Blascheck, Lewis Chuang, Andrew Duchowski, Pernilla Qvarfordt, and Daniel Weiskopf

The miniaturization of optical devices and advances in computer vision, as well as a lower cost point, have led to an increased integration of gaze sensing capabilities in computing systems, from desktop computing to mobile devices and wearables. With these advances in technology, new application areas for gaze sensing are emerging. Eye tracking is no longer restricted to a well-controlled laboratory setting, but moving into everyday settings. When technology makes forays into new environments, there are many questions to be resolved and challenges to be met, from computational to applications and interaction. Ubiquitous gaze sensing and interaction require a framework that can accommodate compatible solutions from data acquisition to signal processing to pattern classification and computer vision to visualization and analytics. Including gaze data into interactive applications requires knowledge of natural gaze behaviors as well as how gaze is coordinate with other modalities and actions.

Therefore, this Dagstuhl Seminar brought together computer scientists and gaze researchers to explore future ubiquitous applications and to identify requirements for reliable gaze sensing technology. Ubiquitous gaze sensing and interaction cannot be achieved by research discipline, but require knowledge and scientific advancement in multiple fields. And, of utmost importance is that researchers from different disciplines meet, interact, and address their common challenges. For this reason, experts in computer graphics, signal processing, visualization, human-computer interaction, data analytics, pattern analysis and classification along with researchers who employ gaze tracking across diverse disciplines attended: geo-information systems, medicine, aviation, psychology, neuroscience, etc. This fostered a dialogue and allowed: (1) computing scientists to understand the problems that are faced in recording and interpreting gaze data, (2) gaze researchers to consider how modern computing techniques could potentially advance their research. In addition, we discussed the ethical and privacy concerns of deploying gaze monitoring devices in everyday scenarios.

The workshop was organized to identify identifying possible **scenarios** and pinpointing the associated **challenges** of developing and deploying ubiquitous gaze sensing during the first day. Challenges identified by multiple scenarios, or the ones that were considered to be significant were the focus of in-depth cross-disciplinary groups. These challenges were discussed on the second day. In three sessions taking place during the day, five challenges were debated. “Data Privacy” and “Gaze + X” were two of the most important topics and received multiple dedicated sessions of discussion due to the high interest of the participants.

On the third day the Dagstuhl Seminar finally discussed future work and how to get the research community engaged in researching the various interesting topics covered. Some of the suggestions were to organize workshops at conferences and organizing a special issue focused on ubiquitous gaze sensing. Several of the discussion groups started brainstorming on papers covering the important topics raised at the workshop.

## 2 Table of Contents

### Executive Summary

<i>Tanja Blascheck, Lewis Chuang, Andrew Duchowski, Pernilla Qvarfordt, and Daniel Weiskopf</i> . . . . .	78
---	----

### Scenarios

Everyday Use of Wearable Gaze Trackers <i>Andreas Bulling, Lewis Chuang, Kenneth Holmqvist, Radu Jianu, David P. Luebke, Diako Mardanbegi, Thies Pfeiffer, and Laura Trutoiu</i> . . . . .	82
Going Places <i>Amy Alberts, Hans-Joachim Bieg, Tanja Blascheck, Sara Irina Fabrikant, Enkelejda Kasneci, Peter Kiefer, Michael Raschke, Martin Raubal, and Daniel Weiskopf</i> . . . . .	83
Healthcare <i>M. Stella Atkins, Roman Bednarik, Leslie Blaha, Nina Gehrer, and Eakta Jain</i> . . . . .	84
Playing & Learning <i>Maria Bielikova, Andrew Duchowski, Hans Gellersen, Krzysztof Krejtz, Kuno Kurzhals, Radoslaw Mantiuk, and Pernilla Qvarfordt</i> . . . . .	86

### Challenges

Gaze + X <i>Amy Alberts, M. Stella Atkins, Hans-Joachim Bieg, Leslie Blaha, Lewis Chuang, Andrew Duchowski, Nina Gehrer, Hans Gellersen, Kenneth Holmqvist, Eakta Jain, Radu Jianu, Krzysztof Krejtz, David P. Luebke, Radoslaw Mantiuk, Thies Pfeiffer, Pernilla Qvarfordt, Martin Raubal, and Laura Trutoiu</i> . . . . .	87
Intent and Prediction <i>Amy Alberts, M. Stella Atkins, Roman Bednarik, Andreas Bulling, Andrew Duchowski, Sara Irina Fabrikant, Nina Gehrer, Eakta Jain, Peter Kiefer, and Daniel Weiskopf</i> . . . . .	88
Novel Interaction Paradigms <i>Amy Alberts, Hans Gellersen, Kenneth Holmqvist, Krzysztof Krejtz, David P. Luebke, Diako Mardanbegi, and Laura Trutoiu</i> . . . . .	88
Data Privacy <i>Roman Bednarik, Maria Bielikova, Tanja Blascheck, Andreas Bulling, Sara Irina Fabrikant, Eakta Jain, Peter Kiefer, Kuno Kurzhals, David P. Luebke, Diako Mardanbegi, Michael Raschke, and Daniel Weiskopf</i> . . . . .	89
Ubiquitous Gaze-based Guidance and Recommendation Systems <i>Maria Bielikova, Leslie Blaha, Tanja Blascheck, Radu Jianu, Kuno Kurzhals, Thies Pfeiffer, Michael Raschke, and Martin Raubal</i> . . . . .	89

### Open problems

Toward a Ubiquitous Gaze-based Interaction Model <i>Amy Alberts</i> . . . . .	90
Eyegaze Tracking in Medicine <i>M. Stella Atkins</i> . . . . .	91



Gaze-based Attention and Intention Recognition: Potentials and Challenges	
<i>Roman Bednarik</i> . . . . .	93
Gaze Sensing in (Automated) Vehicles	
<i>Hans-Joachim Bieg</i> . . . . .	95
Utilizing Eye Tracking Data for User Modeling in Personalized Recommendation	
<i>Maria Bielikova</i> . . . . .	98
Mixed-initiative Sensemaking Enabled by Ubiquitous Gaze Sensing and Interaction	
<i>Leslie Blaha</i> . . . . .	100
Pervasive Eye Tracking and Visual Analytics	
<i>Tanja Blascheck</i> . . . . .	103
Inferring the Deployment of Limited Attentional Resources	
<i>Lewis Chuang</i> . . . . .	106
Short-term Gaze-based User Intent	
<i>Andrew Duchowski</i> . . . . .	109
The Potential of Gaze-based Training in Psychotherapy	
<i>Nina Gehrler</i> . . . . .	111
Eye Movement as Design Material	
<i>Hans Gellersen</i> . . . . .	113
Why USGI Needs Better Eye Trackers	
<i>Kenneth Holmqvist</i> . . . . .	115
Who Watches the Watchmen: Eye Tracking in XR	
<i>Eakta Jain</i> . . . . .	117
Gaze-driven Education: Sensing, Understanding, Intervention, and Adaption	
<i>Radu Jianu</i> . . . . .	120
From Lab to the Real World: Eye Tracking Grows Up	
<i>Enkelejda Kasneci and Michael Raschke</i> . . . . .	122
Challenges in Gaze-based Intention Recognition	
<i>Peter Kiefer</i> . . . . .	125
Gaze Language: A New Channel of Communication in Augmented Reality	
<i>Krzysztof Krejtz</i> . . . . .	127
Communicating Visualization with Gaze-guided Storytelling	
<i>Kuno Kurzhals</i> . . . . .	130
Gaze as a Service for Ubiquitous Gaze Sensing and Augmented Reality	
<i>David P. Luebke</i> . . . . .	132
Basic Explicit Gaze-based Interaction Techniques in VR/MR	
<i>Diako Mardanbegi</i> . . . . .	134
Don't Make Me Click: Immersive Information Spaces at a Glance	
<i>Thies Pfeiffer</i> . . . . .	136
Envisioning Gaze-informed Interaction	
<i>Pernilla Quarfordt</i> . . . . .	139
Detecting Mindless Gaze	
<i>Martin Raubal</i> . . . . .	143



Challenges and Opportunities of Gaze Sensing in Pervasive Visual Analytics  
*Daniel Weiskopf*. . . . . 145

Conclusion and Outlook . . . . . 147

Participants . . . . . 148

### 3 Scenarios

All scenarios were grounded on the assumption that gaze sensing technology (e.g., eye tracking) were available and working reliably everywhere. The workshop participants brainstormed a large number of scenarios where ubiquitous gaze sensing could be used for the study of human-behavior or to enhance and enrich interaction with computing systems. The participants pitched scenarios to each other in a speed-dating pitch. The different scenarios were consolidated and voted on to extract scenarios that well exemplified ubiquitous gaze sensing in action. In the end, the workshop attendees selected four scenarios: Going Places, Healthcare, Work & Play, and Everyday Use of Wearable Gaze Trackers, to flesh out opportunities and challenges for realizing the scenarios.

Each scenario was discussed in smaller groups, with a focus on describing the scenario in a future setting, identifying relevant research questions, and determining assumptions on technology advancement within the scenarios.

#### 3.1 Everyday Use of Wearable Gaze Trackers

*Andreas Bulling (MPI für Informatik – Saarbrücken, DE), Lewis Chuang (LMU München, DE), Kenneth Holmqvist (Universität Regensburg, DE), Radu Jianu (City – University of London, GB), David P. Luebke (NVIDIA – Charlottesville, US), Diako Mardanbegi (Lancaster University, GB), Thies Pfeiffer (Universität Bielefeld, DE), and Laura Trutoiu (Magic Leap – Seattle, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Andreas Bulling, Lewis Chuang, Kenneth Holmqvist, Radu Jianu, David P. Luebke, Diako Mardanbegi, Thies Pfeiffer, and Laura Trutoiu

**Motivation:** Wearable gaze tracking is on the verge of being adopted by the average consumer, but to be fully adopted in everyday use it needs to enable unique and desired services. When working on this scenario, we discussed: (1) what such services might be, (2) how such services could be implemented with a mobile computing system that has access to the user's gaze, and (3) the technical requirements of such an envisioned system.

**Scenarios:** We envisioned a daily scenario from waking up to going to bed whereby gaze tracking could facilitate personalized computing services.

- *Gaze interactive media:* the user might be presented with news stories on a smart mirror that scrolls according to the user's gaze.
- *Breakfast preparation:* the user would be prompted if it is noted that the user has missed an ingredient.
- *Search suggestions:* when a user displays visual search behavior prior to leaving the house, the system could suggest potential search locations.
- *Public displays:* when gazing at a remote public display (e.g., arrival times of the bus), the relevant information could be delivered directly to one's personal display.
- *Shopping:* when in a shopping mall, a personal computing system could identify what one is interested in buying to make store recommendations.
- *Task scheduling:* a personal computing system could prompt simple tasks upon noticing, from the user's gaze, that the user is available to perform them, such as checking emails while waiting in a queue.
- *Adaptive environment:* ambient lighting, e.g., blinds could be adjusted in accordance to pupil dilation.

- *Social interactions*: gaze paired with face recognition could assist the user in recognizing a familiar acquaintance as well as provide additional information, e.g., name of spouse.
- *Journaling*: a recollection of one's daily events and interactions could be presented at the end of the day to trigger the user's memories during journaling.

**Research Questions:** What are the everyday functions that wearable gaze tracking could serve? How is gaze sensing and analysis location- and context-dependent? How do we integrate analytics from gaze sensing with other services and computing systems? What implications for does ubiquitous gaze sensing have for data privacy and security?

**Assumptions:** To realize the scenarios, we identified the following challenges:

- *Computer vision*: context-dependent applications will depend heavily on computer vision, i.e., object and scene recognition.
- *Reliability*: a personal computing system will have to be cognizant of the precision of current gaze estimates, given environmental luminance and other related factors, prior to making a recommendation.
- *User acceptance*: there will be concerns related to privacy, utility, as well as form factors.
- *Form factors*: gaze tracking should be lightweight, non-obtrusive, and does not obscure field-of-view.
- *Power consumption*: the device should not require more than one charge per day.
- *Multi-modal interaction*: gaze input should be coupled with other inputs to ensure robust inferences.
- *Centralized/distributed computing*: there will be a need for a computing infrastructure that allows for secure interaction between one's personal computing device and others.

## 3.2 Going Places

Amy Alberts (Tableau Software – Seattle, US), Hans-Joachim Bieg (Robert Bosch GmbH – Stuttgart, DE), Tanja Blascheck (INRIA Saclay, FR), Sara Irina Fabrikant (Universität Zürich, CH), Enkelejda Kasneci (Universität Tübingen, DE), Peter Kiefer (ETH Zürich, CH), Michael Raschke (Blickshift GmbH – Stuttgart, DE), Martin Raubal (ETH Zürich, CH), and Daniel Weiskopf (Universität Stuttgart, DE)

**License** © Creative Commons BY 3.0 Unported license

© Amy Alberts, Hans-Joachim Bieg, Tanja Blascheck, Sara Irina Fabrikant, Enkelejda Kasneci, Peter Kiefer, Michael Raschke, Martin Raubal, and Daniel Weiskopf

**Motivation:** Ubiquitous gaze sensing and interaction will have a major impact on future mobility. Eye tracking devices will enable pedestrians, cyclists, car drivers, etc. to enhance their skills through training, for localization, or performance improvement, e.g., based on where a person is looking additional information could be depicted. In addition, collected data from a crowd of people can help shape future cities by integrating gaze information while planning urban projects.

**Scenarios:** The following scenarios were discussed to illustrate ubiquitous gaze sensing for going places:

- Usage of ubiquitous gaze sensing when the car is the main means of transportation: training, spatial cognition, self localization / memory, performance improvement (e.g., Formula 1, Uber, taxis).
- Usage of ubiquitous gaze sensing to ensure or enhance safety, e.g., monitoring/vigilance (sleeping driver), health, advertisement, elderly.

- Usage of ubiquitous gaze sensing for urban planning, e.g., diagnostics, managing traffic.
- Usage of ubiquitous gaze sensing with autonomous cars: e.g., using gaze as interaction; looking outside and the car knows what you are looking at (restaurant).

**Research Questions:** How can eye tracking assist traffic participants (e.g., pedestrians, cyclists, car drivers) in the future?

**Assumptions:** The following assumptions are made that have to be fulfilled for this scenario:

- Robust gaze tracking in the car and while cycling.
- Outdoor conditions do not cause problems (e.g., sunlight, glasses, calibration).
- Problem-free integration of many different sensors (e.g., GSR, EEG, head orientation, vehicle sensors).
- Adaptable to multiple environments (e.g., urban, city, highway, forest).

### 3.3 Healthcare

*M. Stella Atkins (Simon Fraser University – Burnaby, CA), Roman Bednarik (University of Eastern Finland – Joensuu, FI), Leslie Blaha (Pacific Northwest National Lab. – Richland, US), Nina Gehrer (Universität Tübingen, DE), and Eakta Jain (University of Florida – Gainesville, US)*

**License** © Creative Commons BY 3.0 Unported license

© M. Stella Atkins, Roman Bednarik, Leslie Blaha, Nina Gehrer, and Eakta Jain

**Motivation:** Ubiquitous gaze sensing and interaction have the potential to transform medical practices in a number of ways. Gaze patterns are known to change depending on physical and mental conditions, hence gaze sensing can provide diagnostic information not available to health professionals today. Beyond diagnostics, healthcare professionals are engaged in a number of different tasks where gaze plays an important role. The medical setting, however, is quite unique so applications need to be specially targeted to be successful.

**Scenarios:** The following scenarios were discussed to illustrate this scenario:

- Passive monitoring of patients enables continuous monitoring and longitudinal data for diagnostics of health status and evaluations of treatment efficacy. Analysis pushed to the sensors provides continuous analysis, not just continuous data collection. Personalized analytics might enable feedback directly to the patients. This could be done in healthcare establishments (e.g., hospitals, nursing facilities) as well as home and work environments.
- Virtual doctors with realistic and expressive gaze behaviors will be available for mental health evaluation and therapy sessions. Generating high-fidelity simulated behavior is important for garnering patient trust and providing effective feedback. Avatars may be customizable to specific populations by providing appropriate affective and conversational cues.
- Ubiquitous gaze sensing of doctors and health providers provides continuous monitoring of performance. This can be used as a data stream for decision support systems. It provides a record of a provider's observations which can be leveraged for second opinions and record keeping. Expert behaviors can be captured and leveraged in case evaluation and in teaching of other providers.
- Teams of healthcare providers are provided new awareness of each other's activities through gaze sensing. For highly coordinated situations, like emergency triage or surgery, gaze data provides information about more of the situation to providers who

need to coordinate care. Gaze-based interactions provide another method of inputting information to a provider system or record of notes, allowing providers to keep their hands on the patients.

- Virtual health collaboration, or leveraging of mixed reality, will become a possibility. Remote expertise might be brought in to assist. Information from the primary surgeon's gaze can be sent to the remote expert. Gaze-based interactions for the remote expert can control the view or cameras, providing needed information.

**Research Questions:** How will ubiquitous gaze sensing and interaction play into future medical domain applications? Within the breakout group, we discussed the medical domain from three different perspectives:

1. How will ubiquitous gaze sensing impact patients?
2. How will ubiquitous gaze sensing impact care givers?
3. How will ubiquitous gaze sensing change medical care practices or training?

**Assumptions:** The following assumptions are made that have to be fulfilled to support pervasive future healthcare applications:

- We will have an established legal framework that addresses privacy, especially compliance with medical privacy regulations (e.g., the Directive on Data Protection in the European Union or the Health Information Portability and Accessibility Act (HIPAA) in the USA).
- We have established the diagnosticity of gaze data for intended medical applications.
- We have established models for gaze metrics related to diagnostic tests and treatments.
- Diagnostics are robust independent of data collection methods (e.g., wearable eye tracker, desktop cameras, cameras integrated into environment or toys/objects).
- When virtual (avatar) healthcare providers are involved, they behave in a believable and trustworthy manner (e.g., make realistic eye movements).
- Settings and variable conditions (e.g., changes in lighting, glasses, calibration) do not include gaze sensing performance.
- Gaze sensing methods are adaptable to different age ranges and health conditions (e.g., children, adults, elderly; mobile and bed-ridden patients).
- Technology is available for home use and clinical use.
- Technology is user-friendly, requiring minimal setup and maintenance from patients, and no calibration.
- Bandwidth, battery power, and data storage issues are solved.
- The speed of analytics is fast enough to move beyond gaze position inferences.
- Multi-scale, multi-resolution eye tracking is possible, and adaptable according to diagnostics needed.

### 3.4 Playing & Learning

*Maria Bielikova (STU – Bratislava, SK), Andrew Duchowski (Clemson University, US), Hans Gellersen (Lancaster University, GB), Krzysztof Krejtz (SWPS University of Social Sciences and Humanities, PL), Kuno Kurzhals (Universität Stuttgart, DE), Radoslaw Mantiuk (West Pomeranian Univ. of Technology – Szczecin, PL), and Pernilla Qvarfordt (FX Palo Alto Laboratory, US)*

**License** © Creative Commons BY 3.0 Unported license

© Maria Bielikova, Andrew Duchowski, Hans Gellersen, Krzysztof Krejtz, Kuno Kurzhals, Radoslaw Mantiuk, and Pernilla Qvarfordt

**Motivation:** Edutainment is an interesting setting in that it allows participants to learn while playing an engaging game. Today these games are limited how they can model users' understanding and level of learning. Gaze has potential in revealing both users' attention and cognitive processes that can be used to improve models of understanding and learnings.

**Scenarios:** Suppose we have a multi-party game, e.g., playing a problem-solving game, where some participants may be playing from a remote location while other play together in the same room. The game could be projected on a shared surface, or represented in VR or traditional displays. The students' gaze is tracked to help communicate with other students or with remote teachers. The system monitors comprehension, gives advice if needed, or calls on the teachers' attention to help the students when they are stuck. The system can detect fatigue, intellectual helplessness, confusion, and tasks adjusted to educational level. It could provide teacher and students with replay with analytics of the learning session so that they can review, discuss, and learn how to improve their performance.

**Research Questions:** How can gaze be used in multi-party scenarios such as (VR) gaming and/or education? When the game is aware of everyone's gaze, how can this be exploited for the benefit of the players in terms of entertainment and learning? When a player's gaze is monitored and visualized, in real-time or as a kind of brief historical scanpath, how could other players or a remote teacher make use of this? How can we model learning from gaze and other modalities?

**Assumptions:** The following assumptions are made that have to be fulfilled for this scenario:

- Recognition and modeling of student/player cognitive state via gaze, actions and visual context, e.g., real-time analysis of comprehension, is solved.
- Real-time detection of mindless gaze as an indication of cognitive fatigue.
- Gaze visualization of multiple people for interpersonal communication.
- Social presence by gaze, may need to learn "gaze language".
- Joint attention is easy to represent/visualize.
- Gaze as additional channel of information is understood.
- Ethical issues non-existent.
- Skill assessment via gaze and eye-hand coordination is understood.
- Learning disability detection (autism, ADHD) is doable.
- Detection of cheating is doable.

## 4 Challenges

Based on the assumptions identified in the scenarios, the workshop set forth to find challenges that cross multiple scenarios. These challenges were selected for the next set of discussions.

### 4.1 Gaze + X

*Amy Alberts (Tableau Software – Seattle, US), M. Stella Atkins (Simon Fraser University – Burnaby, CA), Hans-Joachim Bieg (Robert Bosch GmbH – Stuttgart, DE), Leslie Blaha (Pacific Northwest National Lab. – Richland, US), Lewis Chuang (LMU München, DE), Andrew Duchowski (Clemson University, US), Nina Gehrler (Universität Tübingen, DE), Hans Gellersen (Lancaster University, GB), Kenneth Holmqvist (Universität Regensburg, DE), Eakta Jain (University of Florida – Gainesville, US), Radu Jianu (City – University of London, GB), Krzysztof Krejtz (SWPS University of Social Sciences and Humanities, PL), David P. Luebke (NVIDIA – Charlottesville, US), Radoslaw Mantiuk (West Pomeranian Univ. of Technology – Szczecin, PL), Thies Pfeiffer (Universität Bielefeld, DE), Pernilla Qvarfordt (FX Palo Alto Laboratory, US), Martin Raubal (ETH Zürich, CH), and Laura Trutoiu (Magic Leap – Seattle, US)*

**License** © Creative Commons BY 3.0 Unported license

© Amy Alberts, M. Stella Atkins, Hans-Joachim Bieg, Leslie Blaha, Lewis Chuang, Andrew Duchowski, Nina Gehrler, Hans Gellersen, Kenneth Holmqvist, Eakta Jain, Radu Jianu, Krzysztof Krejtz, David P. Luebke, Radoslaw Mantiuk, Thies Pfeiffer, Pernilla Qvarfordt, Martin Raubal, and Laura Trutoiu

**Research Question:** If we combine gaze with other information, i.e., about context, user’s reaction, actions, and tasks, as well as other sensor information, what extra power in the interpretation of the “gaze + X” do we get?

**Description:** What is the “X”? In our discussion, “X” could be other signals collected in connection with the gaze, such as pupil size, or by other sensors that detect users action, bio-signals, etc.(gaze + sensors). Stepping away from the user, X could also be context defined by the user’s attention. We could, for instance, detect objects, actions, sounds, and speech in the environment and analyze in relation with the gaze (gaze+context). When applying analysis of gaze, it can transform for being an indication of attention at one point in time, to show a fluid behavior. One such example was how a “glance” can be interpreted. Current eye movement classification schemes do not do this very well, yet this information could improve our understanding of user’s understanding and interpretation. Utilizing multiple analytics from the same gaze sensing device could hence been seen as another “X” (gaze + machine learning). When achieving a more complete understanding of gaze in relation to other information sources, we can develop a framework for design interactive system or for creating improved models of human cognition.

## 4.2 Intent and Prediction

*Amy Alberts (Tableau Software – Seattle, US), M. Stella Atkins (Simon Fraser University – Burnaby, CA), Roman Bednarik (University of Eastern Finland – Joensuu, FI), Andreas Bulling (MPI für Informatik – Saarbrücken, DE), Andrew Duchowski (Clemson University, US), Sara Irina Fabrikant (Universität Zürich, CH), Nina Gehrer (Universität Tübingen, DE), Eakta Jain (University of Florida – Gainesville, US), Peter Kiefer (ETH Zürich, CH), and Daniel Weiskopf (Universität Stuttgart, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Amy Alberts, M. Stella Atkins, Roman Bednarik, Andreas Bulling, Andrew Duchowski, Sara Irina Fabrikant, Nina Gehrer, Eakta Jain, Peter Kiefer, and Daniel Weiskopf

**Research Question:** The related questions of intent and prediction explored whether gaze could be used to predict a user’s action. A particularly interesting qualification to these questions was that of temporal scale. That is, how far into the future could we predict a person’s intent, if at all?

**Description:** As an example of a fairly straightforward proof-of-concept in this scenario was the extreme short-term prediction of saccade landing position, which has already been demonstrated to a certain extent. The greater challenge is in the longer timeframe: could we predict the user’s intent on the order of seconds, minutes, hours, or even days? Doing so would require collecting gaze for longer historical periods and clever algorithms for divining intent-based on observed gaze.

## 4.3 Novel Interaction Paradigms

*Amy Alberts (Tableau Software – Seattle, US), Hans Gellersen (Lancaster University, GB), Kenneth Holmqvist (Universität Regensburg, DE), Krzysztof Krejtz (SWPS University of Social Sciences and Humanities, PL), David P. Luebke (NVIDIA – Charlottesville, US), Diako Mardanbegi (Lancaster University, GB), and Laura Trutoiu (Magic Leap – Seattle, US)*

**License** © Creative Commons BY 3.0 Unported license

© Amy Alberts, Hans Gellersen, Kenneth Holmqvist, Krzysztof Krejtz, David P. Luebke, Diako Mardanbegi, and Laura Trutoiu

**Research Question:** For novel interaction paradigms, the interesting question is how do we move beyond basic gaze-based selection and possibly the use of gaze gestures?

**Description:** To a large extent, coming up with novel interaction paradigms depends on the contextual of gaze, e.g., is it in AR or VR, is it looking at a display, or rather is it in the ubiquitous sense where objects have the “power” of detecting gaze (i.e., at them). The latter was a particularly interesting concept termed the Internet of Seeing Things, or IOST. Other scenarios tended to consider head-mounted tracking as in VR or perhaps AR contexts, e.g., how can we use gaze directed at other individuals? Could we also mix in the concept of “Gaze + X” here, as in, when looking at another individual, could gaze direct face recognition modules to identify the other person and then trigger contextual information such as their name, birthday, and other related pieces of information (how many children do they have, if any), etc.



## 4.4 Data Privacy

*Roman Bednarik (University of Eastern Finland – Joensuu, FI), Maria Bielikova (STU – Bratislava, SK), Tanja Blascheck (INRIA Saclay, FR), Andreas Bulling (MPI für Informatik – Saarbrücken, DE), Sara Irina Fabrikant (Universität Zürich, CH), Eakta Jain (University of Florida – Gainesville, US), Peter Kiefer (ETH Zürich, CH), Kuno Kurzhals (Universität Stuttgart, DE), David P. Luebke (NVIDIA – Charlottesville, US), Diako Mardanbegi (Lancaster University, GB), Michael Raschke (Blickshift GmbH – Stuttgart, DE), and Daniel Weiskopf (Universität Stuttgart, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Roman Bednarik, Maria Bielikova, Tanja Blascheck, Andreas Bulling, Sara Irina Fabrikant, Eakta Jain, Peter Kiefer, Kuno Kurzhals, David P. Luebke, Diako Mardanbegi, Michael Raschke, and Daniel Weiskopf

**Research Question:** How can we ensure privacy once gaze sensing becomes pervasive?

**Description:** If people wear gaze sensing technology, we have to ensure their privacy and the privacy of others. Privacy is critical because gaze data can reveal much and highly personal information about the person being tracked, including information about personality and potential medical issues. How can user models obtained from analyzing such data be protected? In addition, we have to educate people how to control the privacy of their gaze data, understand the implications of different levels of privacy protection, and make sure that the underlying models do not have negative implications such as preventing us from looking someplace (e.g., ‘don’t look there’). Furthermore, privacy issues are not restricted to the person wearing a pervasive gaze-sensing device but may include the person’s environment, in particular, other people with whom we are interacting and who might be recorded by the sensing device.

## 4.5 Ubiquitous Gaze-based Guidance and Recommendation Systems

*Maria Bielikova (STU – Bratislava, SK), Leslie Blaha (Pacific Northwest National Lab. – Richland, US), Tanja Blascheck (INRIA Saclay, FR), Radu Jianu (City – University of London, GB), Kuno Kurzhals (Universität Stuttgart, DE), Thies Pfeiffer (Universität Bielefeld, DE), Michael Raschke (Blickshift GmbH – Stuttgart, DE), and Martin Raubal (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license

© Maria Bielikova, Leslie Blaha, Tanja Blascheck, Radu Jianu, Kuno Kurzhals, Thies Pfeiffer, Michael Raschke, and Martin Raubal


**Research Question:** How can gaze be used for ubiquitous gaze-based guidance and recommendation systems?

**Description:** Gaze-based recommendation system use data collected from eye trackers to make recommendations to people. For example, by detecting a persons familiarity or expertise with a new tool or device they can be aided when performing a task. In addition, gaze can be used to detect engagement, activity changes, or context-switching to give the next input while performing a task. These examples also require proper guidance of a person, for example, gaze-guided storytelling techniques. For this, a taxonomy of different scenarios, methods, drawbacks, and benefits as well as the creation of a design space for ubiquitous gaze-based guidance and recommendation systems is required.

## 5 Open problems

### 5.1 Toward a Ubiquitous Gaze-based Interaction Model

*Amy Alberts (Tableau Software – Seattle, US)*

License  Creative Commons BY 3.0 Unported license  
© Amy Alberts

In 2002 a tech article in the New York Times touted the innovation of a small company named FingerWorks. They imagined pieces of glass that could display a keyboard and be controlled by a fingertip. They promised you could “spill your coffee on it” and they keyboard would still work. FingerWorks was eventually bought by Apple and their TouchStream Interaction Model is how we all interact with touch-enabled devices today. The success of the TouchStream model came from solving many human factors, user interaction, and design problems that came with traditional indirection manipulation devices. We are now on the horizon of a new technological breakthrough, that will address the user and design problems we have with touch-enabled devices. Gazed-based user interaction is imminent. However, like touch in 2002, the base interaction model (e.g., select/dismiss, scroll, etc.) for gaze is not known. This paper explores how we might achieve an equally ubiquitous model for gaze-enabled systems of the future.

The most commercially practical application of a gaze-based user interface will be the delivery of information about items in your visual field. Imagine early versions of this where a customer (Alice) walks into a retailer like Nordstrom’s or Whole Foods. Alice is given a pair of glasses to wear when she enters the store. These glasses are computer-vision and gaze-tracking enabled. As she shops around the store, she can see digital indicators of ‘more information’ about items in her field of view. Alice can visually select an item on a shelf (a loaf of artisan bread). The information she sees tells her the price of the loaf, how long it’s been on the shelf, and other infographics about its ingredients and caloric composition. Alice grabs the loaf which dismisses the information she’s seeing and she moves onto the next item in her list.

Alice’s partner is allergic to tree nuts. Alice has the Whole Foods app on her phone in which she indicated this allergy. They are having friends over for dinner and she wants to get a fresh cake for dessert. She approaches the bakery counter and looks at the different cakes. There’s a small indicator drawn over the cakes that she should avoid because those cakes include tree nuts.

Alice is about to check out and she remembers she needs to get some ground coffee. She’s unsure of where coffee is in this store, so she asks (out loud) “where’s the coffee?” She sees an overlay of arrows that indicate the route to the coffee. These arrows adjust and change as she moves through the store.

Core interaction model questions must be addressed to enable the scenarios described here. This interaction model must address human factors considerations that ensure low impact on the human – especially for high volume gestures (e.g., selection). The definition of scenario categories, variables that will affect the reliability of eye tracking technology, and architecture of the software stack will need exploration.

A systematic approach to the gaze-based interaction model should include (but is not limited to) the following

- Comprehensive literature review of the development, consideration, and limitations of existing interaction models (GUI, Touch, Haptics, etc.).

- Identify and develop methodologies to establish human factors, cognitive, and visual system principles that are relevant to a gaze-based interaction model.
- Identify a set of core gestures that must be supported by the gaze-based interaction model (e.g., select/dismiss, scroll, etc.).
- Build acceptance thresholds for successful gaze interaction gestures.
- Build and test a variety of interaction model options to test against the acceptance thresholds.
- Propose a core set of ubiquitous gaze-based gestures.

## 5.2 Eyegaze Tracking in Medicine

*M. Stella Atkins (Simon Fraser University – Burnaby, CA)*

**License** © Creative Commons BY 3.0 Unported license  
© M. Stella Atkins

**Main reference** Benjamin Law, M. Stella Atkins, Arthur E. Kirkpatrick, Alan J. Lomax: “Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment”, in Proc. of the Eye Tracking Research & Application Symposium, ETRA 2004, San Antonio, Texas, USA, March 22-24, 2004, pp. 41–48, ACM, 2004.

**URL** <http://dx.doi.org/10.1145/968363.968370>

Eye trackers are used for medical image perception studying how radiologists make diagnoses in medical images such as CT, MRI, and mammograms. The eye tracking data provides understanding of the visual search process and why errors occur. Eye trackers also are used in surgery, especially for minimally invasive and robot-assisted surgical training where eye-hand coordination is a key factor for good performance. Eye tracking gives insight into how experts differ from novices, and how to improve medical training and monitoring methods. Emerging applications are being developed to integrate eye tracking information towards developing eyegaze-driven decision support systems and to provide gaze contingent control in surgery.

### 5.2.1 Introduction

Developing expertise in radiology and in other clinical visual tasks such as examining a patient to diagnose skin problems, is an important domain where eye tracking can provide valuable information to suggest methods for training and to form effective decision support systems for medical diagnosis. Eye trackers for medical image perception in radiology were pioneered by Drs. Kundel and Nodine [1, 2], where the ultimate aim was to improve diagnosis and reduce the error rates. A recent review details some history and progress [3], concluding that eye tracking can assist in the assessment of expertise, as well as address human errors in visually-based medical decision-making.

Eye tracking research is often performed in the field of minimally invasive surgery (MIS), as MIS is technically much more demanding than open surgery due to the remote interface of the technique with little tactile feedback [4]. MIS training involves practicing simulated surgery tasks such as grasping and reaching objects, using computer simulators in 3D or a physical training box. Eye tracking has revealed differences in the visual behavior between novices and experts performing the same simulated laparoscopic task [5]; experts kept their gaze on the surgical target whereas novices tracked the tool tip. Such knowledge is key to the understanding of how the motor learning process occurs and it elucidates the role of the human visual system on this process. Training also includes novices watching surgery

videos, where eye tracking reveals there is a difference between “watching” and “doing” [6]. Other research addresses gaze during delicate neuro-surgery applications requiring the use of a microscope, to which eye trackers can be attached and used to predict the surgeon’s intent [7].

### 5.2.2 Envisioned Challenges and Solutions

Acquiring quality eyegaze data with experts is a huge challenge, but in 2016, an “Image Perception Lab” was held at the Radiological Society of North America annual conference, whereby the lab invited attendees to volunteer their own time reading data while being eye tracked in an interactive session. This initiative is ongoing, and enables much important data to be collected. Other challenges include developing appropriate models of image perception and intent; at the Dagstuhl seminar it was stimulating to discuss these issues with psychologists and eyegaze practitioners, and consider new psychological models to improve diagnostic performance.

Developing effective visual training for surgery is challenging because of the difficulty of intent prediction. With machine learning we can identify key points in the scene videos for further detailed investigation, and infer cognitive state through eye parameters such as pupil size, and ultimately, how we can use this data to train novices where to look, for improved performance. For gaze contingent control in surgery, we need to take advantage of the surgeons’ 3D vision through eye tracking through microscopes or through binocular vision eg of the Da Vinci robot. Camera control is a problem in many surgeries, as it’s very difficult to synchronize with the surgeon’s movements. Gaze-based camera control is an encouraging approach in robotic surgery [8], which may also be suitable for MIS surgery. As a result of the Dagstuhl seminar, I will be contributing a review section on eyegaze training in medicine, part of a “Gaze-based User Intent” review document.

### References

- 1 Kundel, H. L., Nodine, C. F., Carmody, D.: Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13(3):175–181, 1978.
- 2 Kundel, H. L.: Reader error, object recognition, and visual search. Keynote speech, *Proc SPIE Medical Imaging*, 5372:1–11, 2004.
- 3 Fox, S., Faulkner-Jones, B. E. Eye-Tracking in the Study of Visual Expertise: Methodology and Approaches in Medicine. *Frontline Learning Research*, 5(3):43–54, 2017.
- 4 Gallagher AG, Smith CD, Bowers SP, Seymour NE, Pearson A, McNatt S : Psychomotor skills assessment in practicing surgeons experienced in performing advanced laparoscopic procedures. *J Am Coll Surg*, 197:479–488, 2003.
- 5 Law, B., Atkins, M. S., Kirkpatrick, A. E., Lomax, A. and MacKenzie, C. L. 2004. Eye gaze patterns differentiate skill in a virtual laparoscopic training environment. *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:41–47, 2004.
- 6 Zheng, B., Jiang, X., Bednarik, R., Atkins, M. S. Gaze Characteristics of Video Watching in a Surgical Setting. *Proceedings 2nd Workshop on Eye Tracking and Visualization*, 11-15, 2016.
- 7 Eivazi, S., Hafez, A., Fuhl, W., Afkari, H., Kasneci, E., Lehecka, M., Bednarik, R.: Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta Neurochirurgica*, 159(6):959–966, 2017.
- 8 Fujii, K., Gras, G., Salerno, A., Yang, GZ. Gaze gesture based human robot interaction for laparoscopic surgery. *Medical Image Analysis*, 44:196–214, 2018.

### 5.3 Gaze-based Attention and Intention Recognition: Potentials and Challenges

*Roman Bednarik (University of Eastern Finland – Joensuu, FI)*

**License** © Creative Commons BY 3.0 Unported license

© Roman Bednarik

**Main reference** Roman Bednarik, Hana Vrzakova, Michal Hradis: “What do you want to do next: a novel approach for intent prediction in gaze-based interaction”, in Proc. of the 2012 Symposium on Eye-Tracking Research and Applications, ETRA 2012, Santa Barbara, CA, USA, March 28-30, 2012, pp. 83–90, ACM, 2012.

