DAGSTUHL
REPORTS

Volume 7, Issue 10, October 2017

Dagstuhl Reports, Vol. 7, Issue 10 ISSN 2192-5283

ISSN 2192-5283

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

Publication date
March, 2018

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

Digital Object Identifier: 10.4230/DagRep.7.10.i

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

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

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

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

Report from Dagstuhl Seminar 17401

Quantum Cryptanalysis

Edited by

Michele Mosca[1], Nicolas Sendrier[2], Rainer Steinwandt[3], and Krysta Svore[4]

1 **University of Waterloo, CA,** michele.mosca@uwaterloo.ca
2 **INRIA – Paris, FR,** nicolas.sendrier@inria.fr
3 **Florida Atlantic University – Boca Raton, US,** rsteinwa@fau.edu
4 **Microsoft Corporation – Redmond, US,** ksvore@microsoft.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17401 "Quantum Cryptanalysis." We start out by outlining the motivation and organizational aspects of the seminar. Thereafter, abstracts of presentations given by seminar participants are provided.

Seminar October 1–6, 2017 – http://www.dagstuhl.de/17401
1998 ACM Subject Classification E.3 Code Breaking, F.2 Analysis of Algorithms and Problem Complexity, G.2 Discrete Mathematics
Keywords and phrases computational algebra, post-quantum cryptography, quantum circuit complexity, quantum computing, quantum hardware and resource estimation
Digital Object Identifier 10.4230/DagRep.7.10.1
Edited in cooperation with John M. Schanck

1 Executive Summary

Michele Mosca
Nicolas Sendrier
Rainer Steinwandt
Krysta Svore

Motivation and scope

Like its predecessors, this fourth installment of a Dagstuhl seminar on *Quantum Cryptanalysis* was devoted to studying cryptographic solutions that might be suitable for standardization in the post-quantum setting and to studying quantum attacks against currently deployed cryptographic solutions. Two main thrusts were of particular interest:

Algorithmic innovation. Quantum resources can be used in various way for attacking cryptographic solutions, and the seminar included multiple presentations on exploiting quantum resources for cryptanalytic purposes. Both attacks on symmetric and asymmetric primitives were considered, and there were lively discussions on the feasibility of mounting particular types of attacks. Complementing the presentations on quantum attacks, the program included presentations on advanced classical algorithms, raising the question of identifying possibilities to speed up such classical attack venues through quantum "subroutines."

Quantum resource estimation. It goes without saying that asymptotic improvements are of great interest when trying to tackle computational problems underpinning the security of cryptographic constructions. However, when looking at an actually deployed scheme, quantifying the exact resources (such as the number of qubits) needed by an attacker is relevant to judge the practical impact of a proposed attack strategy. The seminar included presentations on the estimation of resources for attacking some prominent cryptographic schemes.

As expected from a seminar with this title, many talks were indeed devoted to cryptanalysis, but the program also included presentations on establishing provable security guarantees in a post-quantum scenario. With the field becoming more mature, we did not schedule much time for survey talks. However, we did include a presentation on the *status of the development of quantum computers* in the program, thereby helping to get a better idea of potential obstacles when trying to implement quantum cryptanalytic attacks.

Organization

This was the fourth Dagstuhl seminar devoted entirely to quantum cryptanalysis, and as in the prior editions the set of participants included both experts in quantum algorithms and experts in classical cryptography. Some of the participants had already participated in earlier editions of this seminar series, but a number of colleagues attended such a seminar — or any Dagstuhl event — for the first time. In total, we had 42 participants from academia, government, and industry. This time we also included an open problem session in the program, which will hopefully help to stimulate further work in this vibrant research area. In the schedule we tried to leave sufficient time for discussions and for collaborative work in smaller groups. In line with the Dagstuhl tradition, no presentations were scheduled for Wednesday afternoon, and the seminar participants could devote the afternoon to a hike, an excursion, or to their research.

Results and next steps

Over the course of the years, communication and collaboration between the classical cryptographic and the quantum algorithmic research communities has intensified, and many colleagues cross traditional discipline boundaries. As evidenced in the seminar, available quantum cryptanalytic results can go well beyond asymptotic statements and include rather fine-grained resource counts. The seminar covered the analysis of both symmetric and asymmetric primitives, and ongoing efforts toward standardizing quantum-safe cryptographic solutions are likely to stimulate more progress, in particular on the quantum cryptanalysis of asymmetric cryptographic primitives.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Quantum-Secure Symmetric-Key Cryptography Based on Hidden Shifts

Gorjan Alagic (University of Maryland - College Park, US)

Joint work of Gorjan Alagic, Alexander Russell
Main reference Gorjan Alagic, Alexander Russell: "Quantum-Secure Symmetric-Key Cryptography Based on Hidden Shifts", CoRR, Vol. abs/1610.01187, 2016.
URL http://arxiv.org/abs/1610.01187