**URL** <http://dx.doi.org/10.1145/2168556.2168569>

#### 5.3.1 Introduction

Eye gaze is central in social aspects of life such as communication between people. We use gaze both for directing our own attention in interactions with others, and we employ it for social signaling during conversations. Gaze also has a central role in learning from others. For example in early language learning [9], toddlers employ following of speaker’s gaze to obtain cues to resolve ambiguity.

We know that face is a central source of cues to intention recognition [2]. We employ gaze following as a cue to predict the actions of others [6]. Through mapping of the observed action on the internal motor representations of the action, we, effortlessly, initiate motor programs that allow us to direct our gaze to the action in a proactive way.

Finally, as known for instance from competitive games, people can actively jam signals that can be inferred from their own gaze to deceive the opponents and avoid predictability of their intentions.

In communication with interactive systems, one can envision and imagine intelligent architectures capable of similar feats. Such functionality would allow several breakthroughs, in particular, proactive and intelligent interactive systems. Current interfaces still only react to the actions of the users, because miss the predictive capacity that people normally employ in interactions.

Earlier, we designed and implemented a set of studies to systematically verify, whether intention to act can be detected and predicted from gaze [3]. Using a discriminative ML approach at that time, the performance of such system reached about 80% accuracy in an offline mode.

#### 5.3.2 Potentials

There are many instances where a successful prediction of human action would provide tremendous advances. Not only computational agents could be informed by knowing what is the concern of the upcoming action and where it is (i.e., attention and focus prediction), but also what the action related to the object of concern is going to be.

We then will be able to create systems of early warnings, systems capable of predicting and correcting errors, mechanisms for computational resources optimization including foveated displays, proactive agents and assistive technologies.

Some of these technologies have already been introduced. Driver assistance systems are a paramount example. Earlier research focused on employing EEG signals for automotive applications of intention prediction (e.g., [7]), and recently modern computational architectures for early prediction of intention to maneuver have been employed [10]. The future vehicles will benefit from the predictions of the driver’s intentions along with attention to engage assistance systems.

Grasping and reaching for objects is one of the most ubiquitous actions people perform. Prototypes of grasping assistance systems with gaze input both for attention and intention are already being developed [8]; in future similar systems will seamlessly provide support for users in performing both everyday actions, and in specialized critical-domains such as surgical tasks [1] and their training.

### 5.3.3 Challenges and Questions

Challenges are many. Assuming a technical maturity of gaze tracking systems, one of the main challenges related to accurate intent prediction from gaze lies in computation modeling: what representations of gaze are efficacious such that actionable responses can be performed by an artificial intelligence? While there is little doubt that eye-movement data indeed contains intention-related information, currently we have very little understanding what combination of features carries this information, whether and how much these are user and task specific, and what other variables may be at play.

Another challenge that the gaze-based computational intent modeling community will need to address is the fusion of gaze, other signals and contextual information, across multiple time scales. For example, it has been found that the lane-change intention needs to be interpreted in regard with the driving situation [4]. Again in driving scenarios, head-pose have been found as more reliable source of intentions than gaze as an early signal [5]. Therefore, finding the combination of various user-based signals at different epochs preceding the action will be crucial.

Interplay between attention and intention, for example to help disambiguate the target of the upcoming action, will also need to be modeled. This in turn implies reliable computer vision insights into the intention modeling. Ergo, the domain will need to be able to embrace and model multiple sources of data to help in intention detection and prediction.

Aside from the modeling and computational challenges, the very definition of intent differs across studies. Previous works seem to implicitly approach intent as the period before the action, during which the action is planned. I believe that we need a further distinction, mainly in the terms of granularity, to allow for deeper understanding and consequently efficient modeling.

On the way towards creation of these architectures, we will need to establish large community-built datasets with accurate annotations. These would optimally include contextual and other physiological signals. Once available, benchmarking challenges can be organized to advance the research and development.

Solutions to these problems will help in answering the questions related to the trade-off between accurate and timely predictions: How early we can predict an upcoming action (through long- and short-term intention recognition) with a reliable accuracy?

### 5.3.4 Concluding Remarks

We only begin to understand the complexity of human intention forming and its automatic recognition using gaze. When sensing advances will provide robust data in a non-intrusive way, the computational part of the domain need to be ready for the technology. The recent advances in machine learning techniques help to overcome past feature engineering burdens and promise to propel the research of gaze-based intention recognition.

## References

- 1 Atkins, M. S., Tien, G., Khan, R. S., Meneghetti, A., Zheng, B.: What do surgeons see: capturing and synchronizing eye gaze for surgery applications. *Surgical Innovation*, 20(3):241–248, 2013.
- 2 Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I.: The Reading the mind in the eyes test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines* 42(2):241–251, 2001.
- 3 Bednarik, R., Vrzakova, H., Hradis, M.: What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In: *Proceedings of the Symposium on Eye Tracking Research & Applications*. ACM:83–90, 2012.
- 4 Beggiato, M., Pech, T., Leonhardt, V., Lindner, P., Wanielik, G., Bullinger-Hoffmann, A., Krems, J.: Lane change prediction: From driver characteristics, manoeuvre types and glance behaviour to a real-time prediction algorithm. In: *UR: BAN Human Factors in Traffic*, Springer:205–221, 2018.
- 5 Doshi, A., Trivedi, M. M.: On the roles of eye gaze and head dynamics in predicting driver’s intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):453–462, 2009.
- 6 Flanagan, J. R., Johansson, R. S.: Action plans used in action observation. *Nature*, 424(6950):769–771, 2003.
- 7 Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., Blankertz, B.: EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5):1–11, 2011.
- 8 Orlov, P., Shafti, A., Auepanwiriyaikul, C., Songur, N., Faisal, A. A.: A gaze-contingent intention decoding engine for human augmentation. In: *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:91:1–91:3, 2018.
- 9 Tomasello, M., Barton, M. E.: Learning words in nonostensive contexts. *Developmental Psychology*, 30(5):639–650, 1994.
- 10 Zyner, A., Worrall, S., Ward, J., Nebot, E.: Long short term memory for driver intent prediction. In: *IEEE International Conference on Intelligent Vehicles Symposium*, IEEE:1484–1489, 2017.

## 5.4 Gaze Sensing in (Automated) Vehicles

*Hans-Joachim Bieg (Robert Bosch GmbH – Stuttgart, DE)*

License © Creative Commons BY 3.0 Unported license  
© Hans-Joachim Bieg

### 5.4.1 Abstract

Following decades of research, gaze tracking in vehicles is slowly becoming a reality, potentially spearheading a more widespread application of gaze sensing for everyday purposes. The present article summarizes the application of gaze sensing for driver state monitoring and highlights some of the key challenges in this application domain.

### 5.4.2 Driver Inattention Monitoring

Online driver state assessment using eye trackers in cars or commercial vehicles has been a long-standing research topic (reviewed by [2]), primarily motivated by potential safety benefits. *Driver inattention* due to distraction has been identified as an important crash risk factor (e.g., [11], see [3] for discussion of constructs such as inattention or distraction).

Eye tracking, and gaze sensing in particular, is useful in identifying driver distraction. In the most basic case, methods for online estimation of driver distraction compare a driver's gaze to the most general situational requirement in driving: keeping one's eyes on the road, i.e., by assessing the frequency of glances through a region of interest that resembles the location of the road [14].

More advanced approaches, offering the potential for a more fine-grained assessment of driver inattention, are conceivable by considering information about the specific driving situation that may already be available in modern vehicles from on-board sensors. Examples include adaptation to the vehicle's velocity and associated steering demand [13] or location of objects in the vicinity of the car to assess whether the driver pays attention to these objects or not [4].

In contrast to approaches where driver distraction is directly assessed through observation of the driver's gaze behavior, another strand of research focuses on estimating the driver's secondary task [1]. Based on this, compatibility with the driving task can be assessed, e.g., from expert ratings or crash risk estimates for the specific task.

With the surge in research on automated vehicles, video-based driver state assessment in general and gaze tracking in particular is explored to assure that the driver conforms to the vigilance requirements of automated driving systems. For example, in partially automated systems (SAE level 2; [12]), drivers are obliged to monitor the automated system, i.e., deviations from lateral and longitudinal control much like in manual driving. In higher automation levels, drivers are freed from this task, but still need to display appropriate attentional behavior when taking over control from the automated system [10].

### 5.4.3 Challenges and Approaches

Video-based interior sensing systems now slowly make their way into the automobile, primarily driven by the demands of automated driving. Gaze tracking enables a range of driver state assessment methods (see previous section) but at the same time poses challenges in regard to robustness and availability.

For example, precise gaze tracking information may not be available at all or very sparsely for some users, due to vision aids or oculomotor limitations (e.g., strabism). This would lock out users from system functions such as the vehicle automation described above. Eye tracking technology has been used as a formidable interaction aid for users with limited manual motor capabilities. It would be ironic if the same technology would establish a new technological obstacle for other users, as applications incorporating gaze information become more ubiquitous. In addition, with myopia on the rise [7], the refinement of methods for increasing tracking performance when tracking users with glasses or contact lenses is crucial for widespread integration of eye tracking in vehicles and other everyday applications.

Apart from groups of users that may be completely excluded from the benefits of eye tracking technology, various personal or situational factors may lead to temporary performance decrements. Tracking may become disrupted by the rims of glasses, hair, clothing, harsh lighting conditions, or behaviors that lead to difficulties in face tracking or extraction of ocular features (e.g., conversing, squinting while smiling).

These issues may be addressed on the level of system and function design, weighing performance against availability and present opportunities for more comprehensive modeling of user attention: Precision requirements can be relaxed to achieve comparable performance for a large variability of users and conditions. For example, systems may content themselves with coarser estimates of driver visual attention, e.g., from head pose information [8] – at the expense of the system's primary detection performance.



In contrast, graded approaches may complement transiently unavailable precise gaze tracking information with information from coarser but more robust inference sources, such as head pose information. Conversely, sparsely available precise gaze tracking information may enhance the ability to interpret head movement behavior by modeling the individual propensity to perform visual orienting either by head or solely by eye [5].

Finally, knowledge about the driving situation as well as other, multi-modal driver information (e.g., vehicle, in-vehicle information system state, or other interior tracking techniques) may be used to assess the driver's state [9]. Driver gaze information may constitute an important but not the only building block, in a more comprehensive assessment.

## References

- 1 Braunagel, C., Stolzmann, W., Kasneci, E., and Rosenstiel, W. Driver-activity recognition in the context of conditionally autonomous driving. In *IEEE International Conference on Intelligent Transportation Systems*, IEEE:1652–1657, 2015.
- 2 Dong, Y., Hu, Z., Uchimura, K., and Murayama, N. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):596–614, 2011.
- 3 Engström, J., Monk, C. A., and US-EU Driver Distraction and HMI Working Group. A conceptual framework and taxonomy for understanding and categorizing driver inattention, 2013.
- 4 Fletcher, L., Loy, G., Barnes, N., and Zelinsky, A. Correlating driver gaze with the road scene for driver assistance systems. *Robotics and Autonomous Systems*, 52(1):71–84, 2005.
- 5 Fridman, L., Lee, J., Reimer, B., and Victor, T. Owl and lizard: Patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–313, 2016.
- 6 Hansen, D. W., and Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.
- 7 Holden, B. A., Fricke, T. R., Wilson, D. A., Jong, M., Naidoo, K. S., Sankaridurg, P., Wong, T. Y., Naduvilath, T. J., and Resnikoff, S. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*, 123(5):1036–1042, 2016.
- 8 Jha, S., and Busso, C. Analyzing the relationship between head pose and gaze to model driver visual attention. In *IEEE International Conference on Intelligent Transportation Systems*, IEEE:2157–2162, 2016.
- 9 Li, N., and Busso, C. Predicting perceived visual and cognitive distractions of drivers with multimodal features. In *IEEE Transactions on Intelligent Transportation Systems*, 16(1): 51–65, 2015.
- 10 Marberger, C., Mielenz, H., Naujoks, F., Radlmayr, J., Bengler, K., and Wandtner, B. Understanding and applying the concept of “driver availability” in automated driving. In *Advances in Human Aspects of Transportation*, N. A. Stanton, Ed., Springer:595–605, 2015.
- 11 Olson, R. L., Hanowsk, R. J., Hickman, J. S., and Bocanegra, J. Driver distraction in commercial operations. Tech. Rep. FMCSA-RRR-09-042, FMCSA, 2009.
- 12 SAE. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE J3016, 2016.
- 13 Schmitt, F., Bieg, H.-J., Manstetten, D., Herman, M., and Stiefelhagen, R. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *IEEE Intelligent Vehicles Symposium*, IEEE:412–418, 2016.
- 14 Victor, T. W., Harbluk, J. L., and Engström, J. a. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Accident Analysis and Prevention*, 8(2):167–190, 2005.

## 5.5 Utilizing Eye Tracking Data for User Modeling in Personalized Recommendation

*Maria Bielikova (STU – Bratislava, SK)*

**License** © Creative Commons BY 3.0 Unported license

© Maria Bielikova

**Main reference** Maria Bielikova, Martin Konopka, Jakub Simko, Robert Moro, Jozef Tvarozek, Patrik Hlavac, Eduard Kuric: “Eye-tracking en masse: Group user studies, lab infrastructure, and practices,” *Journal of Eye Movement Research*, 11(3):6:1–6:15, 2018.

**URL** <http://dx.doi.org/10.16910/jemr.11.3.6>

### 5.5.1 Abstract

Although a lot of attention has been dedicated towards user modeling for personalized recommendation, user model representations and its exposing in the recommendation algorithms, there is still open space on inputs to the user modeling process. Traditionally just mouse, gestures and keyboard inputs are considered. However, gaze presents more detailed and accurate information on the user current activity. It enables to acquire an instant stream of data on the user perception of items being recommended. Moreover, utilizing eye tracking data enables to acquire other important features such as pupil size or head distance highly relevant for the task of recommendation as predictors of affective and attentional states.

### 5.5.2 Introduction

To enable recommender systems to suggest suitable items for a particular user, the recommender system should know user preferences and goals, or should be able to infer it from the user feedback. Traditionally, a user in digital environment is model-based on his/her explicit or implicit feedback [1]. Explicit feedback on user intents, interests, skills and knowledge is hard to acquire, people often are not willing to answer questions and in many scenarios it is not possible to get it either. Even though explicit feedback once given is oftentimes qualified as reliable, in many real scenarios its reliability may be low, especially in cases when users' input is somehow forced. In such situations the users do not provide accurate responses either due inability to do so or because they do not pay attention to or even they may want to provide a false feedback. So, implicit user feedback is heavily used to complement or just replace the explicit one in many situations.

Current recommenders placed in a digital environment use inputs for user modeling based mainly on an infrastructure used for implementation of the recommender system:

- *web-based applications* in various domains (e.g., e-shops, e-books for education, various web services such as flight booking, museums, healthcare systems) – traditional inputs are page visits, mouse movements, clicks, keyboard typing;
- *applications on smart phones and tablets* for similar domains as above mentioned web-based systems – traditional inputs are screens visited, gestures, taps, typing;
- *applications on smart glasses* for various scenarios of everyday life – traditional inputs are images of environment.

These inputs represent evidence of the user's intent, but present just a little help in understanding “why” the user has acted in particular way and what is his/her opinion. Gaze data can not only strengthen evidence for particular intent deduced, but also disprove various assumptions on user behavior, e.g., an interest based on his/her activity (e.g., clicks). This can markedly improve recommendations as we get more reliable implicit ratings of the items for particular user currently computed mainly based on learning to rank algorithms. Considering collaborative applications gaze data bring even more accurate inputs.

### 5.5.3 Envisioned Challenges

We list scenarios and challenges for utilizing eye tracking data as an input for recommendation. They need various levels of gaze as an input from low level signals through features effective for machine learning to inferred knowledge on the user short term or long term characteristics.

Challenges related to the content of eye tracking data:

- detection of user states useful for recommendation – how eye tracking data can help in recognition of confusion, attention, fatigue?
- user skill assessment – how eye tracking data can help in recognizing familiarity with the application, expertise?
- features for machine learning – which features based on eye tracking data can be used for machine learning tasks such as values estimation, classification, clustering, comparing items, finding similar items.

Challenges related to the recommendation algorithms:

- explanation, i.e., presenting reasoning on recommendation to the user using eye tracking data, making recommenders scrutable – should gaze be explicitly presented to the user? can eye tracking data help in increasing trust of users?

Challenges related to processing of eye tracking data following that gaze produces enormous amounts of:

- massive data processing,
- data storage and filtering,
- real time data sharing.

First steps towards ubiquitous gaze are using eye trackers in collaborative or group scenarios. Such scenarios require special infrastructure [2]. Research in this domain has started primarily in educational domain as personalization including recommendations in intelligent tutoring systems is active research area for many years and multiple eye tracker setups are almost exclusively present in educational environments.

### References

- 1 Jawaheer, G. Weller, P. and Kostkova, P. Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback. *ACM Trans. Interact. Intell. Syst.*, 4(2):8:1–8:26, 2014.
- 2 Bielikova, M. Konopka, M. Simko, J. Moro, R. Tvarozek, J. Hlavac, P. Kuric, E. Eye-tracking en masse: Group user studies, lab infrastructure, and practices. *Journal of Eye Movement Research*, 11(3):6:1–6:15, 2018.

## 5.6 Mixed-initiative Sensemaking Enabled by Ubiquitous Gaze Sensing and Interaction

*Leslie Blaha (Pacific Northwest National Lab. – Richland, US)*

**License** © Creative Commons BY 3.0 Unported license

© Leslie Blaha

**Main reference** Baber, C., Cook, K., Attfield, S., Blaha L. M., Endert, A., and Franklin, L. “A conceptual model for mixed-initiative sensemaking”, 2018 CHI Sensemaking Workshop. 2018.

### 5.6.1 Abstract

We desire computational teammates that can recommend relevant or interesting data sources to support situation awareness, understanding, or decision making. We are currently able to create and transmit more data than people can process and make sense of. Much of this data goes unanalyzed when we lack the human resources to examine it all. Because human analysts have limited capacity for processing data, they must leverage the computational efficiency and data storage capabilities of machines. Machines have the potential to process the large sets intractable for humans to find and recommend relevant information for people, but they need guidance from humans to do so. If teamed, these complementary strengths support useful and timely inferences on large volumes of data, especially in dynamic decision environments where time for sensemaking may be limited. I propose a novel interactive machine learning paradigm directly leveraging gaze information to train machine learning to support sensemaking.

### 5.6.2 Gaze-enabled Emergency Response Scenario

Video and picture footage is streaming into the emergency management operations center as a wildfire threatens an urban area. This footage both contains valuable information of emergency planning and is greater in volume and velocity than a person can reasonably attend to. One emergency response traffic coordinator is tasked with monitoring the evacuation process. She must keep traffic flowing, route evacuees safely away from the fire, and deploy first responders efficiently to address problems early. She approaches the interactive data display equipped with ubiquitous gaze sensing and interaction technology. She focuses her gaze on the map icon, blinks twice, and opens up the current traffic feeds overlaid on a map of the urban area. Glancing at the video feed from one of the front-line teams, she visually pulls the video to the map with a slow saccade. The computer recognizes that she is checking the fire forecast against the traffic flow and adds an overlay of green-yellow-red coloring to indicate traffic delays. Computational models offer cones of uncertainty for the fire movement. A weather forecast is suggested, and she rejects the suggestion with a quick nod. Instead, she focuses on the location of Fire Team 3. The fixation brings up a live video feed from the chief's helmet camera. The computer asks if she needs a communication channel. Nodding while fixating on the video feed, the computer video calls the chief for a verbal report of conditions. Registering keywords in the report, the computer extracts relevant video clips from the other Fire Team chiefs across the responder locations. The compiled report of fire status is automatically sent to the responder logistics coordinator to consider additional truck deployments. Simultaneously, the traffic coordinator is provided analytic results indicating a need for a road to be blocked and traffic re-routed. She saccades and fixates on the police icon at the top of the screen, initiating a call to the local police chief. Fixating back on the map location with the forecast analytics, she transmits the information to the police department to initiate traffic re-rerouting. With a triple blink, she closes the maps and logs the activity in the operations center research.

### 5.6.3 Challenges Toward Mixed-Initiative Sensemaking

Ubiquitous gaze sensing and interaction offers novel approaches to enabling mixed-initiative sensemaking on large volumes of streaming data. In dynamic decision environments like the emergency response operations described above, a single person or even a small team does not have the capacity to process all possible data sources to determine which contain important information to assess the situation. However, we can aid humans with computational tools for mixed-initiative sensemaking [1]. Mixed-initiative sensemaking relies on a combination of human and machine intelligence collaborating to complete complex exploratory analysis, reasoning, problem solving, and decision making tasks. Ubiquitous gaze sensing and interaction captures task-related gaze behavior, providing a rich source of information about the data relevant to each individual. We desire intelligent machine analytics using ubiquitous gaze sensing as a key input for providing recommendations about additional data sources or other analytics to help the operators with their tasks. There are a number of computational challenges in the analytics process to achieving mixed-initiative sensemaking. Note that I am assuming the technical capabilities to collect, transmit, and store ubiquitous gaze and interaction data are addressed separately, and herein emphasize the data modeling and interpretation challenges.

Mixed-initiative sensemaking relies on common ground between the human and computational teammates to effectively align the information needs, goals, and interpretations [1]. Establishing common ground is an ongoing process wherein both the human and machine interpret each other's behavior and resolve conflicts. This requires (near) real-time gaze modeling, which constitutes one of the big data challenges for eye movement analytics [2].

Machines need a way to communicate with people in a manner consistent with the sensemaking process and tasks. Visualization of information is an effective way to present the outputs of large scale analytics, and mixed-initiative systems have the potential to suggest relevant data through the visual analytic tools. Additionally, machines need a way to ensure information is presented in a size and structure which people will be able to effectively visually process, which can be informed by analytics of gaze tracking.

To be adaptive to the analyst, the machine needs a way to understand the goals and tasks of the user, to track switches of tasks, and to predict which data is relevant and informative to the current tasks. This is challenging in complex sensemaking where a person may be switching frequently between subtasks like information foraging and information synthesis. Machines need a way to infer the task a person is doing from the gaze and interaction behaviors and predict the tasks to which a person is likely to switch to make those transmissions smooth. Doing so ensures information can be ready for the user when it is needed without long analysis or query delays.

Machines need a way to extract task-meaningful gaze behaviors that should be leveraged to inform the machine analytics. Importantly, there are two sets of analytics informing the mixed-initiative sensemaking process: (1) gaze behavior analytics supporting the modeling of the user, the task, and the efficacy of the recommendation and information presentation processes, and (2) analytics from the data interactions informing the data analysis and further mixed-initiative recommendation processes. These may leverage the same gaze inputs but require different calculations and computations. The machine will need flexible algorithms to determine which behaviors are needed to inform the gaze analytics and user modeling and which are needed to inform the data stream analytics.

#### 5.6.4 Gaze-based Interactive Machine Learning

I propose a new interactive machine learning paradigm to support large-scale data analysis leveraging ubiquitous gaze sensing and interaction for mixed-initiative sensemaking. Gaze information provides a rich context for indicating what information is of interest to the analyst, which can be leveraged as a labels or input to both machine learning and computational cognitive models, in addition to the other situation-related data sources. I propose that a mixed-initiative system that combines computational cognitive models of the operator with machine learning creates a computational teammate that can bootstrap small amounts of imagery or information viewed by the operator to analyze large volumes of situation data to support the sensemaking process.

As an analyst is viewing images, the cognitive model is tracking and analyzing gaze behavior to understand the information that is of interest to the analyst. The model provides an interpretation of how the information supports different aspects of the analyst's tasks, including memory or reasoning. A cognitive model of the analyst can predict future steps or tasks the analyst is likely to take and what information might be needed in the data to support this task. It can interface with machine learning or other artificial intelligence-based analytics to search the larger amounts of available data for additional relevant information the analyst will need to support the sensemaking process. Through visual analytic representations, the system recommend that relevant information to the user at a time and in a manner that will not disrupt the current tasks but is readily available when needed (as predicted by the model and adaptive to real-time gaze modeling). Old information or information that seems irrelevant due to lack of attention can be adaptively removed from the interface.

The gaze data, both raw and interpreted by the cognitive models, becomes another dimension on which the machine can train, like another label set useful for machine learning. A classifier, for example, might use fixation points as an indicator of image features of interest and classify other features into "of interest" and "not of interest" categories. After recording the operator viewing a small amount of imagery, then, the machine can pull additional "of interest" data for the operator from the sources and streams that the operator might not otherwise have the bandwidth to view. Interaction sequences supporting repeated tasks can be predicted with decision trees, so the whole sequence of information is made available with fewer steps. As an interactive learning process, the system can track changes in gaze-based interactions that reflect changes in the tasks or information needs of the operators, to adaptively support sensemaking, maintaining common ground.

Note that interactive learning with cognitive models integrated into the system means that the computational processes can be adaptive to individual operators, because cognitive models are tailored to individual operators. For a team of individuals then, the same types of gaze behavior measured ubiquitously combined with interactive machine learning can facilitate sharing of information between operators. Results of analytics from one process can be recommended to others performing similar tasks or transferred between people when their tasks are interdependent.

Importantly, continuous gaze sensing and interaction with the visual information enables an unobtrusive way to extract information from the viewer to inform the machine processes. As discussed at the Ubiquitous Gaze Sensing and Interaction Seminar, the advances in technology and algorithms to support such continuous gaze sensing is within reach; decreasing costs make it a potential new technology for emergency operations. We take advantage of natural viewing behaviors and advances in machine learning and analytics to support operators in time-critical decision making.

## References

- 1 Baber, C., Cook, K., Attfield, S., Blaha L. M., Endert, A., and Franklin, L. A conceptual model for mixed-initiative sensemaking. *2018 CHI Sensemaking Workshop*, 2018.
- 2 Blascheck, T., Burch, M., Raschke, M. and Weiskopf, D. Challenges and perspectives in big eye-movement data visual analytics. *Big Data Visual Analytics (BDVA)*, IEEE, 2015.

## 5.7 Pervasive Eye Tracking and Visual Analytics

*Tanja Blascheck (INRIA Saclay, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Tanja Blascheck

**Main reference** Tanja Blascheck, Markus John, Kuno Kurzhals, Steffen Koch, Thomas Ertl: “VA<sup>2</sup>: A Visual Analytics Approach for // Evaluating Visual Analytics Applications”, *IEEE Trans. Vis. Comput. Graph.*, Vol. 22(1), pp. 61–70, 2016.

**URL** <http://dx.doi.org/10.1109/TVCG.2015.2467871>

### 5.7.1 Introduction and Motivation

In this open problem statement, I present my vision of a pervasive use of eye tracking technology. In the future, I imagine that people will use eye tracking technology always and everywhere. This leads to a great potential in improving our daily lives but also comes with many new challenges we have to face and overcome. My vision is that people use eye tracking technology in combination with other sensors (e.g., EEG, GSR, GPS) and devices (e.g., smartwatches, smartphones, augmented reality glasses, other wearables) for a quantified self. For example, we can enhance the mobility of ourselves by monitoring our behavior to ensure, for example, safe travels, while worn devices provide a mechanism to give feedback, for example, on the travel direction based on landmark. Another possible scenario is the improvement of visual data exploration, because we generate more and more data every day, and more novices as well as (domain) experts, want to visually analyze their data. However, the challenges we are facing in these scenario are manifold. Therefore, in this open problem statement, I sketch two possible scenarios for a pervasive use of eye tracking data, the challenges associated with these scenarios, and some possible directions for future work to achieve such a pervasive use of gaze data.

### 5.7.2 Scenarios

In the following, I sketch two potential scenarios how pervasive eye tracking data can help people in their daily life. The first example focuses on a pervasive use of eye tracking in the context of mobility, in this case, while riding a bike. The second example focuses on the combination of pervasive eye tracking and visualization, in which data collected from many people is used to enhance the experience of data exploration for an individual.

#### Scenario 1: Pervasive Use of Eye Tracking for Cyclists

Imagine a cyclist named Mary riding her bike down the road on a busy street in a large city. Mary is on her way to a birthday party of a friend who she has never visited before. She has put on her bike helmet, which is equipped with an Electroencephalography (EEG) and an eye tracker with integrated augmented reality glasses, which measures the gaze as well as her head movements. In addition, she is wearing a smartwatch, which measures her Galvanic Skin Response (GSR), which she can use to show information about the direction of travel,

her location as well as surroundings, and which gives her feedback about her current stress level and dangerous situations on her way.

The system analyzes the data collected from the worn sensors in real-time comparing it with the surroundings to give her feedback about which way to go or potentially dangerous situations. For example, the system detects that Mary is checking her watch to see which way to go, highlights an important landmark using augmented reality to guide her in the correct direction and detects that she has not seen the car that is approaching from the left. The system then warns her of the car and sends a message to the autonomously driving car at the same time to communicate this possible impact. The car slows down and Mary, warned by the system, sees that the car is stopping and can safely pass the junction following her route.

### Scenario 2: Gaze Guided Visual Analytics

Sarah is an interested citizen and wants to investigate her communities' energy consumption to make an appropriate choice about which energy company to choose when she moves to a new apartment. As every laptop, hers is equipped with an eye tracker which records her eye movements and interactions while she is inspecting the website of her community. The energy data is represented using multiple visualizations and Sarah starts exploring these visualizations. The system automatically analyzes her gaze patterns and compares them to an underlying user model based on gaze data collected from a large number of other citizens who have visited the page.

First, the system uses her gaze data to estimate her expertise with the visualizations. After discovering that Sarah has not used the website before, the visualizations adapt themselves and display help information next to Sarah's gaze to educate her how to use the visualizations. After she transitions to the next level of expertise the help information fades away and the system starts to make suggestions on how to proceed with the exploration. Based on the collected eye movement data of many citizens a data narrative has been created which is used to guide Sarah through her exploration. Depending on the interests of Sarah and where she is looking this narrative automatically branches into multiple story lines guiding her through them. Based on her eye movements the system analyzes which data Sarah has already explored and gives her feedback about this and what information she might have missed.

### 5.7.3 Envisioned Challenges and Solutions

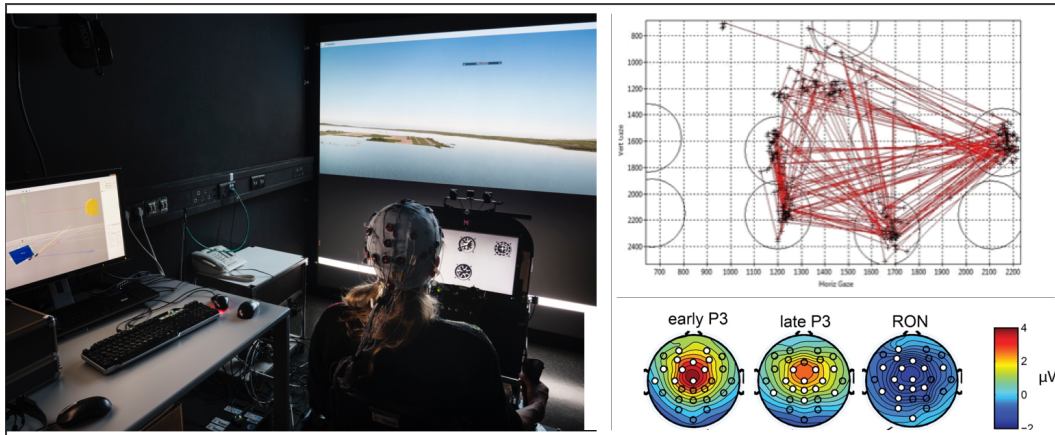
To come closer to these scenarios, we have to overcome different challenges. For the mobility scenario, the most important challenge that has to be solved are more robust eye tracking systems. First, the infrared light from the sun influences the eye tracking glasses and often leads to a loss of the gaze data [10]. In addition, mapping of gaze data for a pervasive use in real-world and real-time systems requires to map the data to known landscapes, for example, using GPS information in combination with the eye tracking data. This combination of different data sources and sensors (e.g., GPS, EEG, GSR) as well as multiple devices (e.g., eye tracking and augmented reality glasses, smartwatch) requires that the data is automatically recorded and synchronized as well as analyzed in real-time [1, 2]. For analyzing and giving visual feedback to the person, novel methods have to be developed that are unobtrusive yet quickly to grasp. For example, smartwatches can be used to display small-scale visualizations about the currently recorded data [4]. However, recording eye movement data in a public environment brings up the question of privacy issues for the person wearing the eye tracking glasses but also for the people that might be recorded while being on the move. Therefore, we have to find ways to ensure privacy.



The challenges for the second scenario include proper guidance when people are using a visual analytics system for data exploration. For example, if a novice is using a novel system containing visualizations, we have to know a person's intention. This requires that we define appropriate user models based on pre-recorded eye movement data, which we can then use as ground truth if a new person is using a system. Another important aspect is to engage and keep people engaged while using such a system. We have to detect from the eye movement data if a person is bored or overwhelmed to counteract this by showing tutorials or switching to an advanced mode. Especially, in scenarios where people are novices and visually illiterate [5] it is important to help them use a novel system. One possibility is to offer appropriate entry points [3] or use storytelling [6, 8] for a compelling narrative. This requires that we develop and provide appropriate guidance mechanisms [7]. Then, we can make sophisticated recommendations [9] to people using appropriate visual feedback about which part of the data to explore next.

## References

- 1 Blascheck, Tanja and John, Markus and Kurzhals, Kuno and Koch, Steffen and Ertl, Thomas. *VA<sup>2</sup>: A Visual Analytics Approach for Evaluating Visual Analytics Applications*. IEEE Transactions on Visualization and Computer Graphics, 22(1): 61–70, 2016.
- 2 Blascheck, Tanja and John, Markus and Koch, Steffen and Bruder, Leonard and Ertl, Thomas. *Triangulating User Behavior Using Eye Movement, Interaction, and Think Aloud Data*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM:175–182, 2016.
- 3 Blascheck, Tanja and MacDonald Vermeulen, Lindsay and Vermeulen, Jo and Perin, Charles and Willett, Wesley and Ertl, Thomas and Carpendale, Sheelagh. *Exploration Strategies for Discovery of Interactivity in Visualizations*. IEEE Transactions on Visualization and Computer Graphics, 2018 (in press).
- 4 Blascheck, Tanja and Besancon, Lonni and Bezerianos, Anastasia and Lee, Bongshin and Isenberg, Petra. *Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches*. IEEE Transactions on Visualization and Computer Graphics, 2018 (in press).
- 5 Boy, Jeremy and Rensink, Ronald and Bertini, Enrico and Fekete, Jean-Daniel. *A Principled Way of Assessing Visualization Literacy*. IEEE Transactions on Visualization and Computer Graphics, 20(12):1963–1972, 2014.
- 6 Boy, Jeremy and Detienne, Francoise and Fekete, Jean-Daniel. *Storytelling in Information Visualizations: Does It Engage Users to Explore Data?*. Proceedings of the Annual ACM Conference on Human Factors in Computing Systems, ACM:1449–1458, 2015.
- 7 Ceneda, Davide and Gschwandtner, Theresia and May, Thorsten and Miksch, Silvia and Schulz, Hans-Jörg and Streit, Marc and Tominski, Christian. *Characterizing Guidance in Visual Analytics*. IEEE Transactions on Visualization and Computer Graphics, 23(1):111–120, 2017.
- 8 Segel, Edward and Heer, Jeffrey. *Narrative Visualization: Telling Stories with Data*. IEEE Transactions on Visualization and Computer Graphics, 16(6):1139–1148, 2010.
- 9 Silva, Nelson and Schreck, Tobias and Veas, Eduardo and Sabol, Vedran and Eggeling, Eva and Fellner, Dieter. *Leveraging Eye-gaze and Time-series Features to Predict User Interests and Build a Recommendation Model for Visual Analysis*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM:13:1–13:9, 2018.
- 10 Trefzger, Mathias and Blascheck, Tanja and Raschke, Michael and Hausmann, Sarah and Schlegel, Thomas. *A Visual Comparison of Gaze Behavior from Pedestrians and Cyclists*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM:104:1–104:5, 2018.



■ **Figure 1** A remote gaze tracking system tracks the eyes of an amateur pilot as she tries to land a fixed-right aircraft (right). Her unpredictable scan path across the flight instruments reveal her levels of anxiety (top-left), and reduced EEG responses to auditory stimuli suggest that she is likely to miss radio messages, i.e., “inattentional deafness” (bottom-left).

## 5.8 Inferring the Deployment of Limited Attentional Resources

Lewis Chuang (LMU München, DE)

License Creative Commons BY 3.0 Unported license  
© Lewis Chuang

### 5.8.1 Introduction

Gaze-tracking systems are increasingly prevalent, not only in laboratories but also in our daily environment. The implementation of gaze-tracking systems in the real-world could either be personal, such as head-worn devices, or not, such as those that are implemented in public display systems. In either case, gaze trackers collect data on when the user is looking at what, and how long for. No more, no less. How should we translate this data to draw meaningful inferences of how the user is acquiring and processing visual information? This is necessary in order for us to design computing systems that can be aware of their users needs.

In my lab <http://humanmachinesystems.org>, we are interested in understanding how limited attentional resources are deployed during user interactions with closed-loop machine systems. We regard attentional resources broadly as any physiological resource that is available to the user, that can allocated to selectively increase the gain of an information channel, often times over other channels; Or as William James [8] have said:

Attention . . . is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought, localization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatter brained state which in French is called *distracted*, and *Zerstreuung* in German.

For this, we employ the methods of eye tracking and electroencephalography (EEG) (see Figure 1).

### 5.8.2 Envisioned Challenges

With regards to gaze, this is achieved by moving one's fovea to an area-of-interest (AOI) in the visual scene. This allows spatial information (e.g., edges, contours, object form) to be better resolved and distinguished from the background. This is referred to as overt attention. Many attempts have been made to extract further information from gaze characteristics in order to understand how information might be amplified post-foveation, i.e., covert attention. Unfortunately, our understanding of this is relatively limited. To answer this, it is necessary to first understand the nature of information that serves the purposes of the human observer. This is difficult to answer as human observers are unlikely to be consciously operating on image information, in the same terms as would describe the fovea as an image sensor.

It is popular to treat pupil dilation as an index of cognitive load [2]. However, recent findings show that pupil dilation is not only sensitive to ambient lighting, but to the color of the fixated object itself [12]. Clearly, gaze tracking systems are not intended to infer the reflected luminance and ambient luminance of fixated objects. Thus, the depressing message could be that pupil dilation is not a useful measurement at all. However, this dim (and hasty) prognosis is unwarranted [10]. Rather, we need better models of the task and context that gaze is embedded in, in order to discount the variables that influence our estimation of gaze features [11]. A naive solution could be to couple a scene camera with an eye tracker, which would allow us to normalize pupil dilation to ambient illumination. An object recognition algorithm could allow us to further normalize pupil dilation to the likely color of the car model in the scene image.

To summarize, we cannot make meaningful inferences from what an observer is observing ("in-the-wild") without understanding what the observer considers to be meaningful. Furthermore, what the observer perceives to be meaningful could influence our measured gaze characteristic for reasons other than information processing it. Better models that account for context are necessary to allow for meaningful inferences from recorded gaze features.

### 5.8.3 Envisioned Solutions

We believe that gaze movements are planned actions of a goal-oriented observer that seek out task-relevant information. Thus, it would stand to reason that the predictability of eye movements reflect the ability of the observer to execute this plan. Indeed, we have recently reported that highly anxious "*pilots*" were more likely to generate chaotic and unpredictable eye movements across their flight instruments, especially when they experienced high cognitive load [1]. This is in line with the predictions of attentional control theory [5], which suggests that high anxiety levels and working memory load can compromise executive function. With this in mind, it is possible that ubiquitous gaze tracking systems could be employed to adapt the computing work environment to the user's state.

The EEG response to physical events could also reflect resource availability at the cortical level. In other experiments, we demonstrated that increasing the difficulty of a visuomotor control task diminished the EEG response to irrelevant sounds (i.e., environment sounds), specifically the novelty P3 [13, 14]. The implication is that high cognitive load can prevent users from noticing unexpected but potentially important events, from a calendar notification to emergency warnings. This is termed *inattentional deafness* [4], which could be rectified if the appropriate notifications are presented for the correct context. Another use of EEG could be to validate the design of novel computing systems, in particular those designed to support cognitive work. In another example, we employed EEG to confirm that an *in situ* haptic assembly system that employed augmented reality to reduce visuospatial working

memory load, did in fact target the same neural correlates as a standardized test for the same cognitive process [9].

To understand how relevant information is managed and attended to post-fixation, it might be optimal to combine both gaze tracking and EEG methods, e.g., [7, 6], as well as a scene camera to infer the context that this activity is embedded in. More importantly, it is important to ensure that results that are derived under laboratory settings are robust and can be generalized to more realistic settings that reflect the variability of real-world settings [3]. As I cast my gaze towards the future, I envision a scenario where a robust understanding of what data means will enable us to delegate to computing systems, the task of making sense of the copious gaze (+ EEG + etc.) data that we are continuously recording of our daily experience.

## References

- 1 Jonathan Allsop, Rob Gray, Heinrich Bülthoff, and Lewis Chuang. Eye movement planning on single-sensor-single-indicator displays is vulnerable to user anxiety and cognitive load. *Journal of Eye Movement Research*, 10(5), 2017.
- 2 Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 1982.
- 3 Lewis L Chuang, Christiane Glatz, and Stas Krupenia. Using eeg to understand why behavior to auditory in-vehicle notifications differs across test environments. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ACM:123–133, 2017.
- 4 F. Dehais, M. Causse, François Vachon, N. Regis, E. Menant, and Sébastien Tremblay. Failure to Detect Critical Auditory Alerts in the Cockpit: Evidence for Inattentive Deafness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(4):631–644, 2014.
- 5 Michael W Eysenck, Nazanin Derakshan, Rita Santos, and Manuel G Calvo. Anxiety and cognitive performance: attentional control theory. *Emotion*, 7(2):336–353, 2007.
- 6 N. Flad, H. H. Bülthoff, and L. L. **Chuang**. Combined use of eye tracking and EEG to understand visual information processing. In *Proceedings from the International Summer School on Visual Computing (VCSS 2015)*, Fraunhofer Verlag:115–124, 2015.
- 7 N. Flad, T. Fomina, H. H. Bülthoff, and L. L. **Chuang**. Unsupervised clustering of EOG as a viable substitute for optical eye tracking. In Michael Burch, Lewis Chuang, Brian Fisher, Albrecht Schmidt, and Daniel Weiskopf, editors, *Eye Tracking and Visualization: Foundations, Techniques, and Applications*, Mathematics and Visualization, Springer:151–168, 2016.
- 8 William James. *The principles of psychology*. Read Books Ltd, 2013.
- 9 Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L Chuang. Assessing the cognitive demand of assistive systems at assembly workplaces using electroencephalography. *Proceedings of the ACM on Human-Computer Interaction – EICS*, 2018 (accepted).
- 10 Sebastiaan Mathôt. Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), 2018.
- 11 Vsevolod Peysakhovich, François Vachon, and Frédéric Dehais. The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology*, 112:40–45, 2016.
- 12 Bastian Pflöging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, ACM:5776–5788, 2016.

- 13 Menja Scheer, Heinrich H Bülthoff, and Lewis L Chuang. Steering Demands Diminish the Early-P3, Late-P3 and RON Components of the Event-Related Potential of Task-Irrelevant Environmental Sounds. *Frontiers in Human Neuroscience*, 10:73, 2016.
- 14 Menja Scheer, Heinrich H. Bülthoff, and Lewis L. Chuang. Auditory task irrelevance: A basis for inattentive deafness. *Human Factors*, 1-13, 2018 (in press).

## 5.9 Short-term Gaze-based User Intent

Andrew Duchowski (Clemson University, US)

**License** © Creative Commons BY 3.0 Unported license

© Andrew Duchowski

**Main reference** Andrew T. Duchowski: “Gaze-based interaction: A 30 year retrospective”, Computers & Graphics, Vol. 73, pp. 59–69, 2018.

**URL** <http://dx.doi.org/10.1016/j.cag.2018.04.002>

### 5.9.1 Introduction

For gaze sensing and interaction to become ubiquitous, use-case scenarios should be developed. Although a myriad tasks may be proposed for everyday use of gaze-based interaction, one compelling assumption made in these scenarios is prediction of the user’s intent through analysis of their gaze. For example, when looking at an object, a lamp say, the lamp should infer the user’s intent, (turning on the lamp). This style of interaction was envisioned by Vertegaal [8, 9] as Attentive User Interfaces. Since then, various other situations have appeared where gaze-based intent could offer predictive benefit, e.g., decision-support systems [3].

### 5.9.2 Challenges

In gaze sensing interactive systems, prediction of the user’s intent would need to be inferred over variable periods of time, i.e., over the short-, medium-, or long-term. In the medium-term, for example, gaze-based intent could be exploited for divining the next object to be reached for in, for example, a sandwich-making task. Long-term prediction is likely be more complex than in the short-term. The very short-term could be as short as a few milliseconds, during which the location of the saccade end point could be predicted.

### 5.9.3 Short-Term Prediction of Intent

A fairly straightforward approach to predicting the user’s *short-term* visual intent is to estimate what is going to be fixated next by predicting the saccade endpoint mid-flight. The basic premise dates back to [1] who showed that predicting saccade termination was possible by detecting saccadic peak velocity, and then mirroring the saccade velocity profile.

The assumed symmetry of the velocity profile only holds for small amplitude saccades. As saccade amplitudes increase, the velocity profile assumes a Gamma distribution. That is, the velocity profile of small saccades is symmetrical but is skewed for large saccades, and can be modeled by the expression

$$V(t) = \alpha \left( \frac{t}{\beta} \right)^{\gamma-1} e^{-t/\beta}$$

where time  $t \geq 0$ , and  $\alpha, \beta > 0$  are scaling constants for velocity and duration, respectively. Shape parameter  $2 < \gamma < 15$  determines the degree of asymmetry [7]. When  $\gamma$  is small,

asymmetrical velocity profiles are produced and as  $\gamma$  tends to infinity, the velocity profile assumes a symmetrical (Gaussian) shape.

A more recent approach for saccade endpoint prediction was demonstrated by Arabadzhiyska et al.[2]. They exploit the above observation of saccades obeying ballistic trajectories dependent mainly on saccade amplitude. They then develop and demonstrate an elegant and robust data-driven model that can adequately predict saccade landing position. Model parameters are set by first performing measurements to collect samples of saccades ranging between 5–45 degrees performed by several participants. These samples serve as a kind of look-up table from which saccade characteristics are obtained in real-time use in a foveated display.

#### 5.9.4 Applications: Foveated Displays

Short-term saccade endpoint prediction is particularly well suited to overcoming eye tracking latency associated with foveated rendering. In general, an eye tracker requires at least one frame of eye camera video to compute the gaze point. If there are digital, e.g., finite-impulse, filters in use to analyze the real-time signal, additional latency is incurred proportional to the filter width.

To match perceptibility of a full-resolution display, the foveated central inset should appear within 7 ms of fixation onset [5]. Greater delays (e.g., 15 ms following fixation onset), are detectable but have minimal impact on performance of visual tasks when the radius of the foveal inset is large ( $\geq 4^\circ$ ). Due to saccadic suppression, delays as long as 60 ms do not significantly increase blur detection [6]. Note that the latter pertains to the time following saccade termination (60 ms), the former to time following fixation onset (7 ms). Either way, appearance of the foveal inset must be updated before the update is noticed.

Being able to predict, in the short-term, the user's intent to switch gaze to a new location reduces latency of the foveated central inset. Thus estimation of the user's intent through gaze analysis affords computational savings in terms of graphics performance and reduces potential impairment to perception of the scene being rendered.

#### References

- 1 James Anliker. *Eye Movements: On-Line Measurement, Analysis, and Control*. Eye Movements and Psychological Processes, Lawrence Erlbaum Associates:185–202, 1976.
- 2 Arabadzhiyska, Elena and Tursun, Okan Tarhan and Myszkowski, Karol and Seidel, Hans-Peter and Didyk, Piotr. *Saccade Landing Position Prediction for Gaze-Contingent Rendering*. ACM Transactions on Graphics, 36(4):50:1–50:12, 2017.
- 3 Julia Behrend. *The influence of personality, habits, hierarchy, and role on decision-making in real-life: application to aviation safety*. L'Universite Pierre et Marie Curie, 2018.
- 4 Han Collewyn, Casper J. Erkelens, and Robert M. Steinman. *Binocular Co-Ordination of Human Horizontal Saccadic Eye Movements*. Journal of Physiology, 404:157–182, 1988.
- 5 Lester C. Loschky and George W. McConkie. *User Performance With Gaze Contingent Multiresolutional Displays*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM: 97–103, 2000.
- 6 Lester C. Loschky and Gary S. Wolverson. *How Late Can You Update Gaze-Contingent Multiresolutional Displays Without Detection?*. Transactions on Multimedia Computing, Communications and Applications, 3(4):7:1–7:10, 2007.
- 7 A.J. Van Opstal and J.A.M. Van Gisbergen. *Skewness of Saccadic Velocity Profiles: A Unifying Parameter for Normal and Slow Saccades*. Vision Research, 27(5):731–745, 1987.
- 8 Roel Vertegaal. *Designing Attentive Interfaces*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM:23–30, 2002.
- 9 Roel Vertegaal. *Introduction*. Commun. ACM, 46(3):30–33, 2003.



## 5.10 The Potential of Gaze-based Training in Psychotherapy

Nina Gehrer (Universität Tübingen, DE)

**License** © Creative Commons BY 3.0 Unported license  
© Nina Gehrer

**Main reference** Nina A. Gehrer, Michael Schönenberg, Andrew T. Duchowski, Krzysztof Krejtz: “Implementing innovative gaze analytic methods in clinical psychology: a study on eye movements in antisocial violent offenders”, in Proc. of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA 2018, Warsaw, Poland, June 14–17, 2018, pp. 41:1–41:9, ACM, 2018.

**URL** <http://dx.doi.org/10.1145/3204493.3204543>

### 5.10.1 Abstract

Our eye movements guide attentional processes and play a crucial role in our perception and interpretation of the world. Different impairments in these processes have been associated with a variety of psychological disorders. Thus, the use of eye tracking methods in clinical psychology is a promising approach to gain more insight into perceptual and cognitive impairments underlying the etiology of disorders such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), or psychopathy. Furthermore, eye tracking is not only a useful tool for clinical research but also for psychotherapy. The development of gaze-based training as therapeutic intervention or prevention method is a new research area and a promising avenue for innovative treatment strategies.

### 5.10.2 Introduction

The eyes are the most important interface between our environment and ourselves. The way our gaze scans our surroundings and lingers on certain details determines where we direct our attention, what we perceive and ultimately, how we respond.

In social interactions, detecting and understanding socially important cues is crucial for the functional communication with other individuals and the development of social skills. Thus, we typically show a strong tendency to look at faces and particularly the eyes, which can convey valuable nonverbal information regarding the emotional and cognitive state of the interaction partner [3]. This preference is rooted deeply in our brains and is evident even in infants [4]. Attention-orienting processes to socially salient cues are essential for concepts such as gaze following, joint attention and eye contact, which play a role in the development of higher order social functions such as theory of mind, social bonding and even language acquisition [8]. Accordingly, previous studies have linked impairments in attention orienting to social cues (e.g., the eyes) with various psychiatric conditions, such as autism spectrum disorder (ASD) and psychopathy [9, 2]. Therefore, using eye tracking in clinical research allows us to gain additional insight in psychological processes underlying the perception of social cues and differences that might be associated with dysfunctions or impairments [5].

Furthermore, numerous eye tracking studies have documented deficits in the very basic oculomotor functions in association with psychological disorders [10]. Accordingly, deficiencies in smooth pursuit have been reported in patients with schizophrenia [7]. Further, previous findings in children with ASD or attention deficit hyperactivity disorder (ADHD) have indicated that the associated impairments in inhibition mechanisms might also affect the eye movements [11].

Therefore, investigating eye movements in clinical psychology is a fruitful approach to gain more insight into deficits in oculomotor and attentional processes associated with different psychological disorders.

### 5.10.3 Potential and Challenges

In clinical psychology and psychotherapy, using eye tracking could help to develop a better understanding of etiology and to learn more about underlying processes of various psychological disorders. Further, eye tracking might be useful as an additional tool in diagnostic procedures if eye movement measures could be shown to provide reliable markers. Finally, gaze-based training is a potentially powerful tool for therapeutic interventions that could address impairments and biases of information processing associated with specific disorders.

First attempts to implement gaze-based training already exist for children with ASD and ADHD. For instance, Chuokoskie et al. [1] developed a robust, low-cost, gaze-contingent game system to train specific oculomotor functions in adolescents with ASD in domestic settings and the preliminary results are very promising. Further, Goodwin et al. [6] are conducting a randomized controlled trial to explore the potential of gaze-based training during infancy as a method to prevent the development of ADHD.

However, the relevance of deficient oculomotor functions and impaired attention orienting in associated psychological disorders remains to be clarified and many important questions are pending: Can these impairments be addressed by specific training? Can improvements in these dysfunctional processes lead to improvements in other symptoms? Do improvements during training transfer to real-life settings? Furthermore, there are technical issues that have to be solved in order to facilitate the development and use of gaze-based therapeutic interventions and to reach their full potential. These challenges start with common issues such as accessibility, robustness, and usability and extend to more complex problems. For instance it would be interesting to develop a mobile eye tracking system that automatically recognizes salient social cues (e.g., eyes or faces) in real-life settings and includes gaze-based recommendations or signals to direct attention to these stimuli. Thus, the development of gaze-based training is a new promising research area albeit the more studies are necessary in order to determine the optimal training targets, methods, their usability, generalizability, and durability.

### References

- 1 Leanne Chukoskie, Marissa Westerfield, and Jeanne Townsend. *A novel approach to training attention and gaze in ASD: A feasibility and efficacy pilot study*. *Developmental Neurobiology*, 78(5):546–554, 2018.
- 2 Mark R. Dadds, Yasmeen El Masry, Subodha Wimalaweera, and Adam J. Guastella. *Reduced eye gaze explains “fear blindness” in childhood psychopathic traits*. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(4):455–463, 2008.
- 3 N. J. Emery. *The eyes have it: The neuroethology, function and evolution of social gaze*. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.
- 4 Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H. Johnson. *Eye contact detection in humans from birth*. *Proceedings of the National Academy of Sciences*, 99(14):9602–9605, 2002.
- 5 Nina A. Gehrer, Michael Schönenberg, Andrew T. Duchowski, and Krzysztof Krejtz. *Implementing innovative gaze analytic methods in clinical psychology*. *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:41:1–41:9, 2018.
- 6 Amy Goodwin, Simona Salomone, Patrick Bolton, Tony Charman, Emily J. H. Jones, Andrew Pickles, Emily Robinson, Tim Smith, Edmund J. S. Sonuga-Barke, Sam Wass, and Mark H. Johnson. *Attention training for infants at familial risk of ADHD (INTERSTAARS): Study protocol for a randomised controlled trial*. *Trials*, 17:608, 2016.
- 7 Sam Hutton and Christopher Kennard. *Oculomotor abnormalities in schizophrenia*. *Neurology*, 50(3):604–609, 1998.



- 8 Roxane J. Itier and Magali Batty. *Neural bases of eye and gaze processing: The core of social cognition*. Neuroscience & Biobehavioral Reviews, 33(6):843–863, 2009.
- 9 Warren Jones, Katelin Carr, and Ami Klin. *Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder*. Archives of General Psychiatry, 65(8):946–954, 2008.
- 10 Canan Karatekin. *Eye tracking studies of normative and atypical development*. Developmental Review, 27(3):283–348, 2007.
- 11 Nanda N.J. Rommelse and Stefan Van der Stigchel and Joseph A. Sergeant. *A review on eye movement studies in childhood and adolescent psychiatry*. Brain and Cognition, 68(3):391–414, 2008.

## 5.11 Eye Movement as Design Material

Hans Gellersen (Lancaster University, GB)

License  Creative Commons BY 3.0 Unported license  
© Hans Gellersen

Eye tracking has been long adopted as input device for interaction. The data model of eye trackers makes it convenient to think of them as a pointing device that, much like a mouse, provides a continuous stream of coordinates within a 2D space. Coupled with a display, designers can work with simple abstractions such as points and regions, adopting eye trackers as a black box that hides the intricacies of eye movement. I argue that it is time to open the box – rather than thinking of eye tracking as input device, we should think of eye movement as a material for interaction design.

The conceptual model for gaze interaction appears straightforward: harnessing “what we look at” either as implicit indication of interest a system can respond to, or as an explicit selection of input. However, while a user’s mental model might be the same for looking at an object that sits still in the field of view versus one that is in motion, there are fundamental differences in the underlying eye movement processes. Gazing at an object in motion involves smooth pursuit eye movement, a closed loop behavior that is distinct from the saccadic movement otherwise observed. Importantly, this behavior only occurs when there is a moving stimulus for the eyes to follow. This has profound implications for design – whether a user is looking at a moving object can be robustly detected from the correlation of eye movement with the object’s motion, as implemented in the *Pursuits* technique [6]. As a consequence, eye tracker and visual environment need only be coupled loosely. There is no need for their coordinate systems to be carefully aligned prior to interaction as the input is based on correspondence of eye motion with motion in the environment. This opens up an entirely new design space for eye gaze: content can be made gaze-aware by presenting it in motion [6]; users can be calibrated implicitly to bootstrap gaze pointing [2]; gaze control can be dynamically associated with ubiquitous devices [5]; and animated widgets can be designed for gaze-only control, even with devices as small as smartwatches [1].

The notion of eye tracking as a pointing device has tended to position eye gaze as alternative to manual action. However, there is a natural interplay and complementarity of eye gaze and manual action – eye gaze precedes action and guides manual input. *Gaze&Touch* demonstrated how the respective strengths of the two modalities can be leveraged for multi-modal input, where “gaze selects, and touch manipulates”. This can seamlessly extend multi-touch and gestural interfaces, by applying manual gestures to objects selected by gaze [3]. Where gaze naturally moves ahead of manual action, manual input can be translated to the

gaze location – shifting the frame of reference for manual input, such that the eyes take on the larger and less accurate movement on the interface, while the hand performs smaller and fine-grained input. *Gaze-shifting* also showed that the coincidence of eye gaze and manual input is significant – the same manual input action can take on different meanings, modulated by gaze attention [4]. How to couple gaze and manual action will be particularly intriguing as we move from touch to touchless interactions that at present lack clear conceptual models.

There is much more to gaze that is waiting to be uncovered and developed for interaction design. State of the art eye trackers are single user devices, and there is a vast space to explore multi-user eye gaze, concepts such as mutual gaze and joint attention, and gaze as social signal [7]. As we move from the narrow fields of view that we have in front of desktop interfaces, to interaction with virtual and augmented reality, consideration of gaze will also require a more holistic approach that accounts for head and body movement. Gaze shifts in the real world involve complex interaction of these movements – smaller shifts are performed with the eyes only, whereas larger ones involve head and body movement. How these observations translate to interaction with novel types of display is an open question and fundamental for the design of techniques that couple gaze, head pose, and body pose for natural forms of interface.

## References

- 1 Augusto Esteves, Eduardo Velloso, Andreas Bulling and Hans Gellersen. *Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze*. Proceedings of UIST'15: User Interface Software & Technology, ACM:457–466, 2015.
- 2 Ken Pfeuffer, Mélodie Vidal, Jayson Turner, Andreas Bulling and Hans Gellersen. *Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze*. Proceedings of UIST'13: User Interface Software & Technology, ACM:261–270, 2013.
- 3 Ken Pfeuffer, Jason Alexander, Ming Ki Chong and Hans Gellersen. *Gaze-touch: combining gaze with multi-touch for interaction on the same surface*. Proceedings of UIST'14: User Interface Software & Technology, ACM:509–518, 2014.
- 4 Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang and Hans Gellersen. *Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze*. Proceedings of UIST'15: User Interface Software & Technology, ACM:373–383, 2015.
- 5 Eduardo Velloso, Markus Wirth, Christian Weichel, Augusto Esteves and Hans Gellersen. *AmbiGaze: Direct Control of Ambient Devices by Gaze*. Proceedings of DIS'16: Designing Interactive Systems, ACM:812–817, 2016.
- 6 Mélodie Vidal, Andreas Bulling and Hans Gellersen. *Pursuits: spontaneous interaction with displays based on smooth pursuit eye movement and moving targets*. Proceedings of UbiComp'13: Pervasive and Ubiquitous Computing, ACM:439–448, 2013.
- 7 Mélodie Vidal, Rémi Bismuth, Andreas Bulling and Hans Gellersen. *The Royal Corgi: Exploring Social Gaze Interaction for Immersive Gameplay*, Proceedings of the Annual ACM Conference on Human Factors in Computing Systems, ACM:115–124, 2015.

## 5.12 Why USGI Needs Better Eye Trackers

*Kenneth Holmqvist (Universität Regensburg, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Kenneth Holmqvist

**Main reference** Holmqvist, Kenneth and Andersson, Richard “Eye tracking: A comprehensive guide to methods, paradigms and measures,” Lund: Lund Eye-Tracking Research Institute, 2017.

### 5.12.1 Introduction

None of the UGSI work will be real until we have an eye tracker that can provide good enough data. Loosely based on [1, Chapters 3 and 6], I describe the issues with camera-based VOGs and introduce some of the alternatives.

### 5.12.2 Envisioned Challenges

Imagine an AR system with a classical video-camera-based system. Our user (consumer?) is trying it out. S/he is doing banking tasks, way finding, checking emails or maybe interacts with an augmented character in a sports application. However because the eye tracker is camera-based, it suffers from a whole range of issues that have been well-studied over the last 20 years. Firstly, the software for rendering the menus and the AR visualizations try to predict the saccadic landing positions mid-flight, which in theory is possible, since saccades are ballistic and have a well-known velocity shape. However, the pupil-corneal reflection technique overestimates the true velocity of the eye [9], and always miscalculates the target by half a degree or more, which makes the rendering seem jumpy. Young people in their teens and up to thirty years of age report that the augmented reality objects slide back and forth, several degrees, for no obvious reason, not knowing that it is because of the variable pupil dilation and the motion of the pupil center that follows with changes in pupil dilation [10]. Furthermore, rendering works better for the brown-eye user than for blue-eyed users who often experience instabilities in the image. Users with contact lenses find the AR objects to be jumping back and forth several degrees at high speed, and refuses to wear it. People who happened to be in the sun when using it report a total loss of functions. A user who was in a room with an old, hot light-bulb report the that all objects are off, and move away from her when she tries to look at them. Everyone report a lot of instability in the imagery, and no-one wants to buy this system [1, Chapter 6].

### 5.12.3 Envisioned Solutions

Precision, accuracy and robustness at an unprecedented scale is necessary. Precision must be at the absolute minimum to avoid noise in the image. The closer to 0, the better. It is easy to calculate what the accuracy requirements on the eye tracker are if we know the size of the objects that the user interacts with [3]. I think we should not be satisfied until we can distinguish which line of text the user looks at in an email at a normal reading distance. Then we need below  $0.05^\circ$  average accuracy, to be compared with around  $0.5^\circ$  for the best VOGs – assuming inexperienced participants. For a long time, there has been a mistaken belief that the optimal accuracy is limited by the size of the fovea. This is based on the erroneous assumption that we can use any part of the fovea for detailed inspection of very small objects. However, as shown by [2], it is possible to calibrate a DPI system to much finer accuracy, and once having done that, they find movements of the eye smaller than the diameter of the fovea with an accuracy below  $0.05^\circ$ .

Robustness is the tricky parameter to achieve. Current video-based eye trackers work for 90-95% of the student population in Europe, but as soon as they have an eye physiology that makes the eye cleft more narrow, or droopy eye lids, downward going eye lashes, large and variable pupil sizes, contact lenses, blue eyes, mascara or eye-liner, not to speak about glasses with anti-reflective coating, data are not usable even for eye movement studies with lenient data quality requirements.

One important thing is to stop using the pupil as a feature in eye tracking. The pupil feature is the root to many of the issues; it varies in size (causing offsets), the pupil border is more difficult to detect in infrared for blue-eyed people, and the pupil moves differently than the eye ball during saccades, because the inertia in the lens affects the pupil. None of this is OK if we are building an eye tracker for augmented reality, so we cannot use the pupil feature.

Another priority is to stop using cameras. Video cameras that capture frames of the eye have several drawbacks. Firstly, they require a lot of energy, and more energy for higher resolutions. Second, they are pixel-based, which means that the smooth analog movement of the eye is quantized. This creates artifacts in the data when small movements are recorded, so that at gaze directions, the movements are amplified and at other gaze directions, movements are compressed. This effects seems to be worse for lower camera resolutions and fewer corneal reflections. Furthermore, what is the point of recording and transferring a high-resolution image at a high frame rate of many hundred Hz, when we only use two coordinate values from it? It seems utterly wasteful.

The corneal reflection (aka glint) is a better signal. Analogue recordings of eye movements were used from the very start of eye-movement research [4, 5, 6] and data that we have from that time are excellent, but the eye trackers were difficult to use and uncomfortable for participants. Contemporary video-based eye trackers do use the corneal reflection, so in principle we could recreate the good signals. Alas, the quantizing resolution of the eye cameras typically introduce a lot of noise in the signal of the corneal reflection which make small movements (microsaccades) unreliably measured, and smooth pursuit and saccade waveforms noisy.

In 1973, the Dual-Purkinje Imaging (DPI) eye tracker was introduced. Because of its excellent data quality, the DPI has been a major working horse in psychology labs throughout the 1980s and 1990s. It utilizes the corneal reflection plus the reflection the back of the lens, that is, the first and the fourth Purkinje reflections. A system of lenses and mirrors leads the two reflections to each a quadrant detector, which attempts to alter the mirror to keep the reflection at the center. This energy needed to do that is output as two analog voltage signals, one for horizontal and one for vertical. The DPI is difficult to use, by todays standards, but some of the design choices are worth looking into. In particular, the idea of an analog eye tracker with no sampling frequency and no quantization of measurement space are very appealing features for data to be used in augmented reality. Using the 4th Purkinje reflection is not a good idea, as the 4th Purkinje contributes strongly to the erroneous measurements of velocity of saccades, and their post-saccadic oscillations. The 4th Purkinje also disappears behind the pupil border for large off-center gaze directions, which makes the tracking range small for small-pupil user.

Scleral search coils result in excellent data. [7] pronounced coils to be the gold standard in eye tracking. Recently, they have been found to have a data quality on par with the DPI [8]. However, who would want to wear coils on their eye balls during augmented reality activities? And moreover, the participant must be positioned inside an oscillating magnetic field, which is very impractical.

Retinal eye tracking has been used since the 1950s. No other measurement technique can capture such small movements. However, because existing systems have been variants of ophthalmoscopes, the eye tracker (camera) efficiently blocks the view of the participant. It has never become a technique that could be sold to researchers.

Today, analog micro-electronics such as MEMS and quadrant detectors can do many of the things the DPI could do. We are likely to very soon see very good eye trackers that are based on such techniques.

## References

- 1 Holmqvist, Kenneth and Andersson, Richard. *Eye tracking: A comprehensive guide to methods, paradigms and measures*. Lund: Lund Eye-Tracking Research Institute, 2017.
- 2 Poletti, Martina and Rucci, Michele *A compact field guide to the study of microsaccades: challenges and functions*. Vision Research, 118:83–97, 2015.
- 3 Jacob L Orquin and Kenneth Holmqvist, *Threats to the Validity of Eye-Movement Research in Psychology*, Behavior Research Methods. 50(4):1645–1656, 2017.
- 4 Dodge, R. and Cline, T.S. *The angular velocity of eye movements*, Psychological Review, 8:145–157, 1901.
- 5 Buswell, G. T. *How People Look at Pictures*. Chicago: University of Chicago Press, 1935.
- 6 Yarbus, A. L. *Eye Movements and Vision*. New York, NY: Plenum Press, 1967.
- 7 Collewyn, H. *Vision Research: A Practical Guide to Laboratory Methods*, Oxford: Oxford University Press, 1988.
- 8 Hee-kyoung Ko and D. Max Snodderly and Martina Poletti *Eye movements between saccades: Measuring ocular drift and tremor*, Vision Research, 122:93–104, 2016.
- 9 Ignace Hooze and Kenneth Holmqvist and Marcus Nyström, *Pupil-CR technique is not suitable for studying detailed dynamics of eye movements* Vision Research, 128:6–18, 2016.
- 10 Drewes, J., *Smaller is better: drift in gaze measurements due to pupil dynamics* PLoS ONE, 9(10):1–6, 2014.

## 5.13 Who Watches the Watchmen: Eye Tracking in XR

Eakta Jain (University of Florida – Gainesville, US)

**License** © Creative Commons BY 3.0 Unported license  
© Eakta Jain

**Main reference** Daniel J. Liebling, Sören Preibusch: “Privacy considerations for a pervasive eye tracking world”, in Proc. of the The 2014 ACM Conference on Ubiquitous Computing, UbiComp ’14 Adjunct, Seattle, WA, USA – September 13–17, 2014, pp. 1169–1177, ACM, 2014.

**URL** <http://dx.doi.org/10.1145/2638728.2641688>

### 5.13.1 Introduction

The push towards reliable, affordable, and universal eye tracking is being driven by the promise of headset-based mixed reality (XR), which includes augmented reality (AR) and virtual reality (VR), in particular, social VR. Social XR envisions a future where everyone wears headsets for long periods of time. These headsets could be augmented reality or virtual reality, the important thing being that eye trackers will be built into them. At the very beginning, eye trackers will provide the data needed to improve XR systems at the enabling technology level, such as for foveated rendering and alleviating the vergence-accommodation conflict, and for creating the social avatar’s eyes. After this, eye tracking data will be used to improve user-friendliness of the system, for example, by combining gaze and gesture to improve gesture recognition, or by creating novel user interaction design by leveraging shared attention. Eye-tracking will also be valued for user identification, for example by

iris-scanning, and for improved security, such as continuous authentication via individual specific patterns of eye movements. We will find that these systems can contribute to tracking health and wellness, and become a critical enabler for artificial intelligence-based personalized interventions.

The flip side to these amazing possibilities is that we will allow someone to watch us with an unprecedented intimacy. Eye-tracking data encodes where we are looking, and that is not entirely under our conscious control. Where does a man look when he accompanies his wife shopping? This is clearly private information, said man will argue. Eye-tracking data reveals subtle preferences that can be used for targeted advertisements. In fact, as we will be making our way through virtual and augmented worlds, advertising will shift from clearly demarcated banners to product placement that is integrated into the augmented or virtual experience.<sup>1</sup> Eye-tracking data contains indicators of medical or behavioral conditions [4]. An insurance company might create a social VR app that lets you walk through a hypothetical claims process with a simulated insurance agent, but collects eye movement markers to check for pre-conditions.

In many ways, the scenario above is similar to how we today ‘wear’ our phones, and allow it to collect multi-modal data on a near-continuous basis. Some of this data is used to improve the quality of the service, such as voice data to improve speech recognition for voice commands, and location data to make search results more relevant to the user. The systems and software architects of these platforms created application programming interfaces (API) that passed this information to third party apps to develop new tools and services to make the user’s life better connected, more convenient, safer, and healthier. Though the user is asked to give permission to third party apps to access their data, those developers often request more information than strictly necessary, and users do not fully understand what they are giving up [7, 3, 1]. The interfaces are not necessarily the easiest to navigate and understand [8].

I propose that we think along three vectors before rather than after eye tracking becomes a pervasive sensor in consumer products.

**Social VR platform architects:** As the platforms for VR and AR develop, the platform creators have full control over the data that is collected and then passed up the software stack. They will determine, for example, the handles that will be provided to third party app developers, the resolution of data they will get, and how well the data are separated. In this domain, the data being recorded will be much richer than previous domains. For example, location, whole body gestures, and fine-grained facial tracking will be needed to create compelling avatars. Eye-tracking data will be collected to enable foveated rendering, consistent focus cues, and to replicate the user’s gaze on their social avatar. This last bit is critically different from other domains: even if the user was to turn off foveated rendering, and opt out of gaze-based interaction, they *need* to let the system track their eyes for their avatar in social VR applications. That does not mean they wish to be targeted for advertising based on probabilistic predictions of medical conditions based on eye tracking data for example. From a systems architecture perspective, the open question is: what are the privacy preserving APIs that will control the type and resolution of data that is passed up the software stack to third party apps?

**User experience designers:** The idea that enabling some form of data gathering for the purpose of task A can also enable task B is not necessarily intuitive to users. Eye tracking

---

<sup>1</sup> This shift can already be seen in search results where sponsored links are often interspersed with search results.

as a general term encompasses several data: raw gaze location, vergence, pupil diameter, fixations, saccades, blinks, microsaccades, and so on. A lay user may not make the connection that if the social VR app collects eye tracking data to create her real-time avatar as she hangs out with her friends, and has access to the scenes she was looking at (browsing history), then the app can infer what objects she looked at and for how long. The open questions here relate to the visualization and user experience (UX) innovations that are needed to educate the user and give her the controls to customize what she shares.

**Eye-tracking technologists:** Eye-tracking data is unique because it reveals our interests and preferences as well as other intrinsic characteristics such as affect, age and health. As eye tracking gets built into VR and AR headsets, it should be possible to perform a given task A while guaranteeing that some other task B cannot occur (a privacy guarantee for task B). For example, if high sample rate eye tracking data is smoothed, could we create a believable social avatar without being able to access the markers that might be indicative of degenerative conditions [5]? Several computer science subcommunities have thought about the privacy problem [2, 6]. The eye tracking community needs to think about the tasks that the different eye tracking data can be used for, and the technical frameworks that allow for the separation of different tasks so that one can be turned on and the other turned off. This understanding is critical to enable the responsible use of eye tracking data by platform developers, and the widespread acceptance of this technology by lay users.

## References

- 1 Chia, Pern Hui, Yusuke Yamamoto, and N. Asokan. *Is this app safe?: a large scale study on application permissions and risk signals*. Proceedings of the 21st International Conference on World Wide Web. ACM:311–320, 2012.
- 2 Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. *Our data, ourselves: Privacy via distributed noise generation*. In Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer:486–503, 2006.
- 3 Felt, A. P., Greenwood, K., and Wagner, D. *The effectiveness of application permissions*. Proceedings of the 2nd USENIX Conference on Web Application Development, 7–7, 2011.
- 4 Liebling, Daniel J., and Sören Preibusch. *Privacy considerations for a pervasive eye tracking world*. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM:1169–1177, 2014.
- 5 Orlosky, Jason, Yuta Itoh, Maud Ranchet, Kiyoshi Kiyokawa, John Morgan, and Hannes Devos. *Emulation of physician tasks in eye-tracked virtual reality for remote diagnosis of neurodegenerative disease*. IEEE Transactions on Visualization and Computer Graphics, 23(4):1302–1311, 2017.
- 6 Pittaluga, F., Koppal, S. J., and Chakrabarti, A. *Learning privacy preserving encodings through adversarial training*. arXiv preprint arXiv:1802.05214, 2018.
- 7 Tam, Jennifer, Robert W. Reeder, and Stuart Schechter. *I’m Allowing What? Disclosing the authority applications demand of users as a condition of installation*. Technical report, 2010.
- 8 Wang, Na, Jens Grossklags, and Heng Xu. *An online experiment of privacy authorization dialogues for social applications*. Proceedings of the Conference on Computer Supported Cooperative Work, ACM:261–272, 2013.



## 5.14 Gaze-driven Education: Sensing, Understanding, Intervention, and Adaption

*Radu Jianu (City – University of London, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Radu Jianu

**Main reference** Radu Jianu, Sayeed Safayet Alam: “A Data Model and Task Space for Data of Interest (DOI) Eye-Tracking Analyses”, IEEE Trans. Vis. Comput. Graph., Vol. 24(3), pp. 1232–1245, 2018.

**URL** <http://dx.doi.org/10.1109/TVCG.2017.2665498>

### 5.14.1 Abstract

Cheap yet reliable eye trackers now let us collect gaze-data that is unprecedented in scale and diversity. This opens up novel opportunities to advance learning. The new data can help us understand how students learn and how to design more effective learning materials, as well as provide a way to track learners’ progress and provide tailored feedback. To take advantage of these opportunities a series of challenges need to be tackled.

### 5.14.2 Introduction

Reliable eye trackers have become sufficiently affordable that they can now be fitted to regular workstations. This opens up novel opportunities to advance learning. We could collect extensive naturalistic gaze-data from people using computers to learn in schools, universities, and homes. Interpreting such data could: (i) advance our understanding of how students learn; (ii) inform the design of more effective visual learning materials; (iii) inform instructor interventions by capturing what students look at and missed during particular learning sessions; (iv) and support the design of novel learning systems that adapt automatically to enhance the learning experience of students.

### 5.14.3 Envisioned Challenges

Reaping these benefits hinges on addressing several research challenges. First, we need to collect and store data from many users learning over long periods of time and we need to be able to interpret and make sense of such data effectively. The power of the proposed approach comes from capturing and mining data from hundreds and perhaps thousands of people spending many hours learning using computers in naturalistic settings. In contrast, traditional eye tracking studies dealt with data collected in carefully controlled experiments and at much smaller scales (e.g., a few minutes’ worth of gaze-data collected from several tenths of participants). The methods used to collect, store, and interpret gaze-data collected in traditional experiments cannot scale to and cope with the amount and diversity of gaze-data we would collect in naturalistic settings.

Second, we need to determine how to use gazes collected from single and multiple learners to provide personalized recommendations and feedback. Consider instructors who need to assess a student’s learning and provide guiding feedback and recommendations in a timely manner and with minimal overhead. They need visual and analytic tools that can summarize students’ gazing behavior and capture deviations from normative learning behavior. In other words, we need specialized tools that can support effective diagnostics and interventions in learning.

Finally, we need to find ways in which the next generation of learning systems can use eye tracking to automatically adapt to the needs and progress of individual learners. Specifically, we need to determine how to link individual and multiple learners’ gaze-data to useful adaptations of learning systems, and how these adaptations can be shown to learners in ways that are conducive to learning.



#### 5.14.4 Envisioned Solutions

A possible solution to the collection and interpretation of naturalistic eye tracking data lies in the related concepts of semantic areas of interest (AOIs) [1] and data of interest (DOIs) [4]. If learners watch digital content, we can easily match their momentary gaze-points to specific content-items shown on the screen automatically and in real-time [4]. Examples of content-items include concrete definitions, examples, or illustrative images present in the learning material. To support analysis, such content-items can be annotated with descriptive attributes to capture, for example, the type of learning idiom (e.g., ‘definition’, ‘example’, ‘exercise’), the type of visual representation (e.g., ‘text’, ‘illustration’, ‘animation’), or which learning concept they refer to (e.g., ‘variable’, ‘loops’, ‘expressions’).

In this way, an individual student’s learning behavior can be captured as a collection of richly-annotated content-items viewed over time. This would facilitate novel data-centric interpretations and analyses. Education researchers and instructors could explore their students’ attention data at a level of abstraction that relates to the semantics of the learning materials. For example, a researcher could easily ask if there’s a correlation between learning performance, as indicated perhaps by a quiz, and the type of content students look at (e.g., “Do effective learners look at definitions or examples more?”, “Do they focus on text or images?”). Alternatively, an instructor could check whether a student skipped too quickly over a particular learning concept or whether they systematically don’t pay attention to definitions. The instructor could then provide targeted recommendations to that student.

To build adaptive learning systems that track learners’ attention and progress to provide automatic feedback and adaption, we can draw inspiration from existing research into recommendation systems. A useful distinction is that between content-based filtering and collaborative filtering. A system could look at a single learner’s performance and adapt based on their individual learning profile (e.g., visual learner vs. verbal learner) and learning progress. Or, a system could mine gaze-data from many learners to infer prototypical learner profiles and effective learning patterns, then match individual learners to such prototypes and guide them along personalized learning paths.

It’s important that adaptive responses appear coherent and unobtrusive to learners. It’s useful to distinguish between explicit and implicit feedback. Explicit feedback could take the form of clearly distinguishable messages that recommend a course of action to the learner (e.g., revisit a particular definition). Implicit feedback could take the form of subtle, unobtrusive changes in the learning interface. Examples include repeating and possibly rephrasing concepts when the system detects a learner glanced over them, or changing the saliency or positioning of particular learning content based on a learner’s reading habits. The design space of implicit feedback is broad and worth investigating.

Incipient research efforts in the directions outlined above already exist. Jianu and Blascheck explored gaze analyses that build on the use of semantic AOIs and DOIs [1, 4]. Conati et al.’s work on adaptive learning systems and gaze-adaptive interfaces provides a valuable stepping stone [2, 3, 5]. However, such efforts are still relatively isolated and more work is needed to fulfill the vision outlined above.

## References

- 1 Blascheck, T., Burch, M., Raschke, M. & Weiskopf, D. Challenges and perspectives in big eye-movement data visual analytics. *Big Data Visual Analytics (BDVA)*, IEEE, 2015.
- 2 D. Bondareva, C. Conati, R. Feyzi-Behnagh, J.M. Harley, R. Azevedo, and F. Bouchet. *Inferring learning from gaze data during interaction with an environment to support self-regulated learning*. International Conference on Artificial Intelligence in Education, Springer:229–238, 2013.
- 3 C. Conati, C. Merten, S. Amershi and K. Muldner. *Using eye tracking data for high-level user modeling in adaptive interfaces*. National Conference on Artificial Intelligence, 22(2):1614–1617, 2007.
- 4 R. Jianu and S.S. Alam. *A Data Model and Task Space for Data of Interest (DOI) Eye-Tracking Analyses*. IEEE Transactions on Visualization and Computer Graphics, 24(3):1232–1245, 2018.
- 5 S. Kardan and C. Conati. *Comparing and combining eye gaze and interface actions for determining user learning with an interactive simulation*. In International Conference on User Modeling, Adaptation, and Personalization, Springer:215–227, 2013.

## 5.15 From Lab to the Real World: Eye Tracking Grows Up

*Enkelejda Kasneci (Universität Tübingen, DE) and Michael Raschke (Blickshift GmbH – Stuttgart, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Enkelejda Kasneci and Michael Raschke

**Main reference** Tanja Blascheck, Michael Burch, Michael Raschke, Daniel Weiskopf: “Challenges and Perspectives in Big Eye-Movement Data Visual Analytics”, in Proc. of the Big Data Visual Analytics, BDVA 2015, Hobart, Australia, September 22–25, 2015, pp. 17–24, IEEE, 2015.

**URL** <http://dx.doi.org/10.1109/BDVA.2015.7314288>

Visual search behavior plays a key role in our ability to complete everyday activities. Especially during the last decade, eye tracking technology has been increasingly employed in numerous research studies across several application domains to analyze the eye movements of users. Therefore, the eye tracking community has been intensively working towards methods to move the analysis of eye movements out of the laboratory to encompass activities in the real world. Probably one of most interesting application domains in the last decade has been the driving scenarios, for which numerous studies have investigated eye movements of drivers to identify deficits in visual search patterns or types of hazardous situations that may cause accidents. In addition, in the autonomous driving context, eye tracking has been considered as a non-invasive way for driver observation.

Related to application, few months ago in May 2018, Elon Musk tweeted that eye tracking is an ineffective technology for driving assistance<sup>2</sup> systems. In contrast to this statement, there are three main arguments for the effectiveness of eye tracking: (1) While the costs for the eye tracking hardware are continuously decreasing, (2) eye tracking has developed to a pervasive technology. Moreover, (3) we expect a break through regarding the application of eye tracking related to an emerging combination of this technology with perceptual models, advanced machine learning methods and big data technologies.

First, the costs. In 2009, the cost for lab-based eye trackers have been in the range of more than ten thousand Dollars. Since then, the prices have shown a steep drop and

<sup>2</sup> <https://twitter.com/elonmusk/status/996102919811350528>

will further drop in the near future. Moreover, we believe that within the next ten years, eye tracking will become a widely used standard sensor. Furthermore, in our opinion, eye tracking will have the first break through into the mass market in the automotive sector, followed by VR/AR applications, MedTech, and education.

Especially, the automotive industry forces currently the development of a robust eye detection. Current state of the art eye tracking hardware is not robust with regard to noise and especially with regard to changing illumination conditions [4, 5]. Since current corneal reflection methods are based on infrared LEDs and sensors and sunlight also emits infra-red parts, the quality of eye detection decreases outside laboratory conditions. However, a growing number of new methods to cope with these challenges have been presented. Many of them are using machine learning algorithms to overcome the light condition problem [6, 7]. In light of these recent developments, we believe that eye tracking technology will mature within the next few years and be applicable to such scenarios.

In contrast to the previously mentioned tweet by Musk, eye tracking has shown to be a promising technology in the context of autonomous driving. The next step towards the fully automated driving is the level of conditional automation, where the autonomous system controls the vehicle for a limited time interval. Some recent work, where eye tracking has been employed to observe the driver, has indicated the effectiveness of this technology to ensure the take-over readiness of the driver in critical situations [2, 3]. More specifically, [3] have proposed the first driver assistance system able to classify the take-over readiness of a driver in conditionally automated driving scenarios. This system works preemptively and at high accuracy, where the driver is warned in advance if a low take-over readiness is to expect.

The third prediction of this position paper is that through the application of machine learning methods there will be powerful analysis methods for eye tracking data in the future [1]. These analysis methods will lead to a quantum leap in the development of perception models for a pervasive understanding of the user's visual perception. Recorded eye movements are the input data for calculating probabilities about which visual objects users have recognized in their environment and which objects they have not seen [9, 11]. As soon as such reliable perception models become available and applicable to online scenarios, current challenges in the realm of human-machine interaction will be solved. In addition, in VR and AR, foveated rendering will not only help to decrease power consumption through reduced computational resources, but will enable a realistic rendering of natural scenes and improve user experience. Just by considering these two specific scenarios (driving and VR/AR), we believe that in the near future the eye movements of millions of users will be recorded and analyzed continuously.

The upcoming combination of new hardware solutions, algorithms from artificial intelligence and big data technologies will change the way how humans interact with computers in a fundamental way. User interfaces will be personalized and machines will adapt the human way how to perceive an environment and will learn to empathically interact with their users. A next step will be that machines will use these learnings from human vision itself to optimize their visual perception and artificial thinking processes.

On the way towards this new human-machine interaction paradigm, fundamental questions have to be answered on how we want to use this new technology. During the Dagstuhl seminar Ubiquitous Gaze Sensing and Interaction 2018 we started a discussion about the process of how to define guidelines for the development of eye tracking towards a broadly accepted technology by the society.

This position paper aims to motivate a continuation of this discussion and even to intensify it. We believe that especially discussions about ethical implications and issues of data privacy will be crucial for the further positive development of eye tracking technology and

its acceptance by the society. Since eye tracking will become a pervasive technology, possibly affecting millions of people, its misuse has to be avoided.

The main challenge for such perspective discussions is that to date we lack a clear picture of the technology in the future. However, many scientific prototypes reaching from new eye tracking sensors to methods from artificial intelligence to interpret eye movements have indicated what might be possible with more advanced technology. Furthermore, similar user tracking technological solutions from other fields show that big data analytics in the field of personal data can lead to misuse. In light of available solutions from related areas, a first step in the context of eye tracking will be to compare the data structure and data processing in the other fields with the eye tracking technology.

We believe that it is the time to raise awareness in our community on ethical implications of eye tracking technology and to organize a framework for discussion and working groups that might propose guidelines hand in hand with current technological developments. To start such a process, we propose to undergo the following steps:

1. First, we connect and bring together all relevant players in the community to discuss ethical implications and data privacy issues. Scientists from domains like biometric measurement and genetic research will be asked about challenges in their fields, available solutions, and guidelines. During this first step, the main goal will be to raise awareness regarding possible implications of eye tracking technology in the future society.
2. Second, we bring together a small group of key players to conduct further steps.
3. Going from this first connection of people from science and tech, the organizer group will invite other disciplines, such as philosophy, social sciences, and bio sciences to the discourse.
4. A milestone could be a dedicated event, where the social, ethical and data privacy implications will be discussed. A result of this first milestone will be a first draft of tasks, which have to be done towards the development of eye tracking applications that have a positive impact on the society.
5. The next step is to create a scientific program to study the implications of eye tracking on the society in more detail. During this creation, a manifest will be written and frequently updated to provide international guidelines of using eye tracking for the society. This manifest shall be signed by all relevant players in science and in tech community to underline its importance.

Our hope is that this research program will support the further development of eye tracking and acceptance of this promising technology in the future. We should actively influence its development and discuss this new technology with the society.

We would like to thank the Blickshift team, especially Michael Stoll, for a very detailed discussion of this topic.

## References

- 1 Blascheck, T., Burch, M., Raschke, M. & Weiskopf, D. Challenges and perspectives in big eye-movement data visual analytics. *Big Data Visual Analytics (BDVA)*, IEEE, 2015.
- 2 C. Braunagel, D. Geisler, W. Rosenstiel, E. Kasneci. *Online recognition of driver-activity based on visual scanpath classification*. IEEE Intelligent Transportation Systems Magazine, 9(4):23–36, 2017.
- 3 C. Braunagel, W. Rosenstiel, E. Kasneci. *Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness*. IEEE Intelligent Transportation Systems Magazine, 9(4):10–22, 2017.

- 4 W. Fuhl, M. Tonsen, A. Bulling, E. Kasneci. *Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art*. Machine Vision and Applications, 27(8):1275–1288, 2016.
- 5 W. Fuhl, D. Geisler, T. Santini, W. Rosenstiel, E. Kasneci. *Evaluation of state-of-the-art pupil detection algorithms on remote eye images*. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM:1716–1725, 2016.
- 6 W. Fuhl, T. Santini, G. Kasneci, E. Kasneci. *PupilNet: Convolutional Neural Networks for robust pupil detection*. arXiv preprint arXiv:1601.04902, 2016.
- 7 T. Santini, W. Fuhl, E. Kasneci. *PuRe: Robust pupil detection for real-time pervasive eye tracking*. Computer Vision and Image Understanding, 120(5):40–50, 2018.
- 8 T. Santini, W. Fuhl, E. Kasneci. *Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction*. Proceedings of the Annual ACM Conference on Human Factors in Computing Systems, ACM:2594–2605, 2017.
- 9 E. Kasneci, T. Kübler, K. Broelemann, G. Kasneci. *Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth*. Computers in Human Behavior, 68:450–455, 2017.
- 10 T. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, E. Kasneci. *SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies*. Behavior Research Methods, 49(3):1048–1064, 2016.
- 11 E. Kasneci, G. Kasneci, T. Kübler, W. Rosenstiel. *Online Recognition of Fixations, Saccades, and Smooth Pursuits for Automated Analysis of Traffic Hazard Perception*. Artificial Neural Networks, Springer:411–434, 2015.

## 5.16 Challenges in Gaze-based Intention Recognition

Peter Kiefer (ETH Zürich, CH)

License © Creative Commons BY 3.0 Unported license  
© Peter Kiefer

### 5.16.1 Introduction

Recent technological developments have made ubiquitous gaze sensing a goal realistically reachable in the near future. This enables new generations of systems that adapt to the user’s gaze in order to provide assistance. It is expected that intelligent assistance can be provided by recognizing cognitive states of a user [1]. While previous work has considered the gaze-based recognition of cognitive states such as interest [2], boredom [3], or cognitive load [4], this extended abstract discusses the recognition of a cognitive state that is on a particularly high cognitive level: intention.

### 5.16.2 Challenges

The following challenges for gaze-based intention recognition can be identified:

**Establishing a common terminology and framework:** the terminology related to cognition is not used consistently throughout the Human Computer Interaction (e.g., [5]) and eye tracking literature (e.g., [6]). This includes terms, such as, cognitive state, intention, plan, activity, action, cognitive load, goal and task. There is a need to review previous work in these and related fields to establish a common ground.

**Selection of gaze features for building the models:** different features of gaze have been suggested for gaze-based activity recognition (e.g., [7]). These need to be considered, combined, and possibly extended for inferring higher-level cognitive states.

**Bringing together short-term and long-term models:** the models for short-term prediction (i.e., in the range of seconds or milliseconds, e.g., [8]) established in the eye tracking and vision research communities need to be combined with models for longer term intention recognition and prediction (i.e., in the range of several minutes) well-known in Artificial Intelligence and Cognitive Science (e.g., [9, 10]).

**Accounting for hierarchical, parallel and interleaved intentions:** intentions can be seen as hierarchical concepts (i.e., an intention can be implemented by several sub-intentions) that occur in parallel (i.e., a subject may have several intentions at the same time) or interleaved (i.e., an intention can be ‘paused’ and superseded by some other intention for a while, but picked up again later) (e.g., refer to [11]).

**Bottom-up vs. top-down:** there is a need to re-visit classic discussions in the literature regarding the benefits, drawbacks and potential combinations of data-driven and model-driven approaches.

**Computational methods and platforms for gaining efficiency:** gaze data come at high frequency, and the acceptable time lag between the occurrence of an intention and the according assistance is small. This calls for efficient algorithms and computing platforms.

**Context-awareness:** adding context to the inference model will benefit the recognition accuracy. In particular, knowing the current situation of the user (such as, being at work or in a restaurant) will help in disambiguating which kinds of intentions are possible in that situation (e.g., [12]).

**Relation to affective states:** the relation between affective states (possibly also inferred from eye movements) and intentions requires further investigation.

**Research practices and infrastructure:** one challenge for the community consists in creating and sharing gaze datasets for different domains, annotated with intentions, which can be used for benchmarking purposes.

## References

- 1 Andreas Bulling and Thorsten O. Zander. Cognition-aware computing. *IEEE Pervasive Computing*, 13(3):80–83, 2014.
- 2 Pernilla Qvarfordt and Shumin Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, ACM:221–230, 2005.
- 3 Peter Kiefer, Ioannis Giannopoulos, Dominik Kremer, Christoph Schlieder, and Martin Raubal. Starting to get bored: An outdoor eye tracking study of tourists exploring a city panorama. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:315–318, 2014.
- 4 Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, ACM:282:1–282:13, 2018. ACM.
- 5 Kristian Lukander, Miika Toivanen, and Kai Puolamäki. Inferring intent and action from gaze in naturalistic behavior: A review. *International Journal of Mobile Human Computer Interaction*, 9(4):41–57, 2017.
- 6 Roman Bednarik, Hana Vrzakova, and Michal Hradis. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:83–90, 2012.



- 7 Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):741–753, 2011.
- 8 Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics*, 36(4):50:1–50:12, 2017.
- 9 Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P. Goldman. *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014.
- 10 John R. Anderson, Michael Matessa, and Christian Lebiere. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462, 1997.
- 11 Peter Kiefer. *Mobile Intention Recognition*. PhD Thesis, Otto-Friedrich-Universität Bamberg, Germany, Springer, 2011.
- 12 Peter Kiefer and Christoph Schlieder. Exploring context-sensitivity in spatial intention recognition. In *Workshop on Behavior Monitoring and Interpretation, 30th German Conference on Artificial Intelligence (KI-2007)*, BMI:102–116, 2007.

## 5.17 Gaze Language: A New Channel of Communication in Augmented Reality

Krzysztof Krejtz (SWPS University of Social Sciences and Humanities, PL)

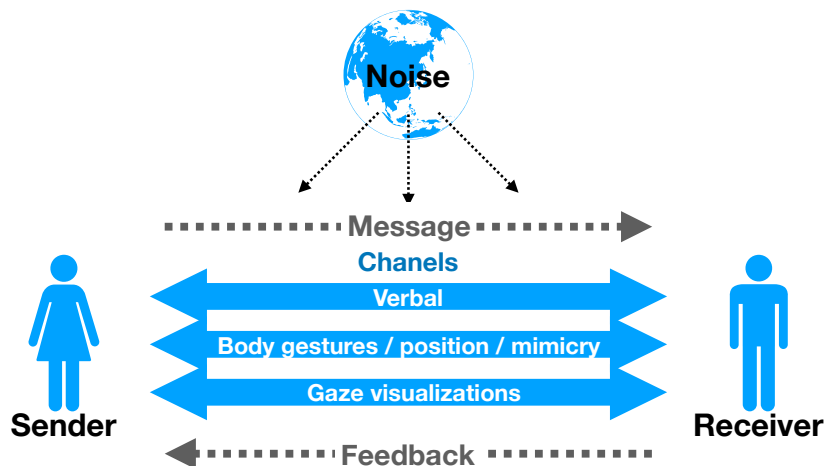
License © Creative Commons BY 3.0 Unported license  
© Krzysztof Krejtz

### 5.17.1 Introduction

This paper focuses on communication challenges during **Play & Learn** scenario in Augmented Reality. In this scenario a group of agents collaborate together by means of Augmented Reality technology to achieve a common goal. Group collaboration is defined as “a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” [7]. There are four elements of the collaboration: situational context, interpersonal interaction, mutual problem understanding, and collaboration effects [1]. The successful collaboration effects requires mutual understanding of a problem, and sharing feelings, attitudes, social norms between group members. Shared Reality Theory [4] claims that mutual understanding and feeling enhance personal connection and involvements with the group [2]. The concept of Shared Reality resulted in an enormous body of literature on collaborative work, learn and play (see e.g., [6]).

The common problem understanding and knowledge construction within the group is achieved through interpersonal dialog, which is often internalized, according the the theory of collaborative learning. The interpersonal dialog requires unrestricted communication with minimized noise and permeable channel(s) of communication [8], see Figure 2.

We postulate that the communication channels may be enriched by the constant monitoring and visualization gaze of the each group agent, see Figure 2. The challenges concerning implementation of the gaze communication are deeply rooted in the psychological and cultural context (see [3]), concerning gaze signals meaning, their influence on interpersonal relationships and self-control.



■ **Figure 2** The Shannon-Weaver communication model enriched with Gaze Communication channel.

### 5.17.2 Gaze Channel Challenges

Gaze social signaling is used by individuals in everyday interpersonal communication. Similarly to gestures, body position and facial expressions or mimicry the gaze role is supportive to the main (usually verbal) channels of communication. It supports verbal channel of communication by intentional and unintentional signaling of the sender emotional or mental state and intent (see [5]). For example, quick glance at the watch during a conversation may communicate the intent of finishing the conversation, or looking away from the interlocutor face may communicate the intent of changing the topic of the conversation.

Gaze social signaling is also used for supporting meta-communication (establishing and communicating the relationship between interlocutors) and is strongly dependent on cultural context. For instance, looking into the interlocutor's eyes may reflect a dominant position, challenging of the partner, or ensuring a good report between them. It has to be stressed that the gaze signaling, similarly to the body gesture communication, has no clear meaning without situational and cultural context.

Ubiquitous gaze monitoring and its online visualization may foster the role of gaze social signaling, changing it into a potentially important communication channel. The gaze visualizations may foster recognition of mental and emotional states between group members. And help them to establish shared reality within the group. As a result faster and more accurate solutions (in the context of a problem solving groups), deeper learning (in context of learning groups) or higher satisfaction from the game (in the context of a group play) may be expected.

Social presence awareness and self-control could also foster the focus on a task and help in establishing useful group norms (see [2]). Seeing that most students focus their visual attention on learning material may help need of learning (a norm of working hard). Noticing that most group members focus their attention on task, in problem solving groups, may trigger the desire for the solution finding.

However, applied to different social situation online display of the gaze may cause new problems. First of all, the continuous visualization of other people gaze may trigger strong social presence awareness and in turn induce higher self-control, fear of being evaluated by others. This may be especially important for socially anxious participants who may



undermine their performance because of their fear of social evaluation. Second, the ability of understanding the explicit gaze signals may be limited. The meaning of different gaze signals in group communication will need to be established in the broader process of social negotiations and learning.

### 5.17.3 Envisioned Solutions

The solutions for the sketched challenges require mainly explicit and implicit training of new social skills of communication with the use of own gaze visualization and reading of the other group members social signaling by visualized gaze. The training for teachers or experts of the use of new social signaling can be prepared and implemented as additional program in the educators curriculum. The training for broader audience can be implemented in a series of gaming apps where the basic elements of own gaze control and others gaze signal understanding may be embed and required for achieving a game goals.

The challenges need also technical solutions which, for example, will allow for momentary disengagement from the group work (switching off the following of others gaze or displaying the own gaze). That would be specially important in the first adaptation phase of the online gaze monitoring and visualization technology to the classrooms, task solving groups or multiuser games situations.

### References

- 1 Pierre Dillenbourg. What do you mean by collaborative learning? In Pierre Dillenbourg, editor, *Collaborative-learning: Cognitive and Computational Approaches*, Elsevier:1–19, 1999.
- 2 Gerald Echterhoff, E. Tory Higgins, and John M. Levine. Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science*, 4(5):496–521, 2009.
- 3 William B. Gudykunst. *Bridging differences: Effective intergroup communication*. Sage Publications Inc, 2004.
- 4 Curtis D. Hardin and E. Tory Higgins. Shared reality: How social verification makes the subjective objective. In R. M. Sorrentino and E. T. Higgins, editors, *Handbook of Motivation and Cognition, Vol. 3: The Interpersonal Context.*, Guilford Press:28–84, 1996.
- 5 J. Kunecke, A. Hildebrandt, G. Recio, W. Sommer, and O. Wilhelm. Facial emg responses to emotional expressions are related to emotion perception ability. *PLoS ONE*, 9:1–11, 2014.
- 6 J. M. Levine, L. L. Thompson, and D. M. Messick. *Shared Cognition in Organizations: The Management of Knowledge*. Organization and Management Series. Taylor & Francis, 2013.
- 7 Jeremy Roschelle and Stephanie D. Teasley. The construction of shared knowledge in collaborative problem solving. In Claire O'Malley, editor, *Computer Supported Collaborative Learning*, Springer:69–97, 1995.
- 8 C. E. Shannon, W. Weaver, R. E. Blahut, and Joseph T. Tykociner. *The Mathematical Theory of Communication*. Number Teil 11 in Illini books. University of Illinois at Urbana-Champaign Library and University of Illinois (Urbana) and Warren Weaver, 1963.

## 5.18 Communicating Visualization with Gaze-guided Storytelling

*Kuno Kurzhals*

**License** © Creative Commons BY 3.0 Unported license  
© Kuno Kurzhals

**Main reference** Chao Tong, Richard C. Roberts, Rita Borgo, Sean Walton, Robert S. Laramée, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, Xiaojuan Ma: “Storytelling and Visualization: An Extended Survey”, *Information*, Vol. 9(3), p. 65, 2018.

**URL** <http://dx.doi.org/10.3390/info9030065>

### 5.18.1 Abstract

Storytelling for visualization is important for multiple reasons: (1) communicating information to people, (2) providing guidance to understand complex data coherences better, and (3) motivating people to engage with the data. From simple infographics to complex visual analytics systems, visualization research in recent years indicates a growing interest of this topic. To this point, storytelling in visualization is realized by static summaries and animations that can be influenced by interaction. We discuss the possibilities of applying eye tracking data as an alternative interaction modality. Such a gaze-guided approach has the advantage that it can individually adapt to the users attention with or without explicit interaction.

### 5.18.2 Introduction

The dissemination of results plays an important role in all research fields. For the communication with different target audiences, the means of presentation also vary from descriptive statistics, summary reports, and visualization to support findings. With the application of visualization, it is often easier to convey facts and circumstances to a broader audience than with just statistical results that require a certain degree of expertise from the audience. This idea of visualizing numbers and concepts to tell data stories is present and commonly known from infographics [4, 7].

Over the last years, the importance of data storytelling was also emphasized for interactive visualization and visual analytics in scientific [9, 5] and information visualization [1, 6]. A recent overview of existing techniques is provided in the survey by Tong et al. [8]. According to Kosara and Mackinlay, “Presentation—specifically, its use of elements from storytelling—is the next logical step in visualization research and should be a focus of at least equal importance with exploration and analysis” [3]. To achieve this, the authors list, among other aspects, interaction, annotation, and highlighting as important future research directions. With interaction, self-running presentations can be extended to individual experiences of data exploration. With eye tracking technology, it is possible to approximate the users’ current attention focus and react to this information. Hence, our focus is on the question: “How can gaze data be incorporated to enhance storytelling for interactive data visualization?” We discuss challenges and scenarios related to this question and how they could be addressed in the future.

### 5.18.3 Envisioned Challenges

When looking at gaze as an input parameter for interaction, it has to be differentiated between eye tracking for explicit (e.g., as a mouse replacement) and implicit (e.g., an attentive display) input [2].

For explicit input, many scenarios replace the mouse by a gaze cursor as a freehand alternative. Consequently, all related issues, in particular the Midas touch problem have

to be addressed when used for interaction with a visualization. One important question here is, how should the visualization react to the current gaze input? On typical scenario could be a public display without touch interface. Here, a narrative presentation could introduce the user to the data and the related circumstances. Then, the user is free to select individual components for further exploration. A direct conversion of established desktop interfaces is not always possible, due to the mentioned issues. Especially navigation through the visualization might be cumbersome without appropriate adjustments to the gaze input.

A more promising direction is the implicit use of gaze to interact with storytelling. Here, the system can make subtle changes to the presentation without the user noticing it. We identified three scenarios that seem promising for further investigation:

**Gaze-guidance:** The visualization can emphasize specific elements to guide the user's attention during presentation. In contrast to a static design, the system can actually identify if the visual cue was sufficient or has to be intensified (e.g., by flickering highlights).

**Attendance-based adjustment:** Measuring the gaze distribution and other related metrics, the system can react with respect to the user's attendance. If the user needs time to explore the presented visualization, or if the gaze data indicates low attention, the presentation can be adjusted accordingly, for example by slowing it down, or including more of the aforementioned guidance.

**Branching stories:** Both examples before assume a fixed storyline. For some cases, it might be beneficial to provide branching paths that adjust to the user, for example to explain an issue in detail. If the system is able to derive an assessment of the subjective understanding based on the gaze data, it can derive from the main story and provide additional explanations to help with the communication of facts. Vice versa, short cuts in the storyline can be taken if the system recognizes that the user is interested in one specific aspect.

We think that explicit and implicit input play an important role for gaze-guided storytelling. While explicit input will be necessary in exploration-focused scenarios, implicit input for guidance and subtle adjustments can significantly enhance storytelling based on animated presentations.

## References

- 1 Nahum Gershon and Ward Page. What storytelling can do for information visualization. *Communications of the ACM*, 44(8):31–37, 2001.
- 2 Robert JK Jacob. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in Human-Computer Interaction*, 4:151–190, 1993.
- 3 Robert Kosara and Jock Mackinlay. Storytelling: The next step for visualization. *IEEE Computer*, 46(5):44–50, 2013.
- 4 Jason Lankow, Josh Ritchie, and Ross Crooks. *Infographics: The power of visual storytelling*. John Wiley & Sons, 2012.
- 5 Kwan-Liu Ma, Isaac Liao, Jennifer Frazier, Helwig Hauser, and Helen-Nicole Kostis. Scientific storytelling using visualization. *IEEE Computer Graphics and Applications*, 32(1):12–19, 2012.
- 6 Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010.
- 7 Mark Smiciklas. *The power of Infographics: Using pictures to communicate and connect with your audiences*. Que Publishing, 2012.
- 8 Chao Tong, Richard Roberts, Rita Borgo, Sean Walton, Robert S. Laramée, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, et al. Storytelling and visualization: An extended survey. *Information*, 9(3):65:1–65:42, 2018.

- 9 Michael Wohlfart and Helwig Hauser. Story telling for presentation in volume visualization. In *Proceedings of the 9th Joint Eurographics/IEEE VGTC Conference on Visualization*, IEEE:91–98, 2007.

## 5.19 Gaze as a Service for Ubiquitous Gaze Sensing and Augmented Reality

*David P. Luebke (NVIDIA – Charlottesville, US)*

License  Creative Commons BY 3.0 Unported license  
© David P. Luebke

### 5.19.1 Abstract

Augmented reality offers tremendous promise, but must be coupled with a minimal interface that avoids overwhelming the user with information or requiring cumbersome input. Gaze sensing will be a key component of such interfaces. When designing such an interface, we should think of gaze as a *service* to which an ecosystem of apps and devices can subscribe. Because of the uniquely personal and sensitive nature of gaze data, the community should consider the possible granularities at which that data could be provided or withheld, and the privacy implications for such systems using gaze.

### 5.19.2 Introduction

Imagine a crowded event, such as a large party or a reception at a conference, full of activity—bands playing, conversations everywhere, people known and unknown. You meet the eyes of a person across the room; they smile and start walking toward you. You unobtrusively tap a finger ring with your thumb and textual information appears directly above the person’s head. Perhaps you don’t know the person well: this floating label is a virtual name card, reminding you of their name and affiliation. Perhaps you see this person often: the label is a calendar reminder of your lunch meeting tomorrow, or the last couple of texts you exchanged a few hours ago. Perhaps you are actively working on a project with the person; a brief summary of their recent commits to your shared codebase appears. By the time you reach each other and begin talking, your shared context—for they too have this informational superpower—has been established.

Augmented reality (AR) offers the promise of superimposing information on your view of the world, with much industrial and academic research targeting a form factor ultimately as fashionable (or covert) as a pair of glasses. Meanwhile, major leaps are enabling artificial intelligence (AI) to analyze, recognize, and understand your environment. AI is, or will soon be, capable of recognizing the people in the room, the words in their conversations, the social groupings and postures, and the song the band is playing. However, still missing is the user interface to make all that superimposed information useful. Our AR-equipped partygoer does not want their view cluttered with virtual name tags hovering over every person or transcribed speech bubbles from every conversation, any more than a first responder in a crisis situation—a firefighter, fire chief, or medic—wants labels on every bystander or every distant siren or fellow responder.

A useful ARAI system should understand the specific task of the specific user, inferring and presenting only the information needed for that user and that task—while also understanding the high-level goals of the user well enough to flag important or anomalous information that

may require changing tasks. Such a system should also predict likely actions for the task and moment, and require absolutely minimal input to enact them—a design principle known as “Do-What-I-Mean” or DWIM. How can even the most advanced AI system predict the user’s attention and intent sufficiently? The crucial missing element is gaze: gaze sensing, augmented with various context both internal (such as EEG, ECG, GSR, pupillometry, pulse, etc.) and external (first-person camera or video feed, location services, user history, etc.). Such augmented gaze data—sometimes called “gaze+X”—will prove a crucial element of future augmented reality interfaces.

### 5.19.3 Challenges

This vision presents many challenges. The gaze tracking itself must be robust, working under almost all conditions (daylight, indoor, nighttime, driving through dappled light with flashes of bright sunlight and shade, etc.) for almost all users (myopic, presbyopic, nystigmatic, amblyopic, etc.). Consumer scenarios will require all-day battery life and a vanishingly unobtrusive form factor; professional scenarios (such as our first responder) may need special hardening such as thermal protection for firefighters. But beyond these hardware challenges lie important system and platform challenges such as handling of privacy and the design of gaze sensing as a service.

### 5.19.4 Envisioned Solutions

*Minimality* will be a key design principle for ARAI systems: require minimal explicit input from the user, and provide minimal output tailored to the user, situation, and task at hand. Modeling user attention and intent from gaze+X will let us minimize the input from the user. Of course gaze sensing, even the nebulous “gaze+X”, is not mind-reading. In the scenario above, the use of a simple hands-free affordance—the tap of a finger ring with the thumb of the same hand—plus the user’s gaze point, plus enough information about the object of the user’s gaze (the other person, and perhaps the fact that their eyes have just met)—gives the additional context needed for a sufficiently advanced and personalized AI system to guess the user’s intent and what options to present. Other scenarios might use speech (“Who is that?”, “What model is that yellow car over there?”, “Where does that door lead?”) or more complex tactile affordances. The point is that gaze provides context vital for reducing the input and cognitive effort required to query or instruct the system.

I believe we should think of gaze as a *service* to which apps can subscribe. Such a service would have many different levels; some examples ranging from least private and personal to most sensitive:

- Basic common gestures (probably provided by the operating system and common across all apps) for direct manipulation of UI elements, selection from menus, etc.
- Objects gazed at, again at different levels:
  - Immediately (at the moment the ring is tapped in the above scenario)
  - In recent history (last few seconds, few minutes, today, etc.)
- People gazed at
  - With a special call-out for the action of meeting somebody’s eyes, signaling interaction
- Statistics on gaze data, such as one might use to measure health, biometric identification, drowsiness, arousal, cognitive load, etc.
- Raw gaze tracks along with the accompanying first-person video feed

Such a service would exist in an ecosystem of apps, both personal and networked. This network implies a framework for handshaking and consensual sharing of gaze data by different participants in the same area; for example a teacher or trainer could use gaze from students or trainees to better evaluate their understanding. Finally, the community must articulate the privacy concerns, accounting for all the various granularities of gaze data referenced above, and propose solutions to protect privacy and educate users about the risks and benefits of sharing gaze data.

## 5.20 Basic Explicit Gaze-based Interaction Techniques in VR/MR

*Diako Mardanbegi (Lancaster University, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Diako Mardanbegi

**Main reference** Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, Hans Gellersen: “Gaze + pinch interaction in virtual reality”, in Proc. of the 5th Symposium on Spatial User Interaction, SUI 2017, Brighton, United Kingdom, October 16–17, 2017, pp. 99–108, ACM, 2017.

**URL** <https://doi.org/10.1145/3131277.3132180>

### 5.20.1 Introduction

When the first iPhone was launched, it defined a new vocabulary for interaction with computers using a simple set of gestures. Although those interaction methods have been proposed before iPhone, the gesture-driven touchscreen brought gestures to the mainstream. We envision that when the Virtual Reality (VR) and Mixed Reality (MR) devices become more ubiquitous in the near future, they define a new set of interaction techniques that will soon become the mainstream. The promising synergy between eye tracking and near-eye displays that we are observing today suggests that our eyes and in particular gaze would provide a key input for those interaction techniques and play an important role in interaction in VR/MR [2]. However, there is no standard set of gaze interaction techniques that could support basic interaction tasks such as selecting, dragging, zooming, undo, etc. In this report, we address some of the challenges that need to be more thoroughly addressed before XR devices with eye tracking functionality become ubiquitous.

We are relying so much on our smart phones as our main personal computing device that are always with us and provide fast access to information making our communication possible. In terms of form factor we see that displays have become the main component of these personal computers that not only provide visual content but also used as the main interaction channel where users can directly input their commands via manual input. We envision that in near future, smart phones are replaced with some sort of mixed-reality head-mounted devices that are capable of displaying visual content directly in front of the user’s eye. Thinking about this new form factor, our eyes seem to be playing an important role in communicating with the device not only as an input channel for visual information but also as an output channel that provide an abundance of information about the subject and the environment (e.g., context, visual attention, cognitive load, biometric, fatigue, health, etc.).

We envision a scenario in which the person wakes up in the morning and puts on his/her wearable computer and uses that for all day. The device provides relevant information such as weather, notifications, news headlines, calendar and schedule depending on where the user is looking at or what the user’s mood is. During breakfast, or other daily activities, attention analysis using eye movements could facilitate automatic recognition of the intended activity,

detection of potentially missing steps, and providing supportive information. The user is at the train station, he looks at the information shown on the train schedule display which is far. The device would then assist the user by enlarging the information and making the text readable in the field of view. In a driving or cycling task, the device provides navigation assistance by showing the map or by attention guiding that takes into account the attention information from the other drivers on the road and whether for example the user is not paying attention to the surroundings. In the shopping mall, the user can get offers and suggestions based on the information about gaze and eye movements. We could also think of many possible applications where the device facilitates interaction with others in a party or a social event. The simplest examples would be that the device provides information about other people as a memory assistant. The eye movements and gaze data recorded during the day, could be used for automatic summarization and journaling at the end of the day.

While in many of these examples gaze is used implicitly, we envision that for such a continuous use of a AR/MR technology, it's crucial to have a set of few explicit gaze-supported and hands-free interaction techniques to help performing actions such as pointing and selecting digital information or even objects in real world.

There are two unwanted things that we want to prevent from happening in this exciting moment: a) first is when the gaze-based interaction techniques proposed by the first VR/MR devices are not designed appropriately and the users start getting used to a set of nonintuitive and unnatural eye-based interaction techniques which will be hard to correct later, b) and the second condition is when gaze become more like a service that various third party apps are allowed to subscribe to that and utilize that to perform generic tasks such as selection. This may have a hugely negative impact on the overall user experience because different apps may utilize gaze differently to perform similar tasks. Similar thing happened when the Microsoft Kinect provided gesture recognition for games and the users had to often use the body movements and gestures differently to perform the same kind of task across different games affecting the overall user experience of the technology. The other main challenge is that because of the inaccuracy and unreliability of even state-of-the-art eye trackers defining explicit commands that don't work all the time could result in user frustration affecting the user experience.

The above mentioned challenges are mainly associated with the explicit use of gaze. The problem with addressing these challenges is that the VR/MR technologies are still in a premature state and many of the 3D interaction tasks are not fully defined yet. This suggests that the early VR/MR devices with integrated eye tracking should perhaps focus more on the implicit ways of using gaze (e.g., [5]) and avoid defining interaction techniques that require the users to deliberately use their eyes to control UI elements. In the meanwhile, I believe the community should identify a set of explicit commands that can be commonly used across VR/MR devices and even third-party apps for basic tasks such as selection, scrolling, zooming, etc. I also think that because of the inaccuracy issue, the first explicit gaze interaction applications should not be reliant on gaze data. There are already gaze interaction techniques that can be implemented without the need for precise gaze estimation (e.g., [3, 1, 4]) and I believe such techniques could be good candidates for explicit use. Another suggestion would be that any explicit interaction that relies on gaze should potentially come with an alternative method where users can easily switch between modalities when gaze interaction fails.

## References

- 1 Jalaliniya, S., and Mardanbegi, D. Eyegrip: Detecting targets in a series of uni-directional moving objects using optokinetic nystagmus eye movements. Proceedings of the Annual ACM Conference on Human Factors in Computing Systems, ACM: 5801–5811, 2016.
- 2 Ling, R., Mardanbeigi, D., and Hansen, D.W. Synergies between head- mounted displays and headmounted eye tracking: The trajectory of development and its social consequences. *Living Inside Social Mobile Information*, 131–156, 2014.
- 3 Vidal, M., Bulling, A., and Gellersen, H. Pursuits: spontaneous interaction with displays based on smooth pursuit eye movement and moving targets. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM:439–448, 2013.
- 4 Zhang, Y., Bulling, A., and Gellersen, H. Sideways: A gaze interface for spontaneous interaction with situated displays. Proceedings of the Annual ACM Conference on Human Factors in Computing Systems, ACM:851–860, 2013.
- 5 Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. Gaze + pinch interaction in virtual reality. Proceedings of the 5th Symposium on Spatial User Interaction. ACM:99–108, 2017.

## 5.21 Don’t Make Me Click: Immersive Information Spaces at a Glance

*Thies Pfeiffer (Universität Bielefeld, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Thies Pfeiffer

**Main reference** Jella Pfeiffer, Thies Pfeiffer, Anke Greif-Winzrieth, Martin Meißner, Patrick Renner, Christof Weinhardt: “Adapting Human-Computer-Interaction of Attentive Smart Glasses to the Trade-Off Conflict in Purchase Decisions: An Experiment in a Virtual Supermarket”, in Proc. of the Augmented Cognition. Neurocognition and Machine Learning – 11th International Conference, AC 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 10284, pp. 219–235, Springer, 2017.

**URL** [http://dx.doi.org/10.1007/978-3-319-58628-1\\_18](http://dx.doi.org/10.1007/978-3-319-58628-1_18)

Mixed reality (MR) technologies allow us to create experiences mixing digital and physical content. As current MR has a strong focus on the visual domain, it seems natural to consider eye tracking as one modality that will allow us to swiftly interact with both visualizations and objects in the physical environment alike. General availability of eye tracking is supported, as it emerged to be a key technology for enabling perceived high resolution rendering for virtual reality (VR) and augmented reality (AR) headsets (foveated/gaze-contingent rendering). With low latency eye tracking technology available in future MR systems, all the required technologies for gaze-based interaction will be readily available. The following text outlines a scenario depicting multiple uses of gaze-based interaction in the context of immersive information spaces.

### 5.21.1 Introduction: Knowledge Work in a Mixed Reality Future

When talking about digital objects, the majority of it is information (texts, pictures, videos, 3D objects) that has been digitized or digitally created. Most of this information is either linked to physical entities or has established incarnations in physical form (e.g., books, pictures, products). However, as of today, accessing and in particular manipulating these digital objects require knowledge and tools that are in most cases completely different to those that work in our physical reality. For accessing the information, we will have to bring them up to a dedicated surface on a smartphone, tablet or computer screen and then we will



have a very abstract, rather generic way of interacting with these objects with a very small set of degrees of freedom, often not appropriate to the format of the information.

In a first step, mixed reality devices will get those displays out of our world. We will no longer see dead black screens in offices or black holes on our walls in the living room. Using wearable mixed reality devices, information can be presented anywhere. With current technology prototypes, such as the Microsoft HoloLens, this vision can already be realized to some extent.

Imagine that, while doing your daily routine in the bathroom, e.g., tooth brushing, you could spend the lazy 2-3 minutes with browsing the recent headlines from your preferred news feed. By monitoring your eye movement patterns, the MR system could detect moments of mindless gaze (or anticipatory gazes at a proxy location where the headlines typically are blended in). This would trigger the presentation of the headlines, e.g., as an overlay on top of your mirror, and by monitoring your attention (what you saw/what you mean), you can get abstracts or full texts in one continuous experience without any explicit interactions (what you get).

In another situation, you are reading a text book on statistics. While going through the texts, you encounter a reference to a statistic procedure that you have no experience with. The MR system detects that you slowed down your reading process, interprets that as uncertainty and offers a brief summary about the procedure hovering beyond the text book. You follow this suggestion, read the summary and the system subsequently will provide additional information (examples, figures, etc.) as a trail, which, when followed with your gaze, will unfold a branching network of available information. Such a concept can easily be extended to libraries [1]. Similar ways to present additional information to existing physical entities can be imagined, e.g., in the area of shopping [5].

But not only receiving information, also information giving can be handled by such a system. When interacting with the personal household robot, areas that require vacuuming or dusting could be communicated to the robot just by gazing at the relevant areas. The robot, in turn, may communicate its schedule to allow the humans to adjust it to match the personal plans (e.g., not to be disturbed while reading the book on statistics) [2, 3].

### 5.21.2 Envisioned Challenges

As described above, the technologies that are required to realize this vision are around the corner. Major problems are provisioning of power, form factor and the realization of a robust tracking of eye movements that cover 99.9 percent of the population (to not exclude non-trackable persons). Major challenges are more on the human-computer interaction part: robust and generic models for basic aspects of cognitive processing have to be developed (e.g., detecting information search, information processing, reading, mindless gaze, task switches, etc.) that will form the basic atomic “user events” that can be used to trigger more complex interactions. A key interaction metaphor for such unfolding information interfaces would have to support a generic undo/backtracking command.

The MR system also is required to detect the environment in so far as to be able to blend in the digital information in an appropriate fashion (e.g., so that text is readable for the user and at an accessible position). Some of this is already rudimentarily available in systems such as the Microsoft HoloLens, however, not at a quality level that would be required for a smooth integration.

Completely missing is an extensive experience with the design of interactive objects that are physically not interactable. As no one was able to present text in mid air (except for some experiments, such as the HoloPro) or on available surfaces at a larger scale, there are

not many design guidelines addressing particular problems that go along with such designs. There is, however, knowledge in the area of the design for head-up-displays, augmented reality or advertisement boards that may be tapped in.

### 5.21.3 Envisioned Solutions

The analysis of eye movements will play a major role and be a key enabling technology for a smooth interaction with the digital and physical objects [4]. However, one should not expect a gaze-only interface, but a multi-modal interface that integrates gaze with other modalities, such as speech, gestures, and some controller-based system for high-precision inputs (e.g., using textiles). Monitoring head and gaze orientation together with the scanning of the environment and a digital display technology will already establish a robust human-in-the-loop interaction system.

The proposed solution will thus be that of a gaze-enabled smart glasses system [6, 7] connected to a cloud system with geo-, object- and action-referenced digital information. It will come with a personalized user model (up to basic cognitive and perceptual level).

### References

- 1 Thies Pfeiffer, Friedrich Summann, Jens Hellriegel, Sebastian Wolf, & Christian Pietsch. *Virtuelle Realität zur Bereitstellung integrierter Suchumgebungen*. o-bib. Das offene Bibliotheksjournal, 4(4):94–107, 2017.
- 2 Sebastian Meyer zu Borgsen, Patrick Renner, Florian Lier, Thies Pfeiffer, & Sven Wachsmuth. *Improving Human-Robot Handover Research by Mixed Reality Techniques*. VAM-HRI 2018. The Inaugural International Workshop on Virtual, Augmented and Mixed Reality for Human-Robot Interaction. 2018.
- 3 Patrick Renner, Florian Lier, Felix Friese, Thies Pfeiffer, & Sven Wachsmuth. *WYSIWICD: What You See is What I Can Do*. HRI Companion: ACM/IEEE International Conference on Human-Robot Interaction Companion, ACM:382–382, 2018.
- 4 Jonas Blattgerste, Patrick Renner, & Thies Pfeiffer. *Advantages of Eye-Gaze over Head-Gaze-Based Selection in Virtual and Augmented Reality under Varying Field of Views*. COGAIN : Workshop on Communication by Gaze Interaction, ACM:1:1–1:9, 2018.
- 5 Jella Pfeiffer, Thies Pfeiffer, Anke Greif-Winzrieth, Martin Meißner, Patrick Renner & Christoph Weinhardt. *Adapting Human-Computer-Interaction of Attentive Smart Glasses to the Trade-Off Conflict in Purchase Decisions: An Experiment in a Virtual Supermarket*. In D. D. Schmorrow and C. M. Fidopiastis (Eds.), *Augmented Cognition. Neurocognition and Machine Learning*, Springer:219–235, 2017.
- 6 Thies Pfeiffer, Steven K. Feiner, & Walterio W. Mayol-Cuevas. *Eyewear Computing for Skill Augmentation and Task Guidance*. In A. Bulling, O. Cakmakci, K. Kunze, and J. M. Rehg (Eds.), *Eyewear Computing–Augmenting the Human with Head-Mounted Wearable Assistants*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 6(1):199–203, 2016.
- 7 Thies Pfeiffer. *Smart Eyewear for Cognitive Interaction Technology*. In A. Bulling, O. Cakmakci, K. Kunze, and J. M. Rehg (Eds.), *Eyewear Computing – Augmenting the Human with Head-Mounted Wearable Assistants*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 6(1):177–177, 2016.

## 5.22 Envisioning Gaze-informed Interaction

*Pernilla Qvarfordt (FX Palo Alto Laboratory, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Pernilla Qvarfordt

**Joint work of** Pernilla Qvarfordt, Diako Mardenbegi, Shumin Zhai, David Beymer, Jacob T. Biehl, Gene Golovchinsky, Tony Dunnigan

**Main reference** Qvarfordt, P.: “Gaze-informed multimodal interaction,” In *The Handbook of Multimodal-Multisensor Interfaces*, ACM:365–402, 2017.

**URL** <https://doi.org/10.1145/3015783.3015794>

### 5.22.1 Abstract

Eye tracking technology have developed into smaller and cheaper devices. As a result, usage of eye tracking technology is moving out from the lab and into real world applications. As usage changes, designers of gaze-based interactive system needs to consider how to make the interaction reliable and efficient. Some of these options and factors are outlined here with examples of how they have been designed into some gaze-informed interactive systems.

### 5.22.2 Introduction

Already from the earliest eye tracking research, we know that our gaze is determined by motor control, perceptual and cognitive factors [1]. Our thinking and intentions are reflected in how we look at things in a particular context. In face-to-face communication, we frequently make use of other’s gaze to inform our comprehension of a situation. Although gaze tracking have the potential of being more accurate and provide more powerful interpretation of a users’ intent and interest than a human is capable of, many challenges remain to be solved.

To design gaze-informed interaction requires a good understanding of how gaze and other modalities are synchronized and coordinated. Today, this understanding is partial and specific application domains need to be explored and documented. We need to build models of specific interaction scenarios as well as general models of users’ preferences and interest. Moving towards truly gaze-informed interaction requires a joint research effort of many researchers with different background and expertise. However, to fully achieve gaze-informed interaction that is natural and smooth, care needs to take how the interaction itself is designed. Here, I outline a few principles that can be used for achieving a robust, smooth and interesting interaction with a system that uses gaze input as one input method.

### 5.22.3 Short on Gaze Interaction and Gaze-Informed Interaction

When envisioning gaze-interaction, the first thing often imagined is gaze-based pointing. The argument for gaze-interaction is that when our hands are occupied or cannot be used, we can instead use the gaze as a pointing device. This explicit or active use of gaze in interaction, is without a doubt important in particular for users who have limited abilities to use a mouse, and is what I call gaze interaction. The gaze signal is used to actively influence the outcome of the system, for instance to point at object and items. The argument against gaze pointing is that it is not natural since we use our eyes to look at our surroundings, not to point with. A mayor problem with gaze interaction is how to distinguish these two cases: looking vs. selecting. Methods range from dwell time [2] to adding additional modalities, be a button press [3, 4], a foot pedal press [5], or even performing a tooth click [6].

Gaze-informed interaction on the other hand, views the information obtained from the gaze signal as one source for understanding the users’ interaction. The way we look at the same scene can differ depending on the task we were given or how engaged we are in an

activity. Since our eyes are partly driven by our cognitive processes, we can potentially use differences displayed in eye gaze when engaged in a task or an activity and infer from this signal what those activities and tasks are to provide more appropriate context information to the user. If we can entangle the specific gaze patterns for a particular task, it may be possible to build gaze-informed interactions where it appears as the system can read the users' mind in that it can based on data from eye tracking infer user's intention, preferences or workload. This kind of application would use gaze as a completely implicit input method. However, there are many challenges for reaching that vision. One of them being that we need a really good understanding of the task and user behavior when performing the task. Since this form of interaction is implicit, rather than calling it gaze interaction, I call it gaze-informed interaction since the information contained in the gaze informs the interaction with the system.

#### 5.22.4 Principles for Reliable Gaze-based Interaction

One challenge with gaze interaction and gaze-informed interaction is the eye tracking technology. Eye tracking have improved considerably, but when moving out from the lab to the real world performance issues are amplified since the situation is changed. Lighting is very changing, the users make larger and more frequent movements; both these factors affects the reliability, accuracy and precision of the eye trackers. Although, we believe that eye tracking will become more reliable in the future, the principles outlined here can serve to provide an extra layer of reliability when design gaze-based interaction.

##### Complementing Modalities

Different modalities have different strength and weaknesses. If combined well, the resulting system can become more reliable, efficient or fun to use than each modality used alone. One such example could be MAGIC pointing [11]. In MAGIC pointing, hand and eye works together to make a pointing selection. The long movements with the mouse is performed using eye tracking, while the short precise movements are performed with the mouse. Each modality performs the action that it does with best performance, be it speed or precision.

Mouse pointing does not contain much noise, but when two input methods with fair amount of uncertainty or noise are combined, the result can be increased certainty of user's intended action. For instance, [9] used gaze information to correct errors in speech recognition. Since both speech recognition and gaze data are error prone, Zhang et al. created an N-best list from each modality and used the item highest on both lists as the final result. Slaney et al. [8] used a similar approach for verbal web browsing tasks. Text from the web-page was extracted from regions on the web page where the user looked while speaking. The large vocabulary speech recognizer's N-best list was re-scored based on the attended text. Speech recognition improved by using another noisy signal, eye tracking.

##### Fall Backs and Redundancies

The second principle is to design fall backs and redundancies for when the gaze signal fails. In gaze pointing, no fall back is provided, since it is the primary input method. If the eye tracking fails, the user cannot make any selections. How a fall back should look like depends on the system. In some cases, the fall back is simply to accept that the gaze signal is absent or is noisy. In [13], we designed a note taking system for wearable displays that used gaze to point at areas of interest within an image captured by the wearable system. The user simply looked at the area of interest, signaled the system to start recording a voice memo.

Where the user looked became the anchor of the annotation. If no gaze could be collected, or the gaze were not stable at the moment the recording started, the system would attach the annotation to the complete image. This would result in a less precise annotation, but this result would be better than not being able to make an annotation.

When designing fall backs and redundancies, it is important to consider the cost of eye tracking failures from a user perspective. One example of this is from a system we designed to support visual inspection [14, 15]. It used eye tracking to suggest regions not inspected that match the characteristics of already viewed regions. In visual inspection, finding all treat targets is the most important factor for success. If the system had failed to record an area as viewed, the cost of inspecting it again is low in comparison to missing treat target. In a different system where, for instance, speed is more important, the cost calculation is likely different and the fall back and redundancies built in would also have been designed differently.

### Understanding Context

Within a particular context, the interpretation of the gaze becomes more powerful. A look is not just a look when viewed in context, it can provide an extensive resource for interpretation the user's intention and task progress in an interactive system.

The context can be one of many things, but it is either centered around the user or around the stimuli. Within a user's gaze, multiple signals, such as the pupil size, movement speed and direction, may be extracted and used. Pupil size can be used to detect user's workload [12]. Analyzing gaze patterns, such as consecutive fixations, can provide information of higher level cognitive processes when the user is trying to connect the dots. The user also performs other actions, such as gestures, speech, mouse movements, etc. These actions analyzed together with the gaze can provide a framework for detecting task or task progress. Turning to the stimuli, it contains information as well, e.g., objects, text analysis, etc., that can provide clues to what the user attends to at a specific moment and helps to infer users' task or intention.

How user-center context indicator and stimuli-centered context indicator can work together to provide an adaptive and efficient interaction, is illustrated in a tourist information system we developed [10]. By first looking at how a remote tourist consultant provided information to a tourist using a maps as a visual aid, we identified particular gaze patterns that were telling of users' intention and interest [16]. For example, the tourist consultant often used the tourist's gaze pattern over a map to infer when a particular topic was saturated and it was time to switch to another topic. The tourist gaze patterns often served as indicators of what new topic was of interest. Specific gaze patterns was also found, for instance, the tourist often looked back and forth between two objects identified in the map before asking about distances between them. Based on these finding, we could implement a system that only used gaze to carry out the same conversation as the tourist consultant [10]. Gaze directions, duration and objects, such as bus routes, hotels and attractions, were all used to infer user's interest and specific information need as it changed over time. This example shows the power of the context, using a speech conversation and the visual information in a map, the context could be extracted and modeled so that when only using gaze, the same task could be completed.

### 5.22.5 Ethical Considerations

Gaze-interaction have clear ethical considerations. Since eye trackers collect information of when and where a person is looking is collected and analyzed, an interactive system would make use of sensitive personal data. An anecdote that I encountered when I started to work with eye tracking, was that of a famous HCI researcher testing eye trackers for usability testing and revealed that he spent considerably time looking at a beautiful woman when seemingly reading a web page. Eye tracking can reveal intimate truths about a user that he or she may not consciously be aware of. However, using eye tracking as a user input device can enrich the user interaction and increase the system performance. Balancing user privacy and system performance is important for making an interactive system using eye tracking not only compelling from a performance point of view, but also acceptable from a user privacy point of view.

Ethical and privacy needs to be addressed on all levels in interactive system design. From a privacy perspective, the storing of gaze data is highly sensitive. An interactive gaze-informed system likely does not need to have gaze data stored to perform its function, however, the system might perform better, be more accurate and make better interpretation of the gaze, if user profiles or past interactions are stored and learned from. In an interactive system, the gaze data goes through a number of analytical steps. For each step, the resulting data may be either more or less sensitive. By evaluating ethical and privacy effects on each step, the system designer may be able to find a balancing point that allows the system to retain powerful analysis while preserving the user's privacy and integrity.

### 5.22.6 Conclusions

Gaze-based interactive systems can provide a highly adaptive and unique user experience. However, reaching this goal is not without challenges. Although eye tracking technology is getting ready for challenges outside the lab, the user's behavior outside the lab can make eye tracking challenging. Technology can improve, but to build a reliable gaze interactive system, designers need to think about handling occurrences of incomplete and noisy data. Building in fallbacks, redundancies and utilizing context indicators, it is possible to design engaging gaze interactive systems.

### References

- 1 A.L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.
- 2 R.J.K. Jacob. What You Look at is What You Get: Eye Movement-based Interaction Techniques, *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, ACM:11–18, 1990.
- 3 M. Yamato, K. Inoue, A. Monden, A. and K. Torii and K.-I. Matsumoto. Button Selection for General GUIs Using Eye and Hand Together, *Proc. of the Working Conf. on Advanced Visual Interfaces*, ACM:270–273, 2000.
- 4 M. Kumar, A. Paepcke, and T. Winograd. EyePoint: Practical Pointing and Selection Using Gaze and Keyboard, *Button Selection for General GUIs Using Eye and Hand Together*, ACM:421–430, 2007.
- 5 K. Klamka, A. Siegel, S. Vogt, F. Göbel, S. Stellmach, and R. Dachsel, R. Look & Pedal: Hands-free Navigation in Zoomable Information Spaces Through Gaze-supported Foot Input, *Proceedings of the Annual ACM Conference on Multimodal Interaction*. ACM:123–130, 2015.

- 6 X. Zhao, E. D. Guestrin, D. Sayenko, T. Simpson, M. Gauthier, and Popovic. Typing with Eye-gaze and Tooth-clicks, *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:341–344, 2012.
- 7 M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. Where is “It”? Event Synchronization in Gaze-Speech Input Systems, *Proceedings of the International Conference on Multimodal Interfaces*, ACM:151–158, 2003.
- 8 M. Slaney, R. Rajan, A. Stolcke, A. and P. Parthasarathy. Gaze-enhanced speech recognition, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, IEEE:3236–3240, 2014.
- 9 Q. Zhang, A. Imamiya, K. Go, and X. Mao. Overriding Errors in a Speech and Gaze Multimodal Architecture, *Proceedings of the International Conference on Intelligent User Interfaces*, ACM:346–348, 2004.
- 10 P. Qvarfordt, and S. Zhai. Conversing with the User Based on Eye-gaze Patterns, *Button Selection for General GUIs Using Eye and Hand Together*, ACM:221–230, 2005.
- 11 S. Zhai, C. Morimoto, and S. Ihde. Manual and Gaze Input Cascaded MAGIC Pointing, *Button Selection for General GUIs Using Eye and Hand Together*, ACM:246–253, 1999.
- 12 M. Pomplun, and S. Sunkara. Pupil dilation as an indicator of cognitive workload in human-computer interaction, *Proceedings of the International Conference on HCI*, 542–546, 2003.
- 13 D. Mardenbegi, and P. Qvarfordt. Creating Gaze Annotations in Head Mounted Displays, *Proceedings of the International Symposium on Wearable Computers*, ACM:161–162, 2015.
- 14 P. Qvarfordt, J. T. Biehl, G. Golovchinsky, T. Dunningan. Understanding the Benefits of Gaze Enhanced Visual Search, *Proceedings of the Symposium on Eye Tracking Research & Applications*, ACM:283–290, 2010.
- 15 P. Qvarfordt. Gaze-informed multimodal interaction, *The Handbook of Multimodal-Multisensor Interfaces*, ACM:365–402, 2017.
- 16 P. Qvarfordt, D. Beymer, and S. Zhai. RealTourist – a study of augmenting human-human and human-computer dialogue with eye-gaze overlay, *Proceedings of INTERACT*, Springer:767–780, 2005.

## 5.23 Detecting Mindless Gaze

Martin Raubal (ETH Zürich, CH)

**License** © Creative Commons BY 3.0 Unported license  
© Martin Raubal

**Main reference** Erik D. Reichle, Andrew E. Reineberg, Jonathan W. Schooler: “Eye Movements During Mindless Reading,” *Psychological Science*, Vol. 21(9), pp. 1300–1310, 2010.

**URL** <https://doi.org/10.1177/0956797610378686>

### 5.23.1 Abstract

Eye-tracking data contains mostly fixations, eye movements that stabilize over a stationary object of interest for a certain temporal duration [1]. Thresholds for determining fixations are arbitrary (about 100ms) and we assume that during a fixation people perceive an object meaningfully, which allows us to infer their cognitive processes [3]. This is an assumption though and the question is how often do people fixate objects ‘mindlessly’ (looking through an object or daydreaming), i.e., they fixate only the ‘syntactic Area Of Interest’ but do not relate to its semantics. It is important to detect such mindless gazes because otherwise we would incorrectly infer meaning and cognitive processes.



### 5.23.2 Introduction

Eye-tracking data contains mostly fixations, eye movements that stabilize over a stationary object of interest for a certain temporal duration [1]. Thresholds for determining fixations are arbitrary (about 100 ms) and we assume that during a fixation people perceive an object meaningfully, which allows us to infer their cognitive processes [3]. This is an assumption though and the question is how often do people fixate objects ‘mindlessly’ (looking through an object or daydreaming), i.e., they fixate only the ‘syntactic Area Of Interest’ but do not relate to its semantics. It is important to detect such mindless gazes because otherwise we would incorrectly infer meaning and cognitive processes.

### 5.23.3 Challenges and Research Questions

The first challenge is to come up with a clear definition of mindless gaze. Looking through an object and therefore not perceiving the stimulus is not the same as perceiving the stimulus but semantically misinterpreting it. Can both be defined as mindless gaze? Once a clear definition is reached, several research questions could be tackled by designing an experiment for detecting mindless gaze:

- What constitutes mindless gaze and which methods are best suited for its detection? This question connects to research on eye movements during mindless reading [4].
- Is there a correlation between mindless gaze and galvanic skin response (GSR) (or other bodily measures)?
- When testing which objects people have perceived, how can one distinguish between short-term memory capacity and mindlessness?
- People may fixate objects during a time-critical task but ‘miss them semantically’. Can this be a result of mindless gaze?
- What about tasks where identifying chunks is important (such as when playing chess)? A specific fixation per se is meaningless but successful if the chunk (in chess a meaningful configuration of pieces) is perceived and correctly identified as such.

Several domains and tasks are suitable for an experiment to detect mindless gaze. The experiment must be designed in such a way that allows for testing the participants’ semantic interpretation of fixated objects. Objects that were fixated for a certain amount of time but which participants cannot remember afterwards or attach meaning to, may be classified as belonging to ‘syntactic AOIs’ rather than ‘semantic AOIs’. This allows for distinguishing between meaningful and meaningless AOIs in the sense that the former are being utilized for solving the task at hand. One could, for example, imagine the scenario of an emergency center, where people must solve a cartographic map task [2] under time pressure. The type of task is important, therefore we expect different results depending on whether people must solve a concrete problem versus only explore an area.

It will be interesting to see whether the lack of connections of fixations to ‘interpreted objects’ is sufficient to identify mindless gaze or whether such detection requires data triangulation, e.g., GSR synchronized with the fixations. One can envision several potential application areas for mindless gaze identification, such as learning and education to detect whether pupils are studying or daydreaming.

### References

- 1 Duchowski, A. *Eye Tracking Methodology: Theory and Practice*. Springer, 2017.
- 2 Göbel, F., P. Kiefer, I. Giannopoulos, A. Duchowski and M. Raubal *Improving map reading with gaze-adaptive legends*. Proceedings of the Symposium on Eye Tracking Research & Applications, ACM:1–9, 2018.



- 3 Just, M. and P. Carpenter *A theory of reading: From eye fixations to comprehension*. Psychological Review, 87(4):329–354, 1980.
- 4 Reichle, E., A. Reineberg and W. Schooler *Eye Movements During Mindless Reading*. Psychological Science, 21(9):1300–1310, 2010.

## 5.24 Challenges and Opportunities of Gaze Sensing in Pervasive Visual Analytics

*Daniel Weiskopf (Universität Stuttgart, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Daniel Weiskopf

**Main reference** Kuno Kurzhals, Daniel Weiskopf: “Eye Tracking for Personal Visual Analytics”, IEEE Computer Graphics and Applications, Vol. 35(4), pp. 64–72, 2015.

**URL** <http://dx.doi.org/10.1109/MCG.2015.47>

### 5.24.1 Abstract

Visual analytics systems were traditionally designed for a professional desktop environment. However, there are recent trends to bring visual analytics to other environments, including smartphones, large display walls, or head-mounted displays. With this trend, I expect that visual analytics will become pervasive. I will discuss challenges and opportunities that come with combining pervasive visual analytics and pervasive gaze sensing, in particular, related to gaze sensing technology, gaze-based interaction, evaluation, and privacy.

### 5.24.2 Introduction

Visual analytics has been established as a new direction within visualization that focuses on interactive visual interfaces that facilitate the analysis of complex data [5]. Its strength is the combination of visualization, human-computer interaction, and often some kind of integrated and partially automated data analysis (e.g., using machine learning, data mining, or statistical methods). Originally, most visual analytics systems were designed for a professional workplace, typically in a desktop environment. However, there are recent trends to bring visual analytics to other environments. One example is immersive visual analytics, which puts visual analytics into immersive environments, e.g., with head-mounted displays.

Another scenario is visual analytics on smartphones to support the access to pervasive simulation data “out in the field”: This could be a civil engineer or an architect running and visualizing a simulation of a newly planned extension of a building within the already existing old building, for example, to assess its impact on the surround and discuss that with stakeholders like the users of the building. Another example includes visual support for first responders that need information access on mobile, robust, and lightweight devices.

Another scenario is personal visual analytics [4], i.e., visual analytics on mobile phones targeting data that is collected in a personal setting, which typically includes quantified-self applications.

All these scenarios heavily rely on non-desktop visual analytics, which comes with many challenges that are common to pervasive applications in general. In the following, I will pick a few challenges and research directions that are particularly relevant for pervasive gaze sensing and visual analytics.

### 5.24.3 Challenges and Research Directions

#### Gaze sensing technology

A reliable technological basis is a fundamental requirement that is common to virtually any pervasive gaze sensing application. This is particularly true for pervasive visual analytics because this application will often run for long time spans and in critical (work) environments, i.e., the eye tracking technology should be as unobtrusive and reliable as possible. Also calibration and re-calibration should be simple for the user, ideally, it should be implicit. While there has been much progress in this direction of research and technology, the basis is not yet completely there for pervasive visual analytics. However, with the current speed of development, it is foreseeable that this situation will change in the near to mid-term future, especially since demands come not only from pervasive visual analytics, but from almost all applications of pervasive gaze sensing.

#### Gaze-based interaction

Gaze-based interaction plays a critical role in many examples of pervasive gaze sensing. This is true for pervasive visual analytics as well, for example, for the general problem of explicit interaction by gaze, but also for indirect approaches like foveated rendering.

However, there are some specific challenges, too. For example, immersive visual analytics is still facing the problem of how to interact with the display of spatial and nonspatial data, including the selection of objects in semi-transparent renderings (such as in volume visualization of scalar fields) or abstract displays of networks or high-dimensional data. Another example is the recognition of user intent or activity, which could support user interaction indirectly; here, recognition mechanisms will have to be adapted to visual analytics, which is different from many other applications of pervasive gaze sensing. The third example is human-robot interaction: Pervasive visual analytics will play an important role in scenarios where human-robot interaction is critical, such as data display for an engineer who works in an industry 4.0 factory or at a building construction site with heavy-load robots. Here, the interaction with the visual display should be tightly linked to the interaction with the robotic system. For many of the aforementioned professional applications, we have to consider how interaction can be scaled across different types of devices. For example, the engineer mentioned above may partially collaborate with her or his colleagues in front of a display wall in a meeting room, and partially out in the factory with a head-mounted display or just a smartphone. Finally, the scenario of personal visual analytics comes with further interaction challenges because we have to support it in a casual setting.

#### Evaluation

In general, evaluation is difficult for visual analytics because it has to take into account the various aspects of user involvement: how users perceive, understand, and work with the visual representation. There is even a specific workshop series that addresses novel evaluation methodologies for visualization research: The BELIV Workshops (“Evaluation and Beyond – Methodological Approaches for Visualization”, <https://beliv-workshop.github.io>).

Fortunately, pervasive gaze sensing opens up new opportunities for evaluating visual analytics. While eye tracking has been used as a tool for visual analytics research in general [2], pervasive analytics and gaze sensing will come with additional challenges and opportunities: Can gaze sensing serve as a reliable means of quantifying the effectiveness and efficiency of visual analytics? How can we analyze and understand complex and massive gaze data

collected during in-the-wild or longitudinal experiments? The latter question leads to the problem of data analysis. Here, visual analytics could play a complementary role—as a means of visually analyzing gaze data [1]. However, the unconstrained settings of pervasive gaze sensing lead to hard data analysis issues that will require us to include the analysis of the visual context surrounding the user [3].

### Privacy

Pervasive gaze sensing comes with privacy issues in general because extensive data is collected from the user, but also her or his environment, i.e., others who might be recorded with pervasive camera systems. For pervasive visual analytics, this has also sociological and legal issues because it is often used in a professional setting at the workplace. These issues are complemented by the ones that touch the private sphere in the context of personal visual analytics [4].

### References

- 1 T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum*, 36(8):260–284, 2017.
- 2 K. Kurzhals, B. D. Fisher, M. Burch, and D. Weiskopf. Eye tracking evaluation of visual analytics. *Information Visualization*, 15(4):340–358, 2016.
- 3 K. Kurzhals, M. Hlawatsch, C. Seeger, and D. Weiskopf. Visual analytics for mobile eye tracking. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):301–310, 2017.
- 4 K. Kurzhals and D. Weiskopf. Eye tracking for personal visual analytics. *IEEE Computer Graphics and Applications*, 35(4):64–72, 2015.
- 5 J. J. Thomas and K. A. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

## 6 Conclusion and Outlook

Ubiquitous Gaze Sensing and Interaction turned out to be an engaging Dagstuhl Seminar bringing together researchers and industry with multiple perspectives and backgrounds. The intensity of discussions and the willingness to continue the discussions and create new research partnerships were evident. Several topics attracted much of the participants’ attention. In particular, popular discussions were on Data Privacy and Gaze + X. These two topics were discussed in more than one break-out session with different configurations of researchers and sparked collaborations on papers and idea exchange over traditional academic fields. From this perspective, we achieved what we set out to do when planning this seminar.

The final discussion of the seminar was how to continue the discussions sparked at this seminar and how to include more researchers than those present at these discussions. A number of ideas were put forth, these included organizing a special issue on the topic of Gaze + X, conference workshops to invite new researchers from different fields to continue the discussion, and finishing the papers that were getting started during the workshop.

## Participants

- Amy Alberts  
Tableau Software – Seattle, US
- M. Stella Atkins  
Simon Fraser University –  
Burnaby, CA
- Roman Bednarik  
University of Eastern Finland –  
Joensuu, FI
- Hans-Joachim Bieg  
Robert Bosch GmbH –  
Stuttgart, DE
- Maria Bielikova  
STU – Bratislava, SK
- Leslie Blaha  
Pacific Northwest National Lab. –  
Richland, US
- Tanja Blascheck  
INRIA Saclay – Orsay, FR
- Andreas Bulling  
MPI für Informatik –  
Saarbrücken, DE
- Lewis Chuang  
LMU München, DE
- Andrew Duchowski  
Clemson University, US
- Sara Irina Fabrikant  
Universität Zürich, CH
- Nina Gehrler  
Universität Tübingen, DE
- Hans Gellersen  
Lancaster University, GB
- Kenneth Holmqvist  
Universität Regensburg, DE
- Eakta Jain  
University of Florida –  
Gainesville, US
- Radu Jianu  
City – University of London, GB
- Enkelejda Kasneci  
Universität Tübingen, DE
- Peter Kiefer  
ETH Zürich, CH
- Krzysztof Krejtz  
SWPS University – Warsaw, PL
- Kuno Kurzhals  
Universität Stuttgart, DE
- David P. Luebke  
NVIDIA – Charlottesville, US
- Radoslaw Mantiuk  
West Pomeranian Univ. of  
Technology – Szczecin, PL
- Diako Mardanbegi  
Lancaster University, GB
- Thies Pfeiffer  
Universität Bielefeld, DE
- Pernilla Qvarfordt  
FX Palo Alto Laboratory, US
- Michael Raschke  
Blickshift GmbH – Stuttgart, DE
- Martin Raubal  
ETH Zürich, CH
- Laura Trutoiu  
Magic Leap – Seattle, US
- Daniel Weiskopf  
Universität Stuttgart, DE



# Discipline Convergence in Networked Systems

Edited by

Yungang Bao<sup>1</sup>, Lars Eggert<sup>2</sup>, Simon Peter<sup>3</sup>, and Noa Zilberman<sup>4</sup>

<sup>1</sup> Chinese Academy of Sciences – Beijing, CN, [baoyg@ict.ac.cn](mailto:baoyg@ict.ac.cn)

<sup>2</sup> NetApp Deutschland GmbH – Kirchheim, DE, [lars@netapp.com](mailto:lars@netapp.com)

<sup>3</sup> University of Texas – Austin, US, [simon@cs.utexas.edu](mailto:simon@cs.utexas.edu)

<sup>4</sup> University of Cambridge, GB, [noa.zilberman@cl.cam.ac.uk](mailto:noa.zilberman@cl.cam.ac.uk)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18261 “Discipline Convergence in Networked Systems”. This seminar explored emerging networked system design approaches, seeking to increase performance, efficiency and security through the convergence of disciplines: compute, storage and networking. With technologies such as network function virtualization (NFV) having started the convergence of computing technologies and networking technologies, this seminar discussed new research directions to embrace the convergence of disciplines that used to be mainly isolated in the past.

**Seminar** June 24–29, 2018 – <http://www.dagstuhl.de/18261>

**2012 ACM Subject Classification** Networks → Middle boxes / network appliances, Networks → Data center networks, Computer systems organization → Cloud computing, Computer systems organization → Data flow architectures, Hardware → Emerging architectures

**Keywords and phrases** Big data, cloud computing, computer architecture, networked systems, rackscale computers

**Digital Object Identifier** 10.4230/DagRep.8.6.149

**Edited in cooperation with** Max Plauth


## 1 Executive Summary

*Yungang Bao (Chinese Academy of Sciences – Beijing, CN)*

*Lars Eggert (NetApp Deutschland GmbH – Kirchheim, DE)*

*Simon Peter (University of Texas – Austin, US)*

*Noa Zilberman (University of Cambridge, GB)*

**License**  Creative Commons BY 3.0 Unported license  
© Yungang Bao, Lars Eggert, Simon Peter, and Noa Zilberman

Networked computing systems have reached a watershed, as the amount of networked-data generated by user applications exceeds the processing capability of any single computer. This requires an integrated system design, unlike the traditional layered approaches. This seminar therefore brought together experts from the operating systems, distributed systems, computer architecture, networks, storage and databases communities, to advance the state of the art in discipline convergence in networked systems.

The networking community has advanced in giant leaps, making high bandwidth networking and software-defined networking (SDN) commodity. Furthermore, the advent of network function virtualization (NFV) has started the convergence of computing technologies and networking technologies. The computing community, on the other hand, struggled to



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

Discipline Convergence in Networked Systems, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 149–172

Editors: Yungang Bao, Lars Eggert, Simon Peter, and Noa Zilberman



Dagstuhl Reports

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

overcome power density limitations, resource- efficiency and quality-of-service etc. for cloud computing as well as end host computing (or edge computing), and cannot keep up.

Revolutionary networked system design approaches are now emerging, seeking to increase performance, efficiency and security through the convergence of disciplines: compute, storage and networking. This seminar investigated both hardware and software challenges, and attempted to bridge the gaps between different communities in order to compensate the challenges in some areas with emerging breakthroughs from other areas. Over the course of the 5-day seminar, seventeen presentations were given on various aspects of data center networking. Taking the presentations as input, the workshop then broke into five working groups to discuss research aspects of operating systems, distributed systems, computer architecture, networks, storage, and databases. The talks as well as the outcome of the breakout session and the concluding statements are summarized in this report.

## 2 Table of Contents

### Executive Summary

*Yungang Bao, Lars Eggert, Simon Peter, and Noa Zilberman* . . . . . 149

### Overview of Talks

Hardware acceleration for Data Science as a means to discipline convergence <i>Gustavo Alonso</i> . . . . .	153
The Case for Labeled von Neumann Architecture (LvNA) <i>Yungang Bao</i> . . . . .	153
Distributed Join Processing on Thousands of Cores <i>Claude Barthels</i> . . . . .	154
Efficient Network Communication for Data Access <i>Angelos Bilas</i> . . . . .	154
A Programmable Framework for Validating Data Planes <i>Pietro Bressana and Noa Zilberman</i> . . . . .	155
Compiling for Future Discipline-Converged Systems <i>Trevor Carlson</i> . . . . .	155
Five Ways not to Fool Yourself <i>Tim Harris</i> . . . . .	155
PASTE: A Network Programming Interface for Non-Volatile Main Memory <i>Michio Honda</i> . . . . .	156
How to Make Profit: Building a Heterogeneous HPC System that Takes Advantage from Dynamic Electricity Pricing <i>Timo Hönig</i> . . . . .	157
Research Directions for Edge Computing and Industrial IoT <i>Dirk Kutscher</i> . . . . .	157
Efficient TCP Packet Processing <i>Simon Peter</i> . . . . .	157
Enzian: A Research Computer <i>Timothy Roscoe</i> . . . . .	158
The economics of consumer networks <i>Henning Schulzrinne</i> . . . . .	159
Future Networks Switch Architecture <i>Golan Schzukin</i> . . . . .	159
Technology and Business Challenges at Hyper-scale <i>Leendert van Doorn</i> . . . . .	160
Edge to Cloud <i>Eric Van Hensbergen</i> . . . . .	162
In Network Computing: Truths, Lies and Realities <i>Noa Zilberman</i> . . . . .	162

**Working groups**

Future Systems & Disaggregated Computing <i>Gustavo Alonso, Trevor Carlson, Felix Eberhardt, Matthias Hille, Stefan Klauck, and Max Plauth . . . . .</i>	163
Simulation & Methodologies <i>Trevor Carlson, Yungang Bao, Felix Eberhardt, Stefan Klauck, Max Plauth, Golan Schzukin, and Eric Van Hensbergen . . . . .</i>	164
Edge Computing and IoT <i>Dilma Da Silva, Dirk Kutscher, Jörg Ott, Henning Schulzrinne, Golan Schzukin, Eric Van Hensbergen, and Noa Zilberman . . . . .</i>	165
Tools for Networked Systems <i>Timo Hönig, Gustavo Alonso, Claude Barthels, Angelos Bilas, Matthias Hille, Jacob Nelson, Simon Peter, Timothy Roscoe, and Leendert van Doorn . . . . .</i>	166
Hardware/Software Co-Design – Group 1 . . . . .	168
Hardware/Software Co-Design – Group 2 . . . . .	168
Hardware/Software Co-Design – Group 3 . . . . .	169

**Panel discussions**

Concluding statements <i>Simon Peter, Gustavo Alonso, Yungang Bao, Claude Barthels, Pietro Bressana, Trevor Carlson, Dilma Da Silva, Tim Harris, Matthias Hille, Michio Honda, Timo Hönig, Sue Moon, Jacob Nelson, Dan Ports, Timothy Roscoe, Henning Schulzrinne, Golan Schzukin, Eric Van Hensbergen, and Irene Y. Zhang . . . . .</i>	170
---	-----

<b>Participants . . . . .</b>	<b>172</b>
-------------------------------	------------



### 3 Overview of Talks

#### 3.1 Hardware acceleration for Data Science as a means to discipline convergence

*Gustavo Alonso (ETH Zürich, CH)*

License  Creative Commons BY 3.0 Unported license  
© Gustavo Alonso

Computing Systems are undergoing a multitude of interesting changes: from the platforms (cloud, appliances) to the workloads, data types, and operations (big data, machine learning). Many of these changes are driven or being tackled through innovation in hardware even to the point of having fully specialized designs for particular applications. In this talk I will review some of the most important changes happening in hardware and how they are affecting data processing. I will focus on the need to redesign the entire software stack and the opportunities offered by tailoring the stack to concrete applications. Customized stacks can be a good basis and a solid motivation for discipline convergence. As examples of the latter, I will discuss use cases from our own research that include hardware acceleration for network stacks [1, 2], in-network data processing [3, 4, 5], microservers [6, 7], and machine learning in distributed systems [8].

##### References

- 1 David Sidler, Zsolt István, Gustavo Alonso. Low-latency TCP/IP stack for data center applications. FPL 2016.
- 2 David Sidler, Gustavo Alonso, Michaela Blott, Kimon Karras, Kees A. Vissers, Raymond Carley. Scalable 10Gbps TCP/IP Stack Architecture for Reconfigurable Hardware. FCCM 2015.
- 3 Louis Woods, Jens Teubner, Gustavo Alonso. Complex Event Detection at Wire Speed with FPGAs. PVLDB 2010.
- 4 Zsolt István, Louis Woods, Gustavo Alonso. Histograms as a side effect of data movement for big data. SIGMOD Conference 2014.
- 5 Louis Woods, Zsolt István, Gustavo Alonso. Ibex – An Intelligent Storage Engine with Support for Advanced SQL Off-loading. PVLDB 2014.
- 6 Zsolt István, David Sidler, Gustavo Alonso. Caribou: Intelligent Distributed Storage. PVLDB 2017.
- 7 Zsolt István, David Sidler, Gustavo Alonso, Marko Vukolic. Consensus in a Box: Inexpensive Coordination in Hardware. NSDI 2016.
- 8 Mohsen Ewaida, Gustavo Alonso. Application Partitioning on FPGA Clusters: Inference over Decision Tree Ensembles FPL 2018.

#### 3.2 The Case for Labeled von Neumann Architecture (LvNA)

*Yungang Bao (Chinese Academy of Sciences – Beijing, CN)*

License  Creative Commons BY 3.0 Unported license  
© Yungang Bao

Conventional computer architecture usually expresses and conveys software requirements to the hardware by instruction set architecture (ISA) and virtual memory mechanism, which fail to express emerging requirements such as quality-of-service (QoS) and security. To

address this challenge, we propose a new computer architecture – Labeled von Neumann Architecture (LvNA), which enables a new hardware/software interface by introducing a hardware labeling mechanism to convey software’s semantic information such as QoS and security to the underlying hardware. In this talk, I will revisit tagged architecture initially proposed in the early 1970s. Then I will present LvNA’s principles that are different from tagged architecture and significantly reduce the design and implementation complexity. Finally I will demonstrate a RISC-V based FPGA prototype (a.k.a. Labeled RISC-V) that has been already open-sourced.

### 3.3 Distributed Join Processing on Thousands of Cores

*Claude Barthels (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Claude Barthels

**Joint work of** Claude Barthels, Gustavo Alonso, Torsten Hoeffler, Timo Schneider, Ingo Müller

**Main reference** Claude Barthels, Gustavo Alonso, Torsten Hoeffler, Timo Schneider, Ingo Müller: “Distributed Join Algorithms on Thousands of Cores”, PVLDB, Vol. 10(5), pp. 517–528, 2017.

**URL** <http://dx.doi.org/10.14778/3055540.3055545>

Traditional database operators such as joins are relevant not only in the context of database engines but also as a building block in many computational and machine learning algorithms. With the advent of big data, there is an increasing demand for efficient join algorithms that can scale with the input data size and the available hardware resources. In this talk, we explore the implementation of distributed join algorithms in systems with several thousand cores connected by a high-throughput, low-latency network, show the impact and advantages of modern communication primitives such as RDMA, and discuss the importance of network scheduling.

### 3.4 Efficient Network Communication for Data Access

*Angelos Bilas (FORTH – Heraklion, GR)*

**License** © Creative Commons BY 3.0 Unported license  
© Angelos Bilas

**Joint work of** Angelos Bilas, Pilar Gonzalez-Ferez

This talk discusses our work on how networked storage protocols over raw Ethernet can achieve low host CPU overhead and increase network link utilization for small I/O requests. We first examine the latency and overhead of a networked storage protocol directly over Ethernet and we point out the main inefficiencies. Then, we examine how storage protocols can take advantage of context switch elimination and adaptive batching to reduce CPU and network overhead. We present a system that is able to achieve 14 $\mu$ s host CPU overhead on both initiator and target for small networked I/Os over raw Ethernet without hardware support. Finally, I conclude with some thoughts on the role of host CPU overhead for networked I/O and its impact for fast storage devices.

#### References

- 1 Pilar Gonzalez-Ferez and Angelos Bilas. Reducing CPU and network overhead for small I/O requests in network storage protocols over raw Ethernet. In Proceedings of the 31st International Conference on Massive Storage Systems and Technology (MSST’2015), Santa Clara, CA, USA, June 2015.

- 2 Pilar Gonzalez-Ferez and Angelos Bilas. Mitigation of NUMA and synchronization effects in high-speed network storage over raw Ethernet. *The Journal of Supercomputing*, 72(11), 4129–4159, 2016, ISSN: 1573-0484, DOI: 10.1007/s11227-016-1726-7.
- 3 Pilar Gonzalez-Ferez and Angelos Bilas. Tyche: An efficient Ethernet-based protocol for converged networked storage. In *Proceedings of the 30th International Conference on Massive Storage Systems and Technology (MSST’2014)*, Santa Clara, CA, USA, June 2014.

### 3.5 A Programmable Framework for Validating Data Planes

*Pietro Bressana (University of Lugano, CH)*

*Noa Zilberman (University of Cambridge, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Pietro Bressana and Noa Zilberman

**Joint work of** Pietro Bressana, Robert Soulé, Noa Zilberman

Due to the emerging trend of programmable network hardware, developers have begun to explore ways to accelerate various applications and services. As a result, there is a pressing need for new tools and techniques for debugging network devices. This talk presents NetDebug, a fully programmable hardware-software framework for validating and real-time debugging of programmable data planes. We describe validation use cases, compare our design to alternative solutions, and present a preliminary evaluation using a prototype implementation.

### 3.6 Compiling for Future Discipline-Converged Systems

*Trevor Carlson (National University of Singapore, SG)*

**License** © Creative Commons BY 3.0 Unported license

© Trevor Carlson

There are many issues that come up when we work to build the future datacenter that is able to reduce cost while still maintaining client SLAs. Discipline convergence, between systems and computer architecture, could provide a way forward to accomplish these goals. We propose to expose the fundamental units of potentially parallel work from the application level, as well as the system level, to the datacenter runtime. By developing applications in Domain Specific Languages (DSLs), and mapping these applications onto diverse hardware, we hope to expose potential efficiencies across the entire datacenter stack.

### 3.7 Five Ways not to Fool Yourself

*Tim Harris (Amazon – Cambridge, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Tim Harris

**URL** <https://timharris.uk/misc/five-ways.pdf>

Performance experiments are often used to show that a new system performs better than an old system, and to quantify how much faster it is, or how more efficient it is in the use of some resource. Frequently, these experiments come toward the end of a project and – at

times – seem to be conducted more with the aim of selling the system rather than providing understanding of the reasons for the differences in performance or the scenarios in which similar improvements might be expected. Mistrust in published performance numbers follows from the suspicion that we optimize what we measure or that we measure what we have already optimized.

I will talk about some of the techniques I use in performance evaluation in order to try to get a better understanding of why a system behaves as it does, and why changes I make to the system lead to differences in performance. In part, the aim of these techniques is to be able to explain the performance of the system better when presenting it to other people, but the aim is also to provide me with a better understanding of the system while working on it, and to avoid me fooling myself over why some change has some particular effect.

### 3.8 PASTE: A Network Programming Interface for Non-Volatile Main Memory

*Michio Honda (NEC Laboratories Europe – Heidelberg, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Michio Honda

**Joint work of** Michio Honda, Giuseppe Lettieri, Lars Eggert, Douglas Santry  
**Main reference** Michio Honda, Giuseppe Lettieri, Lars Eggert, Douglas Santry: “PASTE: A Network Programming Interface for Non-Volatile Main Memory”, in Proc. of the 15th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2018, Renton, WA, USA, April 9-11, 2018, pp. 17–33, USENIX Association, 2018.  
**URL** <https://www.usenix.org/conference/nsdi18/presentation/honda>

Costs of persisting data over networks have been dominated by slow access latency to disks or SSDs, and the access methods of them, causing end-to-end latency on the order of hundreds or thousands of microseconds. Therefore, networking whose RTTs over standard TCP/IP take just a few tens of microseconds, was a relatively lightweight component of the end-to-end system. However, emerging non-volatile main memory (NVMM) will change this situation, because durably writing data becomes two-three orders of magnitude faster due to not only physical media speed but also the new access methods. Therefore, network and storage stacks equally stress the end-to-end system, and tight integration of these stacks is required to design efficient systems.

We present PASTE, a new networking interface to build networked storage systems on top of it. It offers run-to-completion processing model across networking and storage layers, and true zero copy by DMA performed directly to named packet buffers on NVMM, while preserving protection and rich set of network protocols provided by the socket APIs today. We benchmark PASTE using Write-Ahead Logging and B+tree, as well as porting it to key value stores and software switch, and show PASTE significantly outperforms well-tuned Linux and the state-of-the art network stack. PASTE is a open source Linux kernel module which does not need to modify the main kernel. FreeBSD support is an ongoing effort. The work appeared in NSDI’18. [1]

#### References

- 1 Honda, Michio, Giuseppe Lettieri, Lars Eggert, and Douglas Santry. *PASTE: A Network Programming Interface for Non-Volatile Main Memory*. In 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). USENIX Association, 2018.

### 3.9 How to Make Profit: Building a Heterogeneous HPC System that Takes Advantage from Dynamic Electricity Pricing

*Timo Hönig (Universität Erlangen-Nürnberg, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Timo Hönig

**Joint work of** Timo Hönig, Christopher Eibel, Adam Wagenhäuser, Maximilian Wagner, Wolfgang Schröder-Preikschat

**Main reference** T. Hönig, C. Eibel, A. Wagenhäuser, M. Wagner, and W. Schröder-Preikschat: “How to Make Profit: Exploiting Fluctuating Electricity Prices with Albatross, A Runtime System for Heterogeneous HPC Clusters” In Proc. of the 8th International Workshop on Runtime and Operating Systems for Supercomputers (ROSS’18), Tempe, AZ, USA, April 12, 2018, pp. 4:1–4:9, ACM, New York, NY, USA, 2018

**URL** <https://doi.org/10.1145/3217189.3217193>

The ongoing evolution of the power grid towards a highly dynamic supply system poses challenges as renewable energies induce new grid characteristics. The volatility of electricity sources leads to a fluctuating electricity price, which even becomes negative when excess supply occurs. This talk discusses a runtime system for heterogeneous HPC clusters that takes advantage of dynamic electricity pricing. To ensure an energy-efficient and economic processing of HPC workloads, the system exploits heterogeneity at the hardware level and considers dynamic electricity prices for runtime decisions.

### 3.10 Research Directions for Edge Computing and Industrial IoT

*Dirk Kutscher (Huawei Technologies – München, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Dirk Kutscher

Edge- and in-Network Computing requires a rethinking of communication abstractions. This talk discusses some of the problems of current edge computing approaches with a particular focus on Industrial IoT. Instead of using client-server protocols for application-layer overlays, we discuss the possibility to conceive computing and networking holistically and propose Networked Computing Platform, providing an empowered data plane for networked computations.

### 3.11 Efficient TCP Packet Processing

*Simon Peter (University of Texas – Austin, US)*

**License** © Creative Commons BY 3.0 Unported license

© Simon Peter

**Joint work of** Simon Peter, Antoine Kaufmann, Tim Stamler, Naveen Sharma, Thomas Anderson, Arvind Krishnamurthy

TCP is widely used for client-server communication in modern data centers and across the Internet. But TCP packet handling is notoriously CPU intensive, accounting for an increasing fraction of server processing time. Techniques such as TCP segment offload, kernel bypass, and RDMA are of limited benefit for the typical small, frequent RPCs. These techniques can also compromise protocol agility, resource isolation, overall system reliability, and complicate multi-tenancy.

In this talk, I propose a unique refactoring of TCP functionality that splits processing between a streamlined fast path for common operations, and an out-of-band slow path. Protocol processing executes in the kernel on dedicated cores that enforce correctness and resource isolation. Applications asynchronously communicate with the kernel through event queues, improving parallelism and cache utilization. I show that the approach can increase RPC throughput by up to 4.1x compared to Linux. The fast-path can be offloaded to a programmable NIC to further improve performance and minimize CPU time for network processing. With hardware offload, data packets are delivered directly from application to application, while the NIC and kernel cooperate to enforce correctness and resource isolation. I show that hardware offload can increase per-core packet throughput by 10.7x compared to the Linux kernel TCP implementation.

### References

- 1 Antoine Kaufmann, Simon Peter, Naveen Kr. Sharma, Thomas Anderson, and Arvind Krishnamurthy. High Performance Packet Processing with FlexNIC. In *21st International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Atlanta, GA, USA, April 2016.
- 2 Antoine Kaufmann, Simon Peter, Thomas Anderson, and Arvind Krishnamurthy. FlexNIC: Rethinking Network DMA. In *15th USENIX Workshop on Hot Topics in Operating Systems (HotOS)*, Kartause Ittingen, Switzerland, May 2015.

## 3.12 Enzian: A Research Computer

*Timothy Roscoe (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license

© Timothy Roscoe

**Joint work of** Reto Achermann, Gustavo Alonso, David Cock, Tobias Grosser, Zsolt Istvan, Amit Kulkarni, Muhsen Owaid, Timothy Roscoe, Zeke Wang

**URL** <http://www.enzian.systems/>

Academic research in rack-scale and datacenter computing today is hamstrung by lack of hardware. Cloud providers and hardware vendors build custom accelerators, interconnects, and networks for commercially important workloads, but university researchers are stuck with commodity, off-the-shelf parts.

Enzian is a research computer being developed at ETH Zurich (in collaboration with Cavium and Xilinx) which addresses this problem. An Enzian board consists of a server-class ARMv8 SoC tightly coupled and coherent with a large FPGA (eliminating PCIe), with about 0.5 TB DDR4 and nearly 500 Gb/s of network I/O either to the CPU (over Ethernet) or directly to the FPGA (potentially over custom protocols). Enzian runs both Bareflish and Linux operating systems. Many Enzian boards can be connected in a rack-scale machine (either with or without a discrete switch) and the design is intended to allow many different research use-cases: zero-overhead run-time verification of software invariants, novel interconnect protocols for remote memory access, hardware enforcement of access control in a large machine, high-performance streaming analytics using a combination of software and configurable hardware, and much more.

By providing a powerful and flexible platform for computer systems research, Enzian aims to enable more relevant and far-reaching work on future compute platforms.

### 3.13 The economics of consumer networks

*Henning Schulzrinne (Columbia University – New York, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Henning Schulzrinne

Most of the key networking protocols and concepts have not changed in the past twenty years. Introductory textbooks written in around 2000 contain basically the same material as today's editions, with only concepts like ATM, X.25 and frame relay disappearing. Similarly, the IETF working group structure from 1992 resembles that of today. On the other hand, browsers, data centers and access networks have acquired new functionality, have emerged as a new area of study and engineering and have dramatically increased speed, respectively. The primary reason is that each of these other domains is dominated by a handful of vendors, at most, and typically makes it easy to maintain backward compatibility. In wide area networking, networks never die, they just slowly slowly drop nodes, with SS7 and fax still having significant commercial value.

Thus, we now have three key industries – rapidly growing data centers, stagnant access networks and disappearing “tier-1” interconnection networks, as transit prices have collapsed. At least in the US, almost all carriers are now trying to acquire businesses other than telecommunications, whether advertising or content. A lot of the Internet traffic has moved from commercial “common carrier” networks to private networks, operated by large CDNs or the hyperscale cloud providers.

For 20 years, carriers have fretted about becoming bit pipes. Initially, they were trying to add value to transport, but repeatedly struggled to make this commercially viable. This realization probably motivated their acquisition of content and advertising.

One of the fundamental challenges of the telecommunications industry is that while users value different kinds of bits, e.g., SMS or video, quite differently, possibly by orders of magnitude, they strongly resist having carriers impose content-based surcharges, particularly as the technical need for differentiation disappeared.

Most of the cost of running networks is management, not hardware, with roughly only 15% attributable to capital expenditures and a large fraction of that spent on civil engineering. Thus, carriers are primarily concerned about automation and reducing staff.

### 3.14 Future Networks Switch Architecture

*Golan Schzudin (Broadcom – Yakum, IL)*

**License** © Creative Commons BY 3.0 Unported license  
© Golan Schzudin

Traditional switches are facing a big challenge of how to keep doubling performance each generation within a reasonable power and cost. Failing to do that will have significant impact on networks cost and power, increasing the number of tiers.

The alternative suggested and already applied in Broadcom DNX fabric switches is to distribute complexity to the edges (TOR/NIC) and keep the fabric switch simple, allowing cost and power improvements.

The simplification includes among others Packet Processing, Traffic Management, Buffering, E2E congestion control and scheduling.



### 3.15 Technology and Business Challenges at Hyper-scale

*Leendert van Doorn (Microsoft Corporation – Redmond, US)*

License  Creative Commons BY 3.0 Unported license  
© Leendert van Doorn

#### Technology and Business Challenges at Hyper-scale

Running a datacenter at scale is all about economics. Like any other business the objective of a Hyperscaler is to increase its scale and increase margins. That means we are constantly on the look out for technologies that help us grow, reduce cost and increase the value of the services that we deliver.

However, there are two opposing forces at work in the datacenter. The cost of the infrastructure is rising faster than the price for the billable goods. So how do we deal with this? We continuously cost optimize our platform designs. We also try to get broad adoption of our platforms by open sourcing our designs within the Open Compute Platform (OCP) project. We utilize our resources more efficiently by increasing the number of VMs per node, use memory/SSD's more efficiently, etc. This goes hand in hand with more flexible customer service level agreements. We are also providing higher-level services, such as data processing and machine learning that provide higher value to our customers.

In general, we try to take advantage of commodity hardware, but we will differentiate the platforms with accelerators where we can provide value.

In this talk I gave an overview of some meta-trends, followed by specific examples for hardware and converged networking trends.

#### Meta Trends

There are a few meta trends that common to most Hyperscalers and that influence their decision making:

1. At the scale and growth we operate at, we cannot be beholden to a single supplier. If a supplier cannot deliver (for whatever reason), we cannot grow. So, we are actively exploring and enabling different suppliers. This applies to CPUs (Intel, AMD, ARM), networking, flash suppliers, memory suppliers, etc.
2. Moore's law is economically dead beyond 5nm. That doesn't mean that silicon technology won't scale, it does mean it is getting more and more expensive to do so at limited power and performance returns. While this trend is still a few years out, it does put a much greater emphasis on born in the cloud applications that take advantage of scale instead of single thread performance.
3. Another way to increase the performance and reduce the power is by taking advantage of application specific accelerators. These include GPGPU's, ML, FPGA's, ASICs, etc.
4. DRAM doesn't really scale anymore while we are putting more and more memory into a node to increase the VM density. So effectively the platform cost keeps increasing. Storage class memory may provide an interesting solution here since it has (long-term) a lower cost than DRAM and it is byte addressable but has orders of higher latency and a durability problem that needs to be overcome.
5. Disaggregation. We are still building datacenters with individual nodes while logically we think of, and manage, them as clusters/rows. Clearly there are optimization opportunities here by looking at clusters/rows more holistically.



6. Security is top of mind with all Hyperscalers. Obviously, platform security hygiene (such as secure boot) is critical, but also are issues such as secure supply chain and most recently isolation guarantees and side channel protection. If Meltdown and Specter have thought us anything, a mono culture is not a good place to be.

### Hardware Technology Trends

There are a few hardware trends that impact our decision making. Rather than going into detail what we do specially, I'm outlining some options.

1. Networking: Every Hyperscaler has some custom software defined networking (SDN) solution. This continues to be a fruitful area of innovation, especially with the convergence of SDN, virtual switch stack, control plane, RDMA, and storage. A perfect example of this development is what AWS did with Nitro (and Mellanox with Bluefield and Broadcom with Stingray). Nitro enables AWS to offload the “expensive” x86 processor by running the bookkeeping functions on a dedicated SOC thereby increasing the number of VM's they can run on the x86 processors (apparently giving them 15% additional capacity per Nitro enabled node).
2. Offloading: The CPU is an expensive resource, so like the AWS Nitro example above, should you move functions that are “data center tax” into an SOC? What about often used data movement functions such as compression and crypto should they be move into an ASIC? What about ML inference engines?
3. Storage Class Memory (SCM): As mentioned earlier, DRAM has hit a cost plateau. Can SCM replace part of the function of traditional DRAM memory?
4. Flash: All hot storage is flash memory today (cold storage uses mechanical disks and even tape). Does it make sense to rethink the way an OS uses storage in its virtual memory infrastructure and couple it much more tightly? Especially with the advent of storage class memory? That said, the cost dynamics between flash and SCM are dramatically different so I think they'll co-exist for a long time. Of course, this is back to future and its worth to revisit some of the Multics ideas.
5. Service Level Agreement (SLA). While not strictly a hardware trend, a lot of design tradeoffs throughout the entire system are directly attributable to SLA parameters. By changing these parameters you'll get more design freedom and potentially provide more value to your customers.

### Converged Networking

In the absence of the clear definition of converged networking, let me provide my own perspective on this topic. From a hyperscaler perspective there are only two interfaces into a node: networking and storage and these two are converging into a single widget. AWS Nitro is a perfect example of this convergence where an SOC provides both a networking interface and a storage interface (NVME) to the host and the guest VMs. The storage is disaggregated and in the AWS case it resides in the row (AWS EBS) although it could be distributed among the nodes or even on a remote block store depending on latency and bandwidth requirements.

With the integration of the control plane into the Smart/Secure NIC, the NIC is becoming the demarcation point between the node and the rest of the infrastructure. For all intent and purpose, it controls the node (reboot, reconfiguration, etc.), provide storage and networking services for bare metal and guest VMs.

In addition, there is a lot of interesting work going on around new types of cheaper and faster interconnects, especially at the T0 level. Examples of this are Gen-Z, new optical interfaces and increased speeds of 400Gbps and beyond.

### 3.16 Edge to Cloud

*Eric Van Hensbergen (ARM – Austin, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Eric Van Hensbergen

**Joint work of** Eric Van Hensbergen, Luis Pena, Geoffrey Blake, Chris Adeniyi-Jones, Girish Birajdar, Edmund Grimely-Evans

Arm is currently researching how to deal with the oncoming glut of data from a trillion Internet of Things (IoT) devices. We are approaching this via a converged infrastructure deploying cloud-like infrastructure for function-as-a-service (FaaS) between edge devices and the cloud for aggregation, analysis, filtering and regional pub/sub mechanisms. We are tackling the challenges in such an approach through hardware/software co-design, leveraging hardware acceleration to minimize overhead in the networking, memory, and storage subsystems while also looking for ways to streamline provisioning, allocation, and dispatch within the resulting distributed system. This has led us to looking at different tradeoffs in existing Operating System abstractions (in particular those dealing with accelerators) and existing microarchitectures which have evolved principally for time-sharing cores versus event-oriented processing dominating today's datacenters. In this talk, I lay out our overall vision for how such systems should work, results from our initial prototypes, and discuss opportunities and challenges for such an approach.

### 3.17 In Network Computing: Truths, Lies and Realities

*Noa Zilberman (University of Cambridge, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Noa Zilberman

**Joint work of** Noa Zilberman, Yuta Tokusashi, Robert Soule, Tu Dang, Pietro Bressana, Han Wang, Ki Suh Lee, Hakim Weatherspoon, Marco Canini, Fernando Pedone, Nik Sultana, Salvator Galea, David Greaves, Paolo Costa, Peter Pietzuch, Andrew W. Moore

In-network computing enables services traditionally running on servers to run within network devices, providing orders-of-magnitude performance improvements. Still, network operators remain skeptical of In-network computing. In this talk I survey several In-network computing design efforts and discuss design trade-offs and their effect on performance, power and feasibility.

#### References

- 1 H. T. Dang, P. Bressana, H. Wang, K. S. Lee, H. Weatherspoon, M. Canini, N. Zilberman, F. Pedone, and R. Soulé. P4xos: Consensus as a network service. Research Report 2018-01, USI, May 2018.
- 2 N. Sultana, S. Galea, D. Greaves, M. Wojcik, J. Shipton, R. Clegg, L. Mai, P. Bressana, R. Soulé, R. Mortier, P. Costa, P. Pietzuch, J. Crowcroft, A. W. Moore, and N. Zilberman. Emu: Rapid Prototyping of Networking Services. USENIX ATC, July 2017.
- 3 Y. Tokusashi, H. Matsutani, and N. Zilberman LaKe: An energy efficient, low latency, accelerated key-value store. CoRR abs/1805.11344, May 2018

## 4 Working groups

### 4.1 Future Systems & Disaggregated Computing

*Gustavo Alonso (ETH Zürich, CH), Trevor Carlson (National University of Singapore, SG), Felix Eberhardt (Hasso-Plattner-Institut – Potsdam, DE), Matthias Hille (TU Dresden, DE), Stefan Klauck (Hasso-Plattner-Institut – Potsdam, DE), and Max Plauth (Hasso-Plattner-Institut – Potsdam, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Gustavo Alonso, Trevor Carlson, Felix Eberhardt, Matthias Hille, Stefan Klauck, and Max Plauth

Key Observations:

- Hardware is changing quickly
- Software is changing much more slowly (if at all) because it is expensive
  - Because cloud software is new (and lacks significant legacy code), there is a chance to take advantage of these systems
  - Both development and maintenance are costly

Location of innovation:

- On-premises
  - Legacy systems using monolithic storage and compute
  - Hardware acceleration is enabled through commercial appliances
- Cloud-based
  - New applications are built to be stateless
  - Using higher-level constructs (Amazon Lambda, TensorFlow, etc.)
  - Not necessarily hardware optimized
  - Data-flow / event driven software
  - Requirements for Cloud-based infrastructure:
    - \* A critical mass of applications and users are needed to enable cloud infrastructure

Vision Statement: A disaggregated computer system, composed of interconnected compute elements (xPU (CPU, GPU, etc.) + NIC), and collected together into a Domain Specific Architecture. Given the Hardware, we could also target the high-level stack (software) using Domain Specific Languages that would map Amazon Lambda-like tasks to hardware implementations. Not all Hardware is treated equally – but, maybe they should be. CPUs have a traditional OS stack, but GPUs, FPGAs, TPUs and xPUs do not.

Open Questions:

- Do we need a system stack (Operating System) on all accelerators (making them 1st-class, standalone citizens)?
- Does each 1st-class citizen require a network connection?
- What is the role of the network?
- How important are resilience and security?

Proposal: Disaggregated 1st-class citizens on the network with storage separated from the computation. Software is constructed with dataflow in mind, given a Domain Specific Construction (DSC) of the Virtual System and Domain Specific Languages (DSLs) to describe the computation in the system.

Aspects to Investigate:

- Description and construction of the system architecture
- Show the viability of the disaggregated machine
- Build a Virtualization Driver (VD) that controls and partitions access to hardware
- Build a Virtualization Controller (VC) that brings together (connects) VDs together into a Virtual Machine (VM)
- Investigate the role of the network (NIC + Switch)
  - Simplification of the network inside the datacenter?
- Hardware aspects to investigate
  - Multi-tenant solutions
  - Hardware threads / context management
  - Context switching
  - Performance isolation
  - Security isolation
  - Virtualization
- Software aspects
  - Support legacy software on the Domain Specific Architectures
  - Provide DSLs to build and connect tasks (like Amazon Lambda)
    - \* DSLs provide optimization of software to the hardware provided
    - \* Compilers target the most optimal solution, but can also target non-optimal matches as needed

## 4.2 Simulation & Methodologies

*Trevor Carlson (National University of Singapore, SG), Yungang Bao (Chinese Academy of Sciences – Beijing, CN), Felix Eberhardt (Hasso-Plattner-Institut – Potsdam, DE), Stefan Klauck (Hasso-Plattner-Institut – Potsdam, DE), Max Plauth (Hasso-Plattner-Institut – Potsdam, DE), Golan Schzukin (Broadcom – Yakum, IL), and Eric Van Hensbergen (ARM – Austin, US)*

**License** © Creative Commons BY 3.0 Unported license  
 © Trevor Carlson, Yungang Bao, Felix Eberhardt, Stefan Klauck, Max Plauth, Golan Schzukin, and Eric Van Hensbergen

Performance Counters:

- Distributed Systems Performance Counters – Does this exist
- NICs and CPUs and GPUs have performance counters
- Datacenter Level: Operations vs. Architecture (different goals, potentially)
- Sampling device counters, new metrics for
  - SFLOW / NetFlow – Statistical Probability – Sampling vs. Tracing
  - SNMP / MIB – Management Information – Doesn't Scale (?)
  - SmartNIC / Microsoft – Datacenters might be different from traditional networking
    - \* Which flows are susceptible to drops (to see if you are provisioning correctly)
- Performance Debugging
  - Standardization in collection and aggregation and

Metrics

- What are the important metrics
- Link utilization

- Tail latency (vs. average, min, max, per flow)
- Metrics on a per-area basis
- Tail latency / (Utilization \* Throughput)
  - Is Tail Latency a QoS issue?
- Retransmission
- Utilization = Theoretical Peak in requests / second (memcached)
- Picking the metrics that are important – SLA is the only thing that is important (?)
- Can your cost of the components factor into how our costs are affected
- Negative Test – Invalid inputs (out-of-specification)
  - Can provide some insights into the bottlenecks and issues
  - Sensitivity Studies – Shows the weak spot in the system – For testing the robustness of the system
  - Orthogonal to CPI-stacks (computer architecture)
- LOGP analysis – can be useful for parallel computing.

#### Simulation

- Modeling
- Sampling
  - Multi-threaded sampling is difficult
  - There are some solutions (BarrierPoint), but they are not yet sufficient for large workloads
- Simulation
  - Parallel Simulation can be interesting – [snipersim.org](http://snipersim.org)
  - Faster simulation solutions exist (zsim), but are not as flexible.
  - Multi-level simulation + sampling

### 4.3 Edge Computing and IoT

*Dilma Da Silva (Texas A&M University – College Station, US), Dirk Kutscher (Huawei Technologies – München, DE), Jörg Ott (TU München, DE), Henning Schulzrinne (Columbia University – New York, US), Golan Schzudin (Broadcom – Yakum, IL), Eric Van Hensbergen (ARM – Austin, US), and Noa Zilberman (University of Cambridge, GB)*

**License** © Creative Commons BY 3.0 Unported license  
 © Dilma Da Silva, Dirk Kutscher, Jörg Ott, Henning Schulzrinne, Golan Schzudin, Eric Van Hensbergen, and Noa Zilberman

What is it? Who wants it? How can value be derived?

- It is not about having proprietary systems with control on tenancy and ownership
- Sectors that may have motivating applications: healthcare, automotive, industrial automation. smart cities, smart farms
  - More traditional enterprise applications such as management of automotive fleet
- Motivation
  - Latency
  - Bandwidth
  - Privacy
  - Services that leverage local data
    - \* Data ownership is all with providers; may enable local competitors
    - \* Analogy of local grocery store versus a global supermarket.

- Two possible views of the edge:
  - Carrier's pursuit for new opportunities:
    - \* Shared tenancy, general computing environment
    - \* Different countries may have different models (single versus multiple carriers)
  - Home gateway services (e.g. Comcast IoT gateway)
- For IoT applications: aggregation and local processing functionality
  - Still storing in the cloud for remote access or wider aggregation
- Service provider may want to provide free service in exchange for access to the data

What are the relevant / interesting problems?

- Identify the appropriate platform(s) and infrastructure
  - Go beyond the current stack of Xeon CPUs
  - Unique platform or diversity?
- Sharing of data and computing among different vendors/providers
- Security and serviceability much harder in a distributed model
- Runtime model
  - Event-based, serverless, micro-services instead of VMs?
  - How to advertise services to the devices, relocation. Is there a central point for fallback?
  - Implications on the type of off-load supported by networking infrastructure?
- Can we have same tools/components in the cloud and at the edge?
  - Different constraints for resource management (e.g. power, thermal constraints)
  - Exploitation of temporal/spatial locality
- How to design software to run on an edge platform
  - What is the impact of doing IoT development easier or harder?
  - How does the development environment differ from existing tools for developing cloud applications?
  - Is it going to be the App Store model?
    - \* Do we need globally unique IDs?
    - \* From login-based to identify-based model
  - Need for fine-grain data authorization schemes (split model for sharing data, e.g., some part can go to X and another part to Y)
  - Is edge computing reinventing services that work well e.g. streaming?

#### 4.4 Tools for Networked Systems

*Timo Hönig (Universität Erlangen-Nürnberg, DE), Gustavo Alonso (ETH Zürich, CH), Claude Barthels (ETH Zürich, CH), Angelos Bilas (FORTH – Heraklion, GR), Matthias Hille (TU Dresden, DE), Jacob Nelson (Microsoft Research – Redmond, US), Simon Peter (University of Texas – Austin, US), Timothy Roscoe (ETH Zürich, CH), and Leendert van Doorn (Microsoft Corporation – Redmond, US)*

**License** © Creative Commons BY 3.0 Unported license

© Timo Hönig, Gustavo Alonso, Claude Barthels, Angelos Bilas, Matthias Hille, Jacob Nelson, Simon Peter, Timothy Roscoe, and Leendert van Doorn

Definition: Converged System

- Converged system must contain something additional, not necessarily a FPGA
- Disaggregated system, no straight line from compute to memory, each system component has access to the network

## Simulation

- Q: Why do people simulate? Is it that people do not have systems of the size they consider?
- A: Systematically explore specific parts of the solution space. Building one thing represents a single point in the solution space. Simulating one thing yields a point in the solution space and the area around it.
- Things become problematic if model of simulator becomes too simple.
- No simulations presented during the solution talks, instead "toy systems" were used. Aren't we able to build realistic simulators?
- Explore and understand small system, then scale and go to bigger systems.
- Backlash on simulations: simulations of large scale system are unrealistic. You can't build a system that scales to the size of the planet if you only use simulation. Advice: do simulation /and/ a verification of the simulation.
- Problem in the systems community: no simulation at all, as second part of the message (verify simulated results) got lost and simulation-only work is not appreciated in the community.
- Recognize that you can get different things from different simulations; get first order information of a simulation for a reasonable design.
- Complexity is a problem for simulations; although we are not doing as complex simulations as nuclear reaction experiments people realize that experiments and simulations with cloud infrastructure is complex and thus, expensive.
- We don't invest enough intellectual efforts to build good simulation.
- Systems people still use their laptops and PCs for simulations, why not using a system at the scale of a supercomputer?

## Applied Simulations

- Q: What are useful tools used at company X?
- A: Snapshotting systems, analyze them from high-level down to RPC; people often do not know what data to capture (e.g. changes in soldering process during manufacturing process that leads to disc failures in racks).
- Q: Does company X share analysis data with research?
- A: No, due to concerns wrt. customer data
- Although hyper-scalers are important, their data is not necessarily required for mid-sized systems; data can be gathered from other systems, too.
- Hyper-scaler problems often are not generic and do not apply to medium size data center.
- Deployment of new systems: build racks, 1000-10000 machines, testing first by putting them into production step by step.

## Limitations

- Simulation is expensive in financial and temporal terms.
- To simulate a 500 cores system: 100us real time requires 4 days of simulation.
- Processor simulations would require 24 hours to boot Windows.
- We, the computer science community, do not think big; other fields (e.g. physics) rush out to get billions out of DoE.
- Q: Why do we not think that big?
- A1: Cultural reasons.
- A2: Different in China (taping out chips on regular basis, millions of dollars each).
- Astro physics projects are spread across many groups and use huge amounts of man-power (including CS programmers) whereas CS must exploit grad students for their research

## 4.5 Hardware/Software Co-Design – Group 1

Different Layers of Abstraction:

- Appliance
- Library (MPI/TensorFlow)
- Low-level API (Portals)
- DSLs
- Co-processor Offload (Command block)
- ISA

Architectural Considerations:

- Synchronous vs. Asynchronous
- Streaming vs. at rest
- Pipeline / Data flow
- Static vs. Dynamic Routing
- Static vs. Dynamic Accelerator Programming

Operating Systems Considerations:

- Security (Necessary but Orthogonal?)
- Provisioning/Workflow
- Performance Isolation (QoS)
- Multi-tenancy concerns

Hardware Interfaces:

- ROCEE/RDMA
- TCP/IP
- Command Block/Active Message
- FIFO
- Coherent Memory
- MMIO
- GPIO

Programming Paradigms:

- Workflow/Dataflow DSL
- Per Accelerator Class DSL

## 4.6 Hardware/Software Co-Design – Group 2

Application requirements / expectations from the network:

- HPC
- Database / Data processing
- Packet processing
- ML

Network needs to have a broad set of instructions to support the workloads mentioned above.

Need for an NISA (Network Instruction Set Arch.)

- Similar concept than ARM/x86/ etc.
- Tool chains and compilers translate high-level languages to NISA
- NISA programs can be pushed into the network
- Different network implementations can implement instructions in different ways (e.g. in different locations, such as the NIC or the switch)



Performance of these modern networks needs to become more transparent

- Network needs to advertise how fast different NISA instructions are
- Tools need to use this information and provide feedback to the user

## 4.7 Hardware/Software Co-Design – Group 3

Starting point of the discussion: how things are done today, both in the hyperscale world and in smaller deployments; and while there may be different concerns, there are a lot of shared ones too.

Optimization of three different resources:

- Programmer effort (education, complexity)
- Provider efficiency (SKU count, etc.)
- CPU efficiency (# VMs per machine)

Traditionally, systems are thought about as layered:

- Application/OS/Hypervisor/Node/Network
- Failures were essentially independent, per machine
- Latency model was simple: within-machine ( $\mu$ s), between machine (ms), the world (s)

Running things in hyperscale clouds is more complex:

- Failures are frequently correlated
- More gradations of latency
- Programs have to think about placement and other concerns to ensure their systems provide the properties they want (performance/reliability)

Perhaps we can't decompose things in such a linear way; perhaps we need to think of overlapping domains:

- Application domains
- Failure domains
- Resource allocation domains
- Security domains
- Administrative domains
- Trust domains

It is unclear what the right “unit” of processing is for in-network computation:

- Pipes/Streams are hard to process in network devices; packets are easiest
- Are “messages” even the right unit?
- If the network does paxos, what's the right unit? Is it a “message” or “round” or something else?

## 5 Panel discussions

### 5.1 Concluding statements

*Simon Peter (University of Texas – Austin, US), Gustavo Alonso (ETH Zürich, CH), Yungang Bao (Chinese Academy of Sciences – Beijing, CN), Claude Barthels (ETH Zürich, CH), Pietro Bressana (University of Lugano, CH), Trevor Carlson (National University of Singapore, SG), Dilma Da Silva (Texas A&M University – College Station, US), Tim Harris (Amazon – Cambridge, GB), Matthias Hille (TU Dresden, DE), Michio Honda (NEC Laboratories Europe – Heidelberg, DE), Timo Hönig (Universität Erlangen-Nürnberg, DE), Sue Moon (KAIST – Daejeon, KR), Jacob Nelson (Microsoft Research – Redmond, US), Dan Ports (Microsoft Research – Seattle, US), Timothy Roscoe (ETH Zürich, CH), Henning Schulzrinne (Columbia University – New York, US), Golan Schzudin (Broadcom – Yakum, IL), Eric Van Hensbergen (ARM – Austin, US), and Irene Y. Zhang (Microsoft Research – Redmond, US)*

**License** © Creative Commons BY 3.0 Unported license

© Simon Peter, Gustavo Alonso, Yungang Bao, Claude Barthels, Pietro Bressana, Trevor Carlson, Dilma Da Silva, Tim Harris, Matthias Hille, Michio Honda, Timo Hönig, Sue Moon, Jacob Nelson, Dan Ports, Timothy Roscoe, Henning Schulzrinne, Golan Schzudin, Eric Van Hensbergen, and Irene Y. Zhang

- It's all economics as usual: companies need to increase scale or margin to stay alive
- Networks are disrupting architecture, distributed systems
- Academic disciplines need to catch-up to discipline convergence
- We should have GPUs everywhere.
- Does the end of Moore's Law mean the end of hyperscalers?
  
- It was great to have folks from systems, networking, PL, etc.
- Let's include storage in the next Dagstuhl seminar.
  
- Lots of thinking for broad range of data storage
- Three areas on an interest:
  - Data encryption
  - Recovery and handling state
  - The cost of operating at large scale
  
- Everyone is struggling with what to put on GPUs / accelerators
- Next 3 years see concrete advancements (general computation, not network dependent)
  
- There's a lot more you don't know than you do know
- In network computing: some of it quite easy on the NIC, but how would you like the network to look in 10 years time?
- The partitioning between NICs and Switches in future networks
- Diverting from just focusing on the server
  
- The challenge of simulating at a large scale
- How do you build a set of DSL in the system while improving/maintaining SLA
- What are the right applications?
  
- Not clear what are the interfaces
- The order of the abstraction layers
- "Making things with the architecture"

- Security
- The current way of thinking of things in a DC is a mess
- We were used to van Neumann model or distributed systems model. Now days we do a mesh of both models.
- Too many details, mix of topics
- Need better definitions for future – need to rethink our language – otherwise will say in ad-hoc nets
- The biggest problem is that CS still works in closed communities – how do we build heterogeneous teams? Challenge for all research institutes
- The effect of the Internet – which does drive the DC. and the concerns in the Internet are different.
- What are the commonalities and differences?
- Diversity in what people call “accelerators”.
- When are accelerators applicable? What to use when? Needs more thought + share with community
- Need abstractions
- There won't a revolution but an evolution.
- Need better use cases
- How to consider non-functional properties (e.g., power)
- DCN different to HPC networking – esp. definition of interfaces
- How the network should look for DB? Where should things be implemented?
- Challenge: implementing the convergence of the systems
- Challenges for researchers: identifying the right research question for a lone researcher
- DC designed hasn't changed a lot since 2012
- New IoT programming model: missed the streams, and how they should be processed
- Past FPGA experience was bad. Possibly need to try again. . .
- There is no dream application, which is disappointing
- Opportunities: we have pocket-technologies, but not how to “communicate” between technologies / language between technologies.
- Finding good collaborators for future work
- Need to think outside the box
- Need to identify good use cases – but limit ourselves to existing use cases
- Surprising how many people care about the network
- Challenge: how to deploy solutions? Need better open source engagement
- Challenges in user experience and QoE – need solutions not just within the DC
- Computer architecture community should learn from other communities
- In comp arch 3 approaches: pipeline, caching, parallelization – can it be used by the networking community as building boxes?
- Combination of DSL and DSA
- Seen more divergence than convergence
- Getting back to the mainframe model called “the data center”
- Need a lot of specialized knowledge – will keep maintaining hyperscalers
- Lack of security

## Participants

- Gustavo Alonso  
ETH Zürich, CH
- Yungang Bao  
Chinese Academy of Sciences –  
Beijing, CN
- Claude Barthels  
ETH Zürich, CH
- Angelos Bilas  
FORTH – Heraklion, GR
- Pietro Bressana  
University of Lugano, CH
- Trevor Carlson  
National University of  
Singapore, SG
- Julian Chesterfield  
OnApp Ltd. – Cambridge, GB
- Dilma Da Silva  
Texas A&M University –  
College Station, US
- Felix Eberhardt  
Hasso-Plattner-Institut –  
Potsdam, DE
- Lars Eggert  
NetApp Deutschland GmbH –  
Kirchheim, DE
- Tim Harris  
Amazon – Cambridge, GB
- David Hay  
The Hebrew University of  
Jerusalem, IL
- Matthias Hille  
TU Dresden, DE
- Timo Hönig  
Universität Erlangen-  
Nürnberg, DE
- Michio Honda  
NEC Laboratories Europe –  
Heidelberg, DE
- Stefan Klauck  
Hasso-Plattner-Institut –  
Potsdam, DE
- Dirk Kutscher  
Huawei Technologies –  
München, DE
- Giuseppe Lettieri  
University of Pisa, IT
- Sue Moon  
KAIST – Daejeon, KR
- Jacob Nelson  
Microsoft Research –  
Redmond, US
- Jörg Ott  
TU München, DE
- Simon Peter  
University of Texas – Austin, US
- Max Plauth  
Hasso-Plattner-Institut –  
Potsdam, DE
- Dan Ports  
Microsoft Research – Seattle, US
- Timothy Roscoe  
ETH Zürich, CH
- Henning Schulzrinne  
Columbia University –  
New York, US
- Golan Schzukan  
Broadcom – Yakum, IL
- Leendert van Doorn  
Microsoft Corporation –  
Redmond, US
- Eric Van Hensbergen  
ARM – Austin, US
- Irene Y. Zhang  
Microsoft Research –  
Redmond, US
- Noa Zilberman  
University of Cambridge, GB



# 10 Years of Web Science: Closing The Loop

Edited by

Susan Halford<sup>1</sup>, James A. Hendler<sup>2</sup>, Eirini Ntoutsi<sup>3</sup>, and Steffen Staab<sup>4</sup>

1 University of Southampton, GB, [susan.halford@soton.ac.uk](mailto:susan.halford@soton.ac.uk)

2 Rensselaer Polytechnic Institute, Troy, US, [hendler@cs.rpi.edu](mailto:hendler@cs.rpi.edu)

3 Leibniz Universität Hannover, DE, [ntoutsi@kbs.uni-hannover.de](mailto:ntoutsi@kbs.uni-hannover.de)

4 Universität Koblenz-Landau, DE, [staab@uni-koblenz.de](mailto:staab@uni-koblenz.de)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18262 “10 years Web Science: Closing the Loop” that took place in Schloss Dagstuhl from 25-29 June 2018. In total, an interdisciplinary team of 22 researchers from computer science, sociology, philosophy and law gathered and discussed on the past, present and future of Web Science and what sort of actions the community should take to stay faithful to its initial mission for societal good. The role of Web Science is more critical than ever given the ever growing impact of the Web in our society.

**Seminar** June 24–29, 2018 – <http://www.dagstuhl.de/18262>

**2012 ACM Subject Classification** Information systems → Collaborative and social computing systems and tools, Information systems → World Wide Web

**Keywords and phrases** Dagstuhl Report, Web Science, Dagstuhl Perspectives Workshop

**Digital Object Identifier** 10.4230/DagRep.8.6.173

**Edited in cooperation with** Fabien Gandon, Bettina Berendt, Katharina Kinder-Kurlanda, Pinelopi Troullinou

## 1 Executive Summary

*Eirini Ntoutsi*

**License** © Creative Commons BY 3.0 Unported license  
© Eirini Ntoutsi

This Dagstuhl Seminar aimed at bringing together researchers from different disciplines related to Web Science, namely computer science, sociology, philosophy and law to discuss on future of Web Science and how it can stay faithful to its initial mission for societal good. Several recent incidents like the online psychological experiment by Facebook have provoked widespread public concern regarding the effect of such experiments and interventions and there is no agreement on expertise and ethics knowledge about how to do Web experimental research.

The Web is a complex sociotechnical system where humans and (intelligent) machines interact in unexpected ways; such hybrid societies of natural and artificial intelligence raise new challenges for Web Science which go beyond technical challenges into ethical, legal and societal implications. The role of Artificial Intelligence in these developments was discussed extensively in terms of both opportunities and risks.



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

10 Years of Web Science: Closing The Loop, *Dagstuhl Reports*, Vol. 8, Issue 06, pp. 173–198

Editors: Susan Halford, James A. Hendler, Eirini Ntoutsi, and Steffen Staab



Dagstuhl Reports

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

Based on the discussions and inputs from all participants, we have split the discussion into three main working groups:

- Working group on innovative methods for Web Science
- Working group on values
- Working group on Web Science and Artificial Intelligence

The group will continue its work in the aforementioned topics and a manifesto is foreseen to be ready by the end of the year.

## 2 Table of Contents

### Executive Summary

<i>Eirini Ntoutsis</i> . . . . .	173
----------------------------------	-----

### Overview of Talks

10 years of Web Science <i>Susan Halford</i> . . . . .	176
Why data science needs web science: reflections from recent research <i>Elena Simperl</i> . . . . .	176
Legal Aspects <i>Nikolaus Forgó</i> . . . . .	177
Social experiments in the Web: The case of Bibsonomy – technical, social and legal implications <i>Andreas Hotho</i> . . . . .	178
What can Web Science give to industry <i>Paolo Parigi</i> . . . . .	178
Web Science: What Next? <i>Ricardo Baeza-Yates</i> . . . . .	179
AI and Society <i>Wendy Hall</i> . . . . .	180
Web Science, Artificial Intelligence and Intelligence Augmentation <i>Fabien Gandon</i> . . . . .	181
New ethics for the web and for the web scientist? <i>Katharina E. Kinder-Kurlanda</i> . . . . .	185
Why formalising fairness won't fix (algorithmic) discrimination ( <i>reloaded</i> ) <i>Bettina Berendt</i> . . . . .	185
World Wide Weapons: Project Maven, Google and Web Ethics <i>Guglielmo Tamburrini</i> . . . . .	186

### Working Groups

Innovative methods for Web Science (Visualization group) <i>Katharina Kinder-Kurlanda, Claudia Müller-Birn, Lynda Hardman</i> . . . . .	187
Working Group on Values <i>Bettina Berendt, Pinelopi Troullinou, Eirini Ntoutsis</i> . . . . .	189
Working Group on Web Science and Artificial Intelligence <i>Fabien Gandon, Oshani Seneviratne, Noshir S. Contractor, David De Roure, Kemal A. Delic, Wendy Hall, Andreas Hotho</i> . . . . .	193

### Panel Discussions


Closing the Loop: a panel discussion moderated by Susan . . . . .	196
---	-----

<b>Participants</b> . . . . .	198
-------------------------------	-----

### 3 Overview of Talks

#### 3.1 10 years of Web Science

*Susan Halford (University of Southampton, UK, [Susan.Halford@soton.ac.uk](mailto:Susan.Halford@soton.ac.uk))*

License  Creative Commons BY 3.0 Unported license  
© Susan Halford

In opening this Dagstuhl seminar this talk focusses on the origins, ambition and achievements of Web Science and – building on theses – the pressing challenges that Web Science must face in the coming decade. Tracing our origins in the early 2000s to both a paradox – despite the phenomenal growth and impact of the Web it was rarely a topic of research in its own right for Computer Science – and an epiphany – that whilst the Web had begun as a set of technical protocols and standards for global information sharing it was rapidly becoming much more, as individuals, communities, businesses and governments took up the opportunities it presented, turning these into unanticipated outcomes that were raising profound economic, social and political questions. The Web was changing the world and the world was changing the Web. Working with, understanding and shaping the future of the Web demanded new ways of working across traditional academic silos, a new form of interdisciplinary expertise. Our success over the past decade has been remarkable: collaborative research grants, major doctoral training programmes, an annual Web Science conference, the Web Science Summer School, the Web Science Trust and a growing network of partner research groups in the Web Science Trust network. The interdisciplinary skills, concepts, and forms of collaboration that we have built over the past decade are essential if we are to face the challenges ahead. Just as in the early 2000s we face a rapidly morphing sociotechnical system, as the Web has led to Big Data, Data Science and the resurgence of Artificial Intelligence, raising profound questions not only about privacy, security and ethics – critical as these are – but also about the future of work, economic and social inclusion, risk, sustainability and global governance, about the kind of society we want for the future. Put this way, there can be no doubt that Web Science is needed now, more than ever!

#### 3.2 Why data science needs web science: reflections from recent research

*Elena Simperl (University of Southampton, GB, [e.simperl@soton.ac.uk](mailto:e.simperl@soton.ac.uk))*

License  Creative Commons BY 3.0 Unported license  
© Elena Simperl

Data science emerged as a term in its own right less than five years ago. Since then it has proven tremendously popular across sectors and organisations around the world looking for new ways to innovate their products, services and operations. The demand for data scientists has never been greater – to be able to source and make sense of data using advanced data science machinery, organisations need new capabilities, drawing upon statistical and computational methods.

The data science community developed a wide understanding of the challenges and limitations their methodological tools. Data science approaches are often much more than a mix of mathematical models – they rely heavily on having access to relevant data, on the quality of this data, and on means that help people use understand their implications in



practical situations. Web science can provide a context and emerging solutions to some of these questions. The Web is one of the largest sources of data ever created, including large repositories of labelled data, for example in the form of user-generated content, system logs, or social media posts, that can be used to train data science models.


Web scientists have worked with these training corpora for many years and have developed an unparalleled understanding of the underlying systems and communities. They were among the first to debate critical issues around the quality of large data sets and the implications of the associated data collection and processing methodologies on the validity and usefulness of analytics. These experiences could and should inform the design of data science algorithms, guide the interpretation of their outcomes, and enrich ongoing discussions around responsible data science and FAIR.

By the methods it uses, data science draws upon insights from different disciplines. Web science is one of the best examples of recent date that show how interdisciplinarity could work – the lessons learned since the seminal article in *Communications of the ACM* that founded the field in terms of research methods, education and impact should be transferred into data science to develop a similarly rich understanding of the opportunities and challenges of true cross-disciplinary work.

We live in a time when data science is plagued by concerns about responsible data sharing and use. These are key challenges that can be addressed only through a concentrated effort that is mindful and serious about interdisciplinarity and defines a new code of values in the digital world. Data scientists should collaborate with Web scientists to learn more about the biases and limitations of Web datasets and platforms and about the broader legal and socio-economic implications of data sharing and algorithmic decision making. I believe this is only way to ensure that the ever-growing trend of datification affecting every single aspect of our lives will have a positive impact over time.

### 3.3 Legal Aspects

*Nikolaus Forgó (University of Vienna, AT, [nikolaus.forgo@univie.ac.at](mailto:nikolaus.forgo@univie.ac.at))*

**License**  Creative Commons BY 3.0 Unported license  
© Nikolaus Forgó


**Main reference** REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

**URL** <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=DE>

The presentation gave an overview on governance issues using legal means in the development of the web. Specific attention was given to data protection and data security law in Europe as it is enshrined in the General Data Protection Regulation, a law that became directly applicable in all EU-member states in May 2018. The history of data protection, leading back to the eighties of the 20th century shows that data protection as always about conflicting fundamental values that are articulated in fundamental rights. The main argument made was that rules that are “clear” enough to regulate technical developments on the web require (some kind of) basic understanding about how to outbalance conflicting fundamental rights. As long as these discussions are not led and not decided, data protection law risks to be too late too abstract and too unclear to produce the certainty Europe needs. The examples taken to illustrate this argument were mainly taken from European research projects in the domain of ICT for health.

### 3.4 Social experiments in the Web: The case of Bibsonomy – technical, social and legal implications

Andreas Hotho (University of Wuerzburg, DE, [hotho@informatik.uni-wuerzburg.de](mailto:hotho@informatik.uni-wuerzburg.de))

License  Creative Commons BY 3.0 Unported license  
© Andreas Hotho

It's more than ten years now, that the transition of the Web from a static to a user-driven and dynamic Web has started. In 2006, we have initiated a social experiment in the form of BibSonomy, a research platform to publicly share, manage and exchange bookmarks and publication. Users annotate or describe these items by tags which enable their efficient retrieval and support the management of large document collections. The collected data results in a lightweight semantic structure contributed to by millions of users and forms a valuable source of knowledge which can be exploited in many different research and application scenarios. For this, Data Mining and Machine Learning methods play a central role and help to extract a multitude of valuable information from the user contributed data.

The talk gives a review of the past 10 years of developing, maintaining, but also analyzing the BibSonomy system. Consequently, besides addressing typical software design issues as well as standard technical development questions, this talk focuses on a set of selected research questions and their results. A concise summary of these results can be found in [1], while the latest related research on behavior analysis and representation learning can be found in [2] and [3], respectively. Additionally, the talk touches on legal aspects and ends with lessons learned. Thus, overall, this talk summarizes and illustrates the story of bringing Web Science to life.

#### References

- 1 Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. The social bookmark and publication management system bibsonomy. *The VLDB Journal*, 19(6):849–875, dec 2010.
- 2 Stephan Doerfel, Daniel Zoller, Philipp Singer, Thomas Niebler, Andreas Hotho, and Markus Strohmaier. What users actually do in a social tagging system: A study of user behavior in bibsonomy. *ACM Transactions on the Web*, 10(2):14:1–14:32, May 2016.
- 3 Thomas Niebler, Luzian Hahn, and Andreas Hotho. Learning word embeddings from tagging data: A methodological comparison. In *Proceedings of the LWDA*, 2017.

### 3.5 What can Web Science give to industry


Paolo Parigi (Airbnb and Stanford University – San Francisco, US, [pparigi@stanford.edu](mailto:pparigi@stanford.edu))

License  Creative Commons BY 3.0 Unported license  
© Paolo Parigi

Ever since Thomas Edison's Menlo Park laboratory or Bell Labs, modern companies have recognized the importance of research for the development of new products. Until recently, most of the research done in industrial research facilities focused on building new physical objects. Exceptions existed but there were indeed exceptions. Against this background, companies like Airbnb, Uber, Lyft, Facebook, LinkedIn, Tinder are investing in research that focuses on learning about people's beliefs and behavior. This shift reflects the fact the products of these companies uses technology in order to facilitate interactions (in person or electronically). In turn, interactions create systems with emergent properties that require the expertise of Web scientists to study. I will illustrate this point using examples taken from my work and that of my collaborators in and outside of academia.

### 3.6 Web Science: What Next?

Ricardo Baeza-Yates (NTENT & Northeastern Univ., USA, [rbaeza@acm.org](mailto:rbaeza@acm.org))

License  Creative Commons BY 3.0 Unported license  
© Ricardo Baeza-Yates

I participated in a Web Science panel held at ECML/PKDD 2009 in Bled, Slovenia. This was my first encounter with this name, just one year after its inception. Then I had many questions: Do we need a new science? How much knowledge do you cover? Must a new science have something new? How hard it is to start? Science of an abstraction? (concepts, tools, and applications) Science of an object? Science in the name? The last two are the same problems that Computer Science has. Today, many of these questions are still valid and I do not know all the answers.

So, does Web Science covers all possible applications of the Web or we should just focus on the Web as communication channel plus the technology behind it? I believe the answer is the second one, which includes web search and data mining, web advertising, web user experience understanding, online social networks, etc. However, this does not include the applications to other fields such as e-health, e-science, e-learning, etc. Even computer networking might be out of the scope as Internet is the infrastructure that allows the Web to exist, but is not really part of the Web.

The Web conceptually could be considered as having three parts: content (including everything such as links), interaction (as result of the people using it), and incentives (the information market that drives the interaction with the content). This implies that Web Science should be important for data scientists, economists, sociologists, ethnographers, web designers, and many kind of software developers.

In the long list of web research problems, I am really concerned about the vicious cycle of bias in the Web [1], understanding users and the content generation process (most users do not contribute [3, 5]), the limits of small web data [2], and the use of machine learning to take fair decisions. The last two implies understanding the trade-offs between quality and data volume [4] as well as understanding how prediction errors are distributed in each problem space (usually this is not analyzed today). This naturally leads to design of ethical codes and accountable algorithms that can be explained and audited.


However, the future can be more complicated if we leave the solution of the societal problems to politicians and lawyers. So, we need to step up and not worry about discrimination and ethics when is already too late.

#### References

- 1 Ricardo Baeza-Yates. Bias on the Web. *Communications of ACM* 61(6), pp. 54–61, 2018.
- 2 Ricardo Baeza-Yates. BIG, small or Right Data: Which is the proper focus? <https://www.kdnuggets.com/2018/10/big-small-right-data.html>.
- 3 Ricardo Baeza-Yates, Diego Sáez-Trumper Wisdom of the Crowd or Wisdom of a Few?: An Analysis of Users' Content Generation. In *ACM Hypertext & Social Media*, pp. 69–74, Guzelyurt, TRNC, Cyprus, Sep 2015.
- 4 Ricardo Baeza-Yates, Zeinab Liaghat. Quality-efficiency trade-offs in machine learning for text processing. In *IEEE BigData 2017*, Boston, USA, Dec 2017.
- 5 Lorena Recalde, David Nettleton, Ricardo Baeza-Yates, Ludovico Boratto. Detection of Trending Topic Communities: Bridging Content Creators and Distributors. In *Hypertext 2017*, pp. 205–213, Prague, Check Republic, Jul 2017.

### 3.7 AI and Society

*Wendy Hall (University of Southampton, GB, wh@ecs.soton.ac.uk )*

License  Creative Commons BY 3.0 Unported license  
© Wendy Hall

Why is there so much hype around AI now? The concept of AI has been in the minds of science fiction writers for centuries. The idea of computers with intelligence started in the UK with Alan Turing's seminal paper *Computing Machinery and Intelligence*, published in 1950. The term AI was coined in the US in the 1950's and has been exciting researchers all over the world ever since but up until now it's progress has been one of exciting leaps forward followed by research funding blights, often called the AI winters. The current leap forward is driven by new developments in deep learning, high performance computing, the internet and the availability of vast amounts of data and on which algorithms can be trained. The current leap forward is also referred to as the Fourth Industrial Revolution and maybe as deeply profound for the future of society as previous industrial revolution. Only time will tell.

Without doubt in this early part of the 21st century, the leading countries in terms of the scale of AI research they support are the US and China, but because of the potential impact of AI on all our futures, other governments are actively considering strategies and policies for adopting and adapting to AI over the coming years. To build on the strengths the UK has in AI and to ensure the UK stays at the forefront of the AI revolution, Jerome Pesenti and I were asked by the UK government in March 2017 to undertake an AI review focussing on job creation and the growth of AI as an industry sector in the UK. This review was published in October 2017.

The review proved timely and was well received. Money was allocated in the UK government budget in November 2017 and the same month it was announced that AI would be one of the key sectors identified for support in the government's Industrial Strategy. There was funding also in the budget for the establishment of a new Centre for Data Ethics and Innovation in the UK, which was established in 2018. The detail of the £1 Billion AI sector deal was published at the end of April 2018 and at the same time the government established the new Government Office for AI.

The recommendations in the review were largely organised under four major themes – data, skills, leadership and adoption. The Office for AI has already started the consultation work for the development of data trusts, which were highlighted in the AI review as mechanisms to support the exchange of data between government, businesses – large and small, and university research labs, to support innovation in and around AI. During the review consultations, companies told us that their biggest problem was getting access to the data they needed to train algorithms, and this has to happen in safe and secure, ethically sound legal frameworks to enable innovation to flourish in a way that is good for society. The role of the new Centre for Data Ethics and Innovation will be crucial in taking this forward in collaboration with the OAI, its implementation partners, and the AI Council.

Implementing the recommendation under skills has also started. UK Research and Innovation (UKRI) launched a call in February 2018 for up to 20 new AI Centres of Doctoral Training to start in October 2019, each of which will train at least 10 students per year for 5 years. The Alan Turing Institute – now the national institute for AI – will manage a commensurate AI fellowship programme, and there will also be programmes to establish industry funded AI and Machine Learning MSc's in UK universities.

However, when we were writing the review we were very conscious that to fill the AI skills gap and to ensure that AI is not only about technology but is also about how AI will

impact society and how society will utilise AI, we included a stream of work to facilitate the development of “conversion” courses that will take students from all disciplines (STEM and non-STEM) as well as students who are re-skilling and give them the expertise they need to work in the AI sector (not necessarily as ML programmers). Running alongside all this will be a campaign to increase diversity in the AI workforce. It is really important that the AI based products and services that are going to become increasingly important in every aspect of our lives in the future are not dominated by any one gender, age group, ethnicity, or culture and are accessible to all.

The UK is not the only government that has developed an AI strategy over the last year and it is interesting to see how the different strategies are panning out. However it will be many years before we can really evaluate what impact government intervention in AI has around the world given the potentially seismic political and global shifts that are bound to come over the next few years. How much and how fast AI is going to become a reality in our lives rather than just science fiction is hard to tell. There is a lot of hype at the moment, but one thing is for sure it will become a matter of global competition.

I would argue however, that it is just as important that we collaborate internationally as compete. AI has the potential to solve or help manage the biggest challenges that society faces in the 21st century but if we pool resources (data, research results, expertise etc.) we could achieve a lot more a lot faster and still enable our companies to compete internationally to sell the products and services that are produced as a result.

AI also has the potential to do a great deal of harm. We must keep at the front of our minds the sociotechnical impact of AI to ensure that we develop AI technologies that are first and foremost for the good of society. I believe the first ten years of Web Science were just the preparation for the real challenges that lie ahead in terms of building a world in which AI plays an ever more significant role in our increasingly interconnected lives.

### 3.8 Web Science, Artificial Intelligence and Intelligence Augmentation

*Fabien Gandon (Inria, Université Côte d’Azur – Sophia Antipolis, FR, [fabien.gandon@inria.fr](mailto:fabien.gandon@inria.fr))*

**License** © Creative Commons BY 3.0 Unported license  
© Fabien Gandon

**Main reference** Gandon, Fabien: “Web Science, Artificial Intelligence and Intelligence Augmentation”, Journal of Seminar Documentation, 1:8, pp. 34–78.

This abstract paper summarizes some challenges and opportunities at the intersection of Web Science, Artificial Intelligence and Intelligence Augmentation.

#### Intelligent approaches to follow and support Web evolution

Initially, the Web was essentially perceived as a huge distributed library of linked pages, a worldwide documentary space for humans. In the mid-90s, with wikis and forums, the Web was re-opened in read-write mode and this paved the way to numerous new social media applications. The Web is now a space where three billion users interact with billions of pages and numerous software. In parallel, extensions of the Web were developed and deployed to make it more and more machine friendly supporting the publication and consumption by software agents of worldwide linked data published on a semantic Web. As a result of all its evolutions, the Web became a collaborative space for natural and artificial intelligence. This raises the problem of supporting these worldwide interactions and forming these hybrid communities. In my talk I presented some of the opportunities and challenges for Web Science in building this evolution of a Web toward a universal space linking all kinds of intelligence.

**AI in classical tasks and problems of the Web**

A first set of challenges can be directly identified from the classical tasks and problems we encounter on the Web e.g. help us search, browse, contribute, etc. The Web already is populated by Web bots but they usually are restricted to certain realms while they could be generalized. For instance we could generalize the bots as the ones of Wikipedia to bots on the open Web designed to monitor and preserve certain characteristics of the Web. We could imagine Web farms for Web AIs hosting autonomous agent that would study, monitor and report on the Web. Problems that could be targeted by these Web bots include: the detection of metrics manipulation, cross-language plagiarisms, centralization or digital divide; the prevention of vandalism or spamming; the generation of links, back links, navigational content beyond search results; etc. These agents would be based on policies and values important to the philosophy of the Web (e.g. seek decentralization, equality of access) to improve its resilience and quality.

**The special relation of AI and data(sets) on the Web**

The open and linked data facet of the Web is a special case of particular importance when considering the links between AI and Web (science) data (science). Artificial Intelligence can be used to assist Web Scientists and vice-versa. Intelligent agent can help us produce, curate, share and maintain corpora and datasets. For instance AI techniques could be designed to check the quality of a dataset and look for bias in it. Inversely, Web Science could produce multidisciplinary methods and tools to certify the quality and characterize training sets to improve the quality of the learning and conclusion made by AIs using them.

**Benevolent AIs for a resilient Web**

The two previous ideas could be generalized to the goal of designing benevolent AIs for the Web. Web agents working to improve users' experience, understanding, awareness and control of their participation and contributions to the Web. For instance, educational AI could help educate Web users in many domains including Web literacy or ethical thinking. Agents could also provide customized descriptions of the context in which a user is, including security, neutrality and privacy notices or his human-computing participation when it occurs. AI could also help users burst our filter bubbles and foster serendipity. On the longer term, benevolent AIs could actively help enforce (human) rights on the Web and be scrutiny agents for important values of the Web.

**AI to help us humans scale and face humanity on the Web**

With the advent of the Web, human individuals also face humanity in all its scale and diversity. Web scientists could design AIs to help humans face humanity on the Web and help us scale to the world-wide web scale. These goal-driven agents could actively participate to the online activity and, for instance, foster linkage, interactions and convergence, bridge, translate, check, or augment our posts and maintain for us an overview of our social context and activity. They could also prevent or report problems such as bullying, harassment and polarization.

**A variety of AIs to absorb the varieties of the Web**

The force of these AIs could also be in their multiplicity and interactions. The law of requisite variety of W.R. Ashby says that "variety absorbs variety" and in our case a diversity of AIs

could be a good way to address the many types of diversity we find of the Web (content, users, contexts, tasks, usages, resources, etc.). In fact more than AI, it is maybe distributed AI that has a rendezvous with the Web and its sciences [6]. Multi-agent systems and distributed AI blackboards are examples of distributed AI architectures which, if merged with the Web architecture would allow for many different kinds of AIs to collaborate worldwide to the benefit of the Web. The AIs and the multi-agent systems would also in return benefit from the Web, its resources and its methods. Following the wiki-way, AIs could be created, edited, crossed, and bred on the Web, socially maintained, copied and versioned: the Web way applied to AI with, for instance, “copy-paste-customize” based contribution to the population of agents. For this to happen, and just as it was the case for the Web, we would need a public domain Web-based AI architecture.

### **Explore and expand all the forms of intelligence on the Web**

The multidisciplinary nature of Web Science also puts it in an ideal position to explore and expand the forms of intelligence on the Web. First, both Web Science and AI are highly multidisciplinary [5] and the multiple disciplines that are common to both fields are as many bridges to make them interact. AI could also be used to operationalize the expertise from each domain into agents that help us providing assistance, reporting or training from the domains they represent. These agents could help us find and support a massively multidisciplinary method and allow us to scale to the multi-disciplinary interactions required by the design and study of the Web. One possibility, for instance, would be for these AIs to produce and maintain boundary artifacts at the frontiers of disciplines. The multidisciplinary domains of Web Science could also be leveraged to identify other ways of simulating, reproducing or engaging intelligence including emotional intelligence, communication skills, imagination, etc.

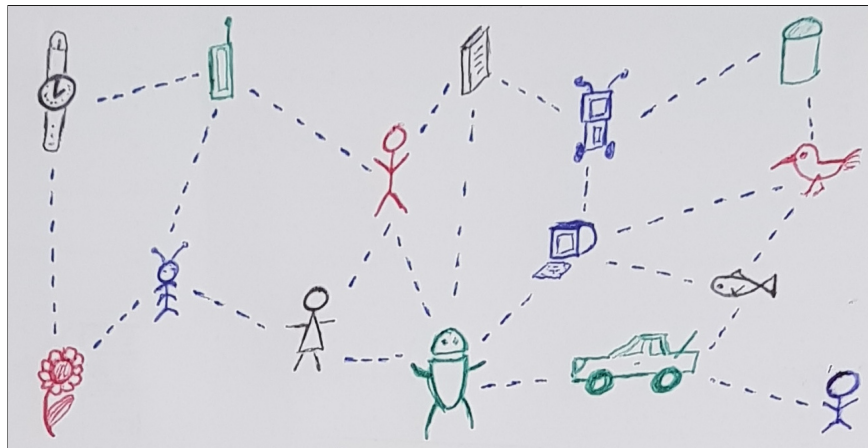
### **Studying and building the hybrid societies of the Web**

Such an evolution as the one described in the previous sections would finally lead Web Science to consider the challenge of studying and designing hybrid societies of natural intelligence and artificial intelligence on the Web. This study would have to include different forms of natural intelligence (e.g. people, connected animals, connected plants) and different forms of artificial intelligence (reasoning, learning, inducing, etc.). The challenge will also be to study their interactions with the resources of the Web (linked pages, linked data, connected objects, etc.) forming the environment of these forms of intelligences. Web Science will have to face the problem of this massive interaction design with the Web and everything it links [7] and AI will have to face the problem of engaging in very different types of interactions with different forms of intelligence including different kinds of AIs [4]. Studying and designing these hybrid societies, from swarms to complex societies with their normative rules, their social constructs, their governance, etc. will be a highly challenging and multidisciplinary task.

### **Towards a Web linking all forms of intelligence**

In Web Science, we should build our research program as a joint effort between Web Science and two research fields born in the 50s: “AI” for Artificial Intelligence [2] and “IA” for Intelligence Amplification [3] and Intelligence Augmentation [1].

To conclude this abstract in one sentence, I would say that a Web Science research agenda must account for the fact that the long term potential of the Web is to augment and link all forms of intelligence.



### References

- 1 Douglas C Engelbart. *Augmenting human intellect: A conceptual framework*. SRI Summary Report AFOSR-3223, Stanford Research Institute, 1962.
- 2 John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. *A proposal for the dartmouth summer research project on artificial intelligence*, august 1955. <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf>.
- 3 Ross Ashby. *Design for an intelligence-amplifier*. Automata studies, 400:215–233, 1956.
- 4 Fabien Gandon, Michel Buffa, Elena Cabrio, Olivier Corby, Catherine Faron Zucker, Alain Giboin, Nhan Le Thanh, Isabelle Mirbel, Peter Sander, Andrea G. B. Tettamanzi, and Serena Villata. *Challenges in Bridging Social Semantics and Formal Semantics on the Web*. In S. Hammoudi, J. Cordeiro, L.A. Maciaszek, and J. Filipe, editors, 5th International Conference, ICEIS 2013, volume 190 of Lecture Notes in Business Information Processing, pages 3–15, Angers, France, July 2013. Springer. <https://hal.inria.fr/hal-01059273>.
- 5 Fabien Gandon. *The three ‘W’ of the World Wide Web call for the three ‘M’ of a Massively Multidisciplinary Methodology*. In Valérie Monfort and Karl-Heinz Krempels, editors, 10th International Conference, WEBIST 2014, volume 226 of Web Information Systems and Technologies, Barcelona, Spain, April 2014. Springer International Publishing. <https://hal.inria.fr/hal-01223236>, doi:10.1007/978-3-319-27030-2.
- 6 Fabien Gandon. *Distributed Artificial Intelligence And Knowledge Management: Ontologies And Multi-Agent Systems For A Corporate Semantic Web*. PhD Thesis, INRIA, Université Nice Sophia Antipolis, November 2002. <https://tel.archives-ouvertes.fr/tel-00378201>.
- 7 Fabien Gandon and Alain Giboin. *Paving the WAI: Defining Web-Augmented Interactions*. In Web Science 2017 (WebSci17), pages 381 – 382, Troy, NY, United States, June 2017. <https://hal.inria.fr/hal-01560180>, doi:10.1038/nbt0609-508



### 3.9 New ethics for the web and for the web scientist?

Katharina E. Kinder-Kurlanda (*GESIS – Leibniz Institute for the Social Sciences, Cologne, DE*, [katharina.kinder-kurlanda@gesis.org](mailto:katharina.kinder-kurlanda@gesis.org))

**License** © Creative Commons BY 3.0 Unported license

© Katharina E. Kinder-Kurlanda

**Main reference** Katharina E. Kinder-Kurlanda, Katrin Weller, Wolfgang Zenk-Möltgen, Jürgen Pfeffer, Fred Morstatter: “Archiving information from geotagged tweets to promote reproducibility and comparability in social media research.” *Big Data & Society* 4(2), 2017.

**URL** <https://doi.org/10.1177/2053951717736336>

The talk looked at new challenges for ethically reflective web science research, and in particular at the challenges of changing data landscapes for epistemology and ethics. There is apparent a lack of transparency and validity, often linked to new arrangements for data access. Proprietary data ‘owners’ play an important, changing and often unclear role. Questions that arise are what knowledge we can gain – and should or should not gain – from various types of data and how the validity of statements can be checked if research processes and properties of data used are mostly opaque and inaccessible for scientific peer review. It is argued that Web Science requires more innovation in research documentation and in approaches to public private partnership arrangements, such as trusted third party models that meet both researchers’ and companies’ interests. Trusted research support infrastructures could play an important role in this as they offer expertise in both data management and long-term preservation. A challenge here lies in the linking of different types of support infrastructure institutions, both established and emerging, and in avoiding the building of a new library (where the data scientists go) next to an existing one (where the theory books are). Haraway’s concept of response-ability was used to argue for a research stance where we as web scientists are a) careful of methodology and proactive about interdisciplinarity b) create spaces for reflection and deliberation c) innovate in the ways to trigger critical thinking about ethics, epistemology and theory in teaching so as to d) not close ourselves off from research subjects, rejecting ‘research at a distance’.

### 3.10 Why formalising fairness won’t fix (algorithmic) discrimination (reloaded)

Bettina Berendt (*KU Leuven, BE*, [bettina.berendt@cs.kuleuven.be](mailto:bettina.berendt@cs.kuleuven.be))

**License** © Creative Commons BY 3.0 Unported license

© Bettina Berendt


The last ten years have seen large developments in analyses of and approaches to mitigating discrimination, bias and unfairness related to (semi-)automated decision making. In this talk, I will argue why a focus on “solving” this problem by formalising fairness and “fixing the algorithms” used in decision making is too narrow. Starting from the classical Myrdal analysis of discrimination’s cumulative causation, I argue that a) the computationally grounded concept of “non-discrimination” or “fairness” neglects core aspects of the real-life goals, which rest on human agency and empowerment, b) algorithms don’t discriminate, people do (and algorithms help), and c) to mitigate algorithm-related or any other discrimination, people are key (and computer science including, but importantly also beyond, algorithms can help). In addition, d) while this is a deeply ethical problem, addressing algorithm-related discrimination should not be left to “ethical approaches” alone, whether by scientists, developers, companies,

or others, but also be informed by, rely on, and cooperate with enforcement by laws and regulations.

The slides of this talk are available at [https://people.cs.kuleuven.be/~bettina.berendt/Talks/berendt\\_2018\\_06\\_28.pdf](https://people.cs.kuleuven.be/~bettina.berendt/Talks/berendt_2018_06_28.pdf).

### 3.11 World Wide Weapons: Project Maven, Google and Web Ethics

*Guglielmo Tamburrini (Universita' di Napoli Federico II – Napoli, IT, [tamburrini@unina.it](mailto:tamburrini@unina.it))*

License  Creative Commons BY 3.0 Unported license  
© Guglielmo Tamburrini

In 2010 Tim Berners-Lee warned us that “The Web, as we know it, is being threatened... by governments – totalitarian and democratic alike – monitoring people’s online habits, endangering important human rights”. This threat is enhanced today by the growing convergence of AI, big data, and Web analytics in the design, development and deployment of increasingly autonomous weapons systems that are endowed with both monitoring and kinetic military capabilities. The Web Science community has to address these threats by reinforcing the ethical pillar of Web Science. This can be achieved by moving from reactive to proactive ethical issues identification and ethical policy development, as well as by expanding both reflective and scientific work on protecting and promoting human rights in the light of web integration with AI, big data, Internet of things and cyber-physical systems.

A specific paradigmatic example of this technological convergence and the ethical problems that it gives rise to is discussed in this talk. This is the Project Maven, undertaken in 2017 by the US Department of Defence (DoD) in cooperation with Google and other Web service actors. The project aims to achieve military target selection from drone video footage and to integrate eventually this capability into Gorgon Head and other powerful surveillance systems.

The organization of the project Maven is used to illustrate the fact that that the so-called military-industrial complex must adapt to the circumstance that today AI development is chiefly happening in the commercial sector. Significant implications of this fact for action towards the protection of fundamental human rights will be emphasized too. Indeed, commercial companies like Google are a complex multi-actor entity, whose employees have manifested deep concern for the protection of human rights and the peaceful character of their company’s activities in the face of their company’s involvement in Project Maven. In particular, this talk will concentrate on a contextual analysis of the ethical concerns manifested by over 3,000 Google’s employees in a letter of April 4th, 2018 addressed to the company’s CEO. These are issues of trust for billions of users all over the world, the company’s moral reputation, and technologically possible, albeit maleficent uses of the developed technological tools beyond their declared aims. Moreover, in agreement with President Eisenhower’s concerns about the lack of democratic transparency and accountability of what he called the military-industrial complex, researchers supporting Google’s employees letter pointed out that Google has moved into this sort of military work without subjecting itself to public debate or deliberation, either domestically or internationally.

Similar ethical issues will be discussed in connection with security and military uses of the Amazon’s face recognition system , ‘Amazon Rekognition’ (<https://aws.amazon.com/rekognition>) , which integrates video footage processing with searches in databases containing tens of millions of faces. These various ethical debates about the embedding of AI and

Web analytics techniques into military applications will be analyzed in connection with ongoing ethical debates about autonomous weapons systems and their underpinnings in both deontological and consequentialist normative ethical theorizing [1].

On the whole, these are urgent and emerging ethical issues for the Web Science research community to address. This community may play a crucial role in promoting ethical dialogue about surveillance and military uses of AI, big data, and Web analytics integration. In particular, one should promote ethical dialogue and understanding between Web scientists working on different sides of «Web experiments» and with different cultural backgrounds, much as the Pugwash conferences have been doing with the international community of nuclear physicists in the wake of the 1955 Einstein-Russell manifesto.

## References

- 1 D. Amoroso; G. Tamburrini, *The ethical and legal case against autonomy in weapons systems*, Global Jurist, 17:3, 1–20, 2017.

## 4 Working Groups

### 4.1 Innovative methods for Web Science (Visualization group)

*Susan Halford (University of Southampton, GB, Susan.Halford@soton.ac.uk)*

*Lynda Hardman (CWI, Amsterdam and Utrecht University, NL, Lynda.Hardman@cwi.nl)*

*Katharina Kinder-Kurlanda (GESIS, Cologne, DE, katharina.kinder-kurlanda@gesis.org)*

*Claudia Müller-Birn (Freie Universität Berlin, Berlin, DE, clmb@inf.fu-berlin.de)*

*Paolo Parigi (Airbnb and Stanford University, San Francisco, US, pparigi@stanford.edu)*

*Elena Simperl (University of Southampton, GB, e.simperl@soton.ac.uk)*

License © Creative Commons BY 3.0 Unported license

© Katharina Kinder-Kurlanda, Claudia Müller-Birn, Lynda Hardman

This working group focused on novel research methods in Web Science to address the following challenges:

- How can we unlock the promise of interdisciplinary research inherent in the vision of Web Science?
- How do we link qualitative and quantitative research results in a more meaningful way?
- How can Web scientists engage with (proprietary) data owners on equal terms?
- How do we address issues around data availability and quality?

#### 4.1.1 Discussed Problems

Three themes were discussed as potential solutions to approach the previous questions:

- Visualization
- Online experiments
- Participatory

##### 4.1.1.1 Visualization can facilitate and empower interdisciplinarity

Data visualisation can help Web scientists from different disciplines work with data more effectively. It creates new ways to engage with data, helping people 'interview' it to make sense of it. It enables a productive dialogue between data, method and theory, for example

in the context of abductive reasoning. Data is often complex either by virtue of its content or quality, or because it is structured or formatted in a particular way. Visualisation helps manage some of this complexity, assisting with hypothesis formulation iteratively.

Data visualisation can also help facilitate collaborations – computer and social scientists could work together to bring data and theory into dialogue from the very beginning. They could look at the same data, ask questions, and present the outcomes in multiple ways in an abductive process. Tool support and inspiration could be drawn from the fields of visual analytics and visual data mining, where questions are derived from visual, more accessible data representations to lead to new hypotheses.

Visualization can also be a way to represent research outcomes to other audiences – academics from other fields, industry, government and the public. It helps communicate complex data and results effectively and as such it has to be an integral part of any research effort, considered and planned from the beginning and not towards the end of a project when preliminary results become available.

The group also touched upon a series of additional topics, which are listed here as they provide important research questions the Web science community could consider:

- Could we map interviewing methods using in social sciences (e.g. expert, semi-structured, open) to data exploration?
- Could we map such methods to computational techniques?
- How would one annotate visualisations of qualitative data with qualitative information?
- What would be the best ways to engage with other relevant disciplines e.g., art, design?

#### 4.1.1.2 Online experimentation and relationships with data owners

This theme looked at how Web scientists could do valid, ethical research, especially but not only in the social sciences and when using Web data. Such efforts aim to understand and bring together core social research questions with engineering methods. The group identified three sub-themes:

- Getting access and creating new sources of data for Web science research;
- Engaging in a critical analysis of ethics, methods and effects; and
- Developing insights and understanding of industry practice

#### 4.1.1.3 Participatory methods

In this theme the group explored the use of participatory methods as a means to

- Learn about people's views and opinions and engage with different groups and communities to develop a broader understanding of the Web and its future
- Improve the quality of critical datasets
- Approach ethical dilemmas

The group acknowledged the importance of a participatory approach to take into account multiple perspectives and encourage dialogue between them. Insights from participatory methods shape research questions and solutions. These methods are very diverse, from citizen (social) science and harnessing new and emerging forms of data (e.g., from social media) to online deliberation methods, citizens' assemblies and the use of AI techniques (for example, to enhance knowledge and understanding of the Web and extending dialogue).

#### 4.1.2 Possible Approaches

Innovation in methods requires:

- new models for accessing and sharing data;
- better mechanisms and tools to share and reuse algorithms within academia
- building Web science data archives, data management tools (for documenting research processes throughout a project's lifetime) and infrastructures;
- novel models and frameworks to collaborate with industry, especially on data sharing;
- actively seeking collaboration across domains and being open and interested in including new disciplines in Web science.

### 4.2 Working Group on Values

*Bettina Berendt (KU Leuven, BE)*

*Eirini Ntoutsis (Leibniz Universität Hannover, DE)*

*Evaggelia Pitoura (University of Ioannina, GR)*

*Steffen Staab (Universität Koblenz-Landau, DE)*

*Guglielmo Tamburrini (University of Naples, IT)*

*Pinelopi Troullinou (The Open University – Milton Keynes, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Bettina Berendt, Pinelopi Troullinou, Eirini Ntoutsis

This working group focused on the values of the Web Science community which stem from and reflect the vision of the Web Science community in accordance with human rights. It is important to acknowledge the distinctive role of Web Science Community (as stated in previous documentation) in the promotion of social goals. Therefore, the values need to reflect this. In response to the call of Sir Tim Berners-Lee to fix the Web, it is now the time to design and develop a training process for web scientists' reflectivity on ethical values. It is critical for Web Science to have/reflect on such values due to the huge impact of the Web in our society.

The discussion in this working group was triggered by the following questions:

- What is the Web we want (what are our values to be reflected on the Web)?
- How do we position ourselves in ethical and societal debates? We cannot be neutral, we need to take a stance adopting the human rights.
- What is the role of the web scientist?
- How do we make these things explicit?
- How should the values and the training process of web scientists' reflectivity on these values become part of a Web Science curriculum?

With these questions in mind, the discussion was organized according to the following topics:

- Current and foreseen future value-related problems in the Web.
- The role of technology, especially of AI as an amplifier of existing problems and as a creator of new value-related problems
- Vision of the Web Science Community on values and ethics in the Web
- A code of ethics for Web Scientists

#### 4.2.1 Discussed Problems

##### 4.2.1.1 Part 1: The value-related problems of the current/future Web

The group discussed a large variety of value-related problems in the Web from online harassment, misinformation, polarization/extremism, attempts to stir the public media attention using “distractions” to the normalization of such sort of behavior under your real name and the rising problem of misconducting behaviour by AI (e.g., Microsoft’s Tay bot).

##### 4.2.1.2 Part 2: The role of technology in amplifying or even triggering value-related problems

User-generated content and Web data in general are among the main enablers for AI, as one of the most successful forms of AI nowadays, machine learning (ML), relies heavily on data. Therefore, whatever biases exist in the data are also reflected in the result of these algorithms or even worse, they are amplified as ML aims to maximize predictive performance rather than ensure fairness. Such implications have been already reported for web search engines (e.g., ranking outputs problems), online services (e.g., Amazon same day prime case), online advertisements etc.

##### 4.2.1.3 Part 3: What are our values? What is the Web we want to have?

When we speak about “values”, we often think of “the protection of human rights”. We endorse the human rights of the Universal Declaration of Human Rights, and their expression in the fundamental rights of our respective constitutions and similar charters. This seemingly simple and easy-to-agree-on statement already shows two key problems that a value-centric ethics for Web Science faces. First, there is only one human/fundamental right that is generally agreed upon to be non-negotiable: human dignity. All the other human rights can be, and regularly are, balanced and traded off against other rights. Second, even the conceptual movement from human rights to fundamental rights show that what is considered “basic” may depend on the jurisdiction. Thus, human rights are not enough as a guideline.

A second layer, and one that is often closer to “implementations” and more amenable to enforcement, are laws. Laws may embody fundamental rights and also already be the product of a societal balancing between different such rights. (An example is the EU’s General Data Protection Regulation.) A challenge for the Web and Web Science is that the laws of different jurisdictions may be incompatible with one another; a reason for optimism is the existence of international laws (many of which are not well-known in the community). On the whole, we observe the need for more awareness of which laws exist and are applicable, and for informed discussions of what legal compliance means (and what value orientation over and above legal compliance is).

There are other values too that play a large role for Web Science, such as universal access to information (see Fabien Gandon’s talk, described in Section 3.8). More debates on these values (as well as their possible limitations, see for example [?], are needed.

Finally, values need to be identified and discussed at different levels: that of the individual researcher (or practitioner), that of platforms, and that of the Web as a whole.

##### 4.2.1.4 Part 4: What can we do?

Agreeing on codes of ethics is a necessary first step. The AoIR Guidelines [1] are an excellent starting point in that they were developed by and meant for *internet* researchers. However, Web Science not only does research on the Web (as a part of the internet), it also and

centrally is about *designing* that very Web. Therefore, we need methodologies to embed ethical concerns/values in design, maintenance, and revision processes. Examples are various forms of (a) value-based/sensitive design (as propagated, for example, in the IEEE Ethically Aligned Design Guidelines [2], of (b) participatory design (for an older but very comprehensive survey, see [3]), and reflective-design [4].

In addition to the initial design, affirmative/corrective actions must be provided for. Here, the question arises based on which values and laws, and “armed” with which techniques, these can be effected. For example, can copyright be used to prevent revenge porn? Finally, design thinking and methodology must provide for newly emerging situations and sudden emergencies, in which reaction speed becomes crucial.

Another field of action is the raising of awareness. This includes awareness about which fields are regulated by law. For example, the activities of bots that produce public speech content (such as Wikipedia bots) are subject to laws on public speech.

Web Science must also move beyond design and awareness-raising. We need to carry out comparative studies between countries and risk assessment exercises. We need to preview and be proactive, trying to predict things that will happen.

There are many phenomena on the Web that can only be understood properly when more disciplines are involved in these studies. These include various “human sciences” such as psychology and anthropology. Web-related phenomena that could be studied better in such interdisciplinary settings include the differences between online and offline behavior and the lack of empathy over the spatial and other distances in the Web/internet, as well as the – possibly mediating – effects of cultural proximity and distance (such as having more empathy with one’s compatriots located in China than with the Chinese themselves).

#### 4.2.2 Raising awareness – a case study on building a new social network system

To exercise our own awareness on problems, we followed a reflective exercise on building a social network system to conduct an ethical analysis (similar to Potter’s box of reasoning). The goal of the exercise is to generate awareness on societal and ethical concerns that are not pre-fixed but emanate from any stage of the project activities (see d’Aquino et. al., 2018)

Our analysis went as follows:

1. Identify the functional requirements of the social network system
2. Identify the stakeholders of the network system
3. Identify the values of the Web Science community and of the different stakeholders (examine conflicting interests)
4. Make a decision on each functional requirement reflecting on the above. For example, (1) the growth of the network as a functional requirement (2) who are the stakeholders of the specific network system e.g., owner, user, non-user, web scientist, advertisers, marketeers etc.) (3) what are the values of each stakeholder and what are the potential conflicts of interest e.g., owner seeks for profit – includes everybody, web science community fights against hate speech circulation – excludes extremists. (4) decide the criteria or not on growth.
5. Reflect on contingent critical situations (e.g., shutting down Microsoft’s Tay bot after it became racist).

### 4.2.3 Conclusions

As the Web plays a tremendous role in all aspects of our life and as each and every one of us is affected by the Web but also impacting the Web (and consequently, the society) via his/her actions, we all agreed that it is now more urgent than ever for the Web Science community to take actions to safeguard our values and promote societal good.

We outline such actions below:

- Raising awareness of the Web Science values and the role and responsibilities of each and every involved entity, from the web scientist to the platforms and the Web as a whole.
- Agree on a code of ethics for Web Scientists, possibly by drawing on existing guidelines and combining/extending them to the specifics of Web Science. Particularly relevant guidelines include: AoIR [1], IEEE [2].
- Integrate values into all stages of the knowledge discovery process, from data collection (e.g., is the dataset used for training representative of the broader population?), to preprocessing (e.g., can I use sensitive attributes or proxies to sensitive attributes in the model?), analysis (e.g., is my model also learning for the minority groups?), evaluation (e.g., is my testing set coming from the same distribution as the training set?) and interpretation of the results (e.g., how the end user personal biases/ preferences affect the interpretation?) as well as adopt model lifecycle management to ensure the model is performing well beyond its predictive performance (e.g., is the deployed model “rejecting” more and more minority instances over the course of its operation?).

### References

- 1 AOIR (Association of Internet Researchers), Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). (2012) <https://aoir.org/reports/ethics2.pdf>
- 2 IEEE (2016). Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. Version 1 for public discussion. [http://standards.ieee.org/develop/indconn/ec/ead\\_v1.pdf](http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf)
- 3 Muller, M.J. & Druin, A. (2003). Participatory Design: The Third Space in HCI. IBM Technical Report. [http://www.watson.ibm.com/cambridge/Technical\\_Reports/2010/TR2010.10%20Participatory%20Design%20The%20Third%20Space%20in%20HCI.pdf](http://www.watson.ibm.com/cambridge/Technical_Reports/2010/TR2010.10%20Participatory%20Design%20The%20Third%20Space%20in%20HCI.pdf). Shorter version in J.A. Jacko & A. Sears (Eds.), *The Human-computer Interaction Handbook* (pp. 1051/1068). Hillsdale, NJ: Lawrence Erlbaum Associates.
- 4 Sengers, P., Boehner, K., David, S., & Kaye, J. (2005). Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (pp. 49–58). New York: ACM.



### 4.3 Working Group on Web Science and Artificial Intelligence

*Fabien Gandon (Inria, Université Côte d’Azur, I3S, CNRS – Sophia Antipolis, FR, fabien.gandon@inria.fr)*

*Oshani Seneviratne (Rensselaer Polytechnic Institute – Troy, New York, USA, senevo@rpi.edu)*

*Noshir S. Contractor (Northwestern University – Evanston, US)*

*David De Roure (University of Oxford, GB)*

*Kemal A. Delic (Hewlett Packard – Grenoble, FR)*

*Wendy Hall (University of Southampton, GB)*

*Andreas Hotho (Universität Würzburg, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Fabien Gandon, Oshani Seneviratne, Noshir S. Contractor, David De Roure, Kemal A. Delic, Wendy Hall, Andreas Hotho

“True, and yet the pattern is fixed. Revolt, suppression, revolt, suppression  
– and within a century Earth will be virtually wiped out as a populated world.

So the sociologists say.”

Baley stirred uneasily. One didn’t question sociologists and their computers.”

—Isaac Asimov, *The Naked Sun* (1957)

#### 4.3.1 Bootstrapping the debate: members of the group and their opening statements

The working group started with a round table where every participant shared some opening statements to start the discussion. We report them here following in the alphabetical order of the participants’ family name:

- *Noshir S. Contractor (Northwestern University – Evanston, US)* proposed to consider the very concrete effect of AI and the Web on work and the work force and the possible futures these two domains are drawing for work. He proposed to consider the role of AI can play in human-agent teaming and personal assistance to team members: “imagine every manager with Watson in his pocket”.
- *David De Roure (University of Oxford, GB)* distinguished up front two subjects: (1) using AI as a method in conducting Web Science and (2) Web Science studying a Web that includes AI. Coming from the symbolic AI school, he reminded us that AI is also about knowledge representation, and that the whole area of the Semantic Web is currently not much discussed in Web Science. He also noted that the subject may also be discussed under many different names “knowledge graphs” and linked data. He expressed a personal interest for a Web Science studying human-agent collectives and social machines including AIs.
- *Kemal A. Delic (Hewlett Packard – Grenoble, FR)* raised the challenge of hyperscale systems and terabytes of Web data and the role of Web Science in the study of the complexity of these hyperscale systems.
- *Fabien Gandon (INRIA Sophia Antipolis, FR)* started by recalling the two meanings of the word “Web” : (1) a standardized software architecture (2) the actual hypermedia we weave worldwide. He believes that both aspects must be part of the Web Science research topics and both raise different research questions with respect to the relation of AI and the Web. As a first set a question he identified classical Web problems and tasks that can benefit from AI including: indexing, searching, browsing, etc. He then identified the special case of data on the Web insisting on both directions e.g. the use of AI to produce the datasets for Web Science studies and inversely the study and design in

Web Science of high quality datasets to be used by AI for training, reasoning, etc. He then moved to the challenge of inventing benevolent AI for the Web (e.g. watchdogs for the Web) and also helping humans face humanity in terms of diversity and scale but also in preventing unwanted behaviours (e.g. bullying). He insisted on the fact that the domain of distributed AI (e.g. multi-agent systems, blackboards) could be instrumental in bridging AI and the decentralized Web.

- *Wendy Hall (University of Southampton, GB)* reminded us that from the beginning, Web Science was meant to be more than the study of the protocols of the Web and that one of the possible names initially considered was “network science” but was already taken. She insisted that Web Science is by nature interdisciplinary and socio-technical on a large scale. She recalled that AI is already used on the Web and will be used hugely as the Web evolves. She also pointed that we needed to talk about the future of the Internet as part of the future of Web Science.
- *Andreas Hotho (Universität Würzburg, DE)* recalled that AI is a broad area already and that this gets even worse when trying to relate it to Web Science. He stressed the importance for him of the specific relation between Machine Learning and Semantic Web to support Web science, including the application to Natural Language Processing (NLP) and Semantic Web mining. He also insisted on the key role of data and their management in the relation between AI and Web Science.
- *Oshani Seneviratne (Rensselaer Polytechnic, US)* raised the challenge of encouraging the benevolent usages of the Web and AI and preventing undesirable uses and effects. This was further specialized as a need to consider how AI can be used to perform good science research that is fair, can be reproduced easily, and the required characteristics for that such as explainable AI and transparent algorithms.

#### 4.3.2 Propositions and reactions: brainwriting pool on Web Science and AI convergence

As a second step, we did the exercise of “Brainwriting Pool”<sup>1</sup> or “Consequences” using some topics mentioned during the first round-table and new topics proposed by the participants to initialize the pool.

This resulted in the following suggestions of research areas for Web Science:

- The role of Web data science to help solve problems of datasets of AI (bias, etc), to automate scrutiny on the Web and generate reports for Web scientists.
- The use of AI techniques to avoid privacy implications and issues.
- The fostering of Web exploration by using AI
- The support of reproducible research and dataset sharing as a core feature in Web Science research (“Web science AIs should be citable in Web Science papers”)
- The need for “laws of Web AI” (cf Asimov) because a Web AI is at least as dangerous as a robot. The need for ethical principles of Web Science research and normative systems for hybrid societies.
- The convergence of digital, physical and artificially generated Web, and the need to use provenance to understand how content is generated
- The possibility of having AIs involved in the management of our Web Science community (e.g. propose programs, topics, etc.)

---

<sup>1</sup> <https://www.mycoted.com/Brainwriting>

- The design of Human-AI interaction and communication; Human-Human, Avatar-Avatar, and AI-AI interactions on the Web
- The goal of regulating subversive Human/AI activity in hybrid solutions
- The provision of an infrastructure for workflows and data as key drivers of Web Science
- The question of how adding AI to the Web re-shapes the future of work on the Web and how the demands of the future of work shapes AI on the Web.

### 4.3.3 Converging on three research directions

The last sessions of the group were dedicated to build a synthesis of the discussions. Three main research directions were drafted to the group identified research questions:

1. **Knowledge Infrastructure and governance of Web:** *extending the Web observatory vision*
  - a. AI infrastructure to study the Web: Using the power of web scale to better understand the Web evolution phenomena.
  - b. AI to detect and counter-attack some undesirable network effects: starting by defining the Web we want, how can AI Help? e.g. consider the “Giant attractors” of the Web either with the global view (crawling) to maintain metrics and then intervene or agents with simple rules pushing emergent behaviors such as an agent purposefully posting links to other platforms to foster linking, decentralizing.
  - c. Workflows and data are key drivers of Web Science;
  - d. Web science AIs should be citable in Web Science papers; Have AIs involved in the management of our Web Science community (propose programs, topics, etc.);
  - e. Explainable AI and linked to accountability
2. **AI in the relation between Web Science and data sets/data lakes (big web science):** *AI for data science on the Web and vice-versa*
  - a. Data science to help solve problems of data sets of AI (bias, etc.) : automate scrutiny on the Web and generate reports for Web scientists, and vice versa AI simulations avoiding privacy implications and issues; Web exploration powered by AI; possible simulation and synthetic data;
  - b. Reproducible research and data set sharing should be a core feature in Web Science research and the need to use provenance to understand how content is generated;
  - c. Data sharing architectures / data trust archives/reserves, met searching, search across sources; Web of archives and search across that Web;
  - d. Reproducibility and secondary use of the data sets
3. **Designing and studying Intelligence forms and Hybrid Web Societies:** *designing artificial Web intelligence*
  - a. Different AI forms: machine learning, knowledge representation and reasoning, etc.;
  - b. The “laws of Web AI” (cf Asimov) since a Web AI is at least as dangerous as a robot. Ethical principles (of Web Science research)?; normative systems for hybrid societies Human-AI interaction and communication;
  - c. Human-Human, Avatar-Avatar, and AI-AI interactions; Regulating subversive Human/AI activity in hybrid solutions; Convergence of digital, physical and artificially-generated Web; How adding AI to the Web, re-shapes the future of work on the Web and how the demands of the future of work shapes AI on the Web
  - d. Involve designers, interaction designers, HCI, interaction with AIs;
  - e. AI to observe AI, AI watchdogs checking on other AIs.

The group also identified a number of general considerations and transversal topics:

- The manifesto could include a lexicon/glossary of preferred terms e.g. “ethically designed agents” instead of “ethical agent”;
- Alternatively the manifesto could select and explain examples of ambiguous terms and expressions to help interdisciplinary interactions;
- Should Web AI be driven by the individual values of the different communities that we represent or do we need a new set of values for the Web?

## 5 Panel Discussions

### 5.1 Closing the Loop: a panel discussion moderated by Susan

In this final session, an interdisciplinary panel of speakers was asked to reflect on the future for Web Science. The panel was comprised of:

- Dr Rob Ackland, the Australian National University, Australia
- Dr Kemal Delic, Industry and Academia
- Prof. Dave de Roure, Oxford University, UK
- Dr Oshani Seneviratne, Rennesler Polytechnic Institute, USA
- Prof. Gugliemlo Tamburinni, University of Naples Federico II, Italy

The discussion included the following points:

- the importance of maintaining methodological breadth in Web Science, to include both critical and applied methodologies;
- the importance of looking ahead, to emerging technologies and the changing formation of the online interactions, e.g. through blockchain and the internet of things;
- the importance of a ‘level playing field’ for data sharing (where at present this is largely limited to bi-lateral relations between major companies and a small number of ivy league Universities; the importance of developing new trusted mechanisms for data sharing, e.g. data trust.

#### 5.1.1 Statements

##### 5.1.1.1 David De Roure

I have recently been studying the broader landscape of social data science and computational social science, and considering how Web Science sits in this interdisciplinary ecosystem. The answer, not surprisingly but very importantly, is that Web Science is about the Web, i.e. a unique socially-constituted system with at least half of humanity participating. Surely this demands study in its own right: Web as an object, an evolving arefact.

During the seminar I have appreciated the nuanced discussions about ethics and normative aspects, which must be part of the study—I am pleased that this is part of the Web Science mission, and that our conversations have moved on from the early strapline “ensuring the social benefit of the Web”.

My new insights gained this week are about the various relationships between AI and the Web: AI in the everyday practice of the Web Scientist, AI as automated web-scale Web Scientist, and Web Science studying a Web with AI inside. We need all three going forward, but I have found these distinctions useful.

Going forward I think it would be useful to discuss creativity in the Web, which is an aspect not widely addressed in Web Science. It is also topical in the context of AI, as we address creative computing but also computational creativity. There would be value in bringing humanists into our discussions.

#### 5.1.1.2 Oshani Seneviratne

I would like to bring to attention an emerging field waiting to be explored by web scientists. Applying Web Science research methodologies for understanding the blockchain ecosystem is a subject that has fascinated me recently. Just like the Web in the early 90's, blockchain technologies are now going through a 10-20 year maturity phase. There are many parallels between the Web and the blockchain. They are both disruptive technologies that has gained massive user adoption through open architectures that promote a giant connected component, and they are both hard to quantify because of their massive scale. Similar to the early Web days, we are starting to see the rise of many applications developed using 'smart contracts' deployed on the blockchain. Some example applications include: Sapien (a democratized social news platform), Steemit (Reddit equivalent), SOLA (a decentralized Social Network with over 700,000 active users), and Indorse (a LinkedIn equivalent, where AI bots can even evaluate your skills). We are starting to see lot of public data generated through 'social' interactions on the blockchain using such applications, even though the original usecase for the blockchain was cryptocurrency.

Therefore, very much like the Web, we will start seeing the need for the next generation of scientists informed by methodologies that are similar to the ones used in Web Science to study blockchain based ecosystems. However, there will also be some challenges, as gathering the data from truly decentralized ecosystems such as those powered by the blockchain will be more challenging. But on the other hand, data sharing using the blockchain will lead to much more transparent and accountable research practises due to the immutable ledger used for data sharing transactions, which will in turn promote good data stewardship. In conclusion, it is my belief that Web Science researchers should pay some attention to this emerging area of research of blockchain based ecosystems; it is a treasure trove in uncharted territory waiting to be explored!

#### 5.1.1.3 Kemal A. Delic

We are living in the age of 'BIG systems': Big Data, Big Infrastructures and Big Algorithms are omnipresent, always on and serving billions of users dispersed around globe. At very abstract level, this represent immensely complex cyber-physical systems which I would call 'fabrics'. This resembles a huge living organism characterized by the large diversity, high dynamics and constantly evolving behavior. Social scientists explored and tried to explain wide variety of Web phenomena during the last decade, but a new playing field is waiting for the scientific breakthroughs, deeper insights and surprising discoveries. Torrents of data flow around high-speed networks, keeping data centers around globe busy and quickly consuming space and filling up data repositories for analytic usage later. Present reality is the peta-scale world, whilst recent advances here will be heralding the exa-scale range. Those data volumes will require special algorithms – and most likely will be AI based – to enable the rise of an entire new kind of instrument – Exascale Scientific Computing (ESC). This will also represent the next big challenge for academic, industrial and government communities. It is envisaged that such kind of systems will become reality in 2020-2030-time horizon. Ultimately, this will be an extraordinary opportunity for the power of Web to explore and analyze itself for scientific, industrial and commercial purposes. Just as the invention of the microscope and the telescope have changed the course of humanity, I am strongly convinced that a web instrument will provide the foundations for the next chapter of Web Science history – evolving into the exciting, intricate and mysterious sphere of the Science of Complex Hyperscale Systems.



## Participants

- Robert Ackland  
Australian National University –  
Canberra, AU
- Ricardo A. Baeza-Yates  
NTENT – Carlsbad, US &  
Northeastern Univ., US
- Bettina Berendt  
KU Leuven, BE
- Noshir S. Contractor  
Northwestern University –  
Evanston, US
- David De Roure  
University of Oxford, GB
- Kemal A. Delic  
Hewlett Packard – Grenoble, FR
- Nikolaus Forgó  
Universität Wien, AT
- Fabien Gandon  
INRIA Sophia Antipolis, FR
- Susan Halford  
University of Southampton, GB
- Wendy Hall  
University of Southampton, GB
- Lynda Hardman  
CWI – Amsterdam and Utrecht  
University, NL
- Andreas Hotho  
Universität Würzburg, DE
- Katharina E. Kinder-Kurlanda  
GESIS – Köln, DE
- Claudia Müller-Birn  
FU Berlin, DE
- Wolfgang Nejdl  
Leibniz Universität Hannover,  
DE
- Eirini Ntoutsi  
Leibniz Universität  
Hannover, DE
- Paolo Parigi  
Airbnb and Stanford University –  
San Francisco, US
- Evaggelia Pitoura  
University of Ioannina, GR
- Oshani Seneviratne  
Rensselaer Polytechnic Institute –  
Troy, US
- Elena Simperl  
University of Southampton, GB
- Steffen Staab  
Universität Koblenz-Landau, DE
- Guglielmo Tamburrini  
University of Naples  
Federico II, IT
- Pinelopi Troullinou  
The Open University –  
Milton Keynes, GB