Recent results of Kaplan et al., building on previous work by Kuwakado and Morii, have shown that a wide variety of classically-secure symmetric-key cryptosystems can be completely broken by quantum chosen-plaintext attacks (qCPA). In such an attack, the quantum adversary has the ability to query the cryptographic functionality in superposition. The vulnerable cryptosystems include the Even-Mansour block cipher, the three-round Feistel network, the Encrypted-CBC-MAC, and many others. In this work, we study simple algebraic adaptations of such schemes that replace $(\mathbb{Z}/2)^n$ addition with operations over alternate finite groups–such as $\mathbb{Z}/(2^n)$–and provide evidence that these adaptations are qCPA-secure. These adaptations furthermore retain the classical security properties (and basic structural features) enjoyed by the original schemes. We establish security by treating the (quantum) hardness of the well-studied Hidden Shift problem as a basic cryptographic assumption. We observe that this problem has a number of attractive features in this cryptographic context, including random self-reducibility, hardness amplification, and–in many cases of interest–a reduction from the "search version" to the "decisional version." We then establish, under this assumption, the qCPA-security of several such Hidden Shift adaptations of symmetric-key constructions. We show that a Hidden Shift version of the Even-Mansour block cipher yields a quantum-secure pseudorandom function, and that a Hidden Shift version of the Encrypted CBC-MAC yields a collision-resistant hash function. Finally, we observe that such adaptations frustrate the direct Simon's algorithm-based attacks in more general circumstances, e.g., Feistel networks and slide attacks.

3.2 Improved Combinatorial Algorithms for the Inhomogeneous Short Integer Solution Problem

Shi Bai (Florida Atlantic University - Boca Raton, US)

Joint work of Shi Bai, Steven D. Galbraith, Liangze Li, Daniel Sheffield
Main reference Shi Bai, Steven D. Galbraith, Liangze Li, Daniel Sheffield: "Improved Combinatorial Algorithms for the Inhomogeneous Short Integer Solution Problem", IACR Cryptology ePrint Archive, Report 2014/593, 2014.
URL https://eprint.iacr.org/2014/593.pdf

We discuss algorithms for the inhomogeneous short integer solution problem: Given (A, s) to find a short vector x such that $Ax \equiv s \pmod{q}$. We consider algorithms for this problem due to Camion and Patarin; Wagner; Schroeppel and Shamir; Minder and Sinclair; Howgrave-Graham and Joux (HGJ); Becker, Coron and Joux (BCJ). Our main results include: Applying the Hermite normal form (HNF) to get faster algorithms; A heuristic analysis of the HGJ and BCJ algorithms in the case of density greater than one; An improved cryptanalysis of the SWIFFT hash function; A new method that exploits symmetries to speed up algorithms for Ring-ISIS.

3.3 Low-communication parallel quantum multi-target preimage search

Gustavo Banegas (TU Eindhoven, NL)

Joint work of Gustavo Banegas, Daniel J. Bernstein
Main reference Gustavo Banegas, Daniel J. Bernstein: "Low-communication parallel quantum multi-target preimage search", IACR Cryptology ePrint Archive, Report 2017/789, 2017.
URL https://eprint.iacr.org/2017/789

The most important pre-quantum threat to AES-128 is the 1994 van Oorschot–Wiener "parallel rho method", a low-communication parallel pre-quantum multi-target preimage-search algorithm. This algorithm uses a mesh of p small processors, each running for approximately $2^{128}/pt$ fast steps, to find one of t independent AES keys $k_1, \ldots, k_t$, given the ciphertexts $\mathrm{AES}_{k_1}(0), \ldots, \mathrm{AES}_{k_t}(0)$ for a shared plaintext 0. NIST has claimed a high post-quantum security level for AES-128, starting from the following rationale: "Grover's algorithm requires a longrunning serial computation, which is difficult to implement in practice. In a realistic attack, one has to run many smaller instances of the algorithm in parallel, which makes the quantum speedup less dramatic." NIST has also stated that resistance to multi-key attacks is desirable; but, in a realistic parallel setting, a straightforward multi-key application of Grover's algorithm costs more than targeting one key at a time. This paper introduces a different quantum algorithm for multi-target preimage search. This algorithm shows, in the same realistic parallel setting, that quantum preimage search benefits asymptotically from having multiple targets. The new algorithm requires a revision of NIST's AES- 128, AES-192, and AES-256 security claims

3.4 Classical Proofs for the Quantum Security of Classical Hash Functions

Serge Fehr (CWI - Amsterdam, NL)

Hash functions are of fundamental importance in cryptography, and with the threat of quantum computers possibly emerging in the future, it is an urgent objective to understand the security of cryptographic hash functions in the light of potential future quantum attacks. To this end, we reconsider the notion of the so-called collapsing property of hash functions, as introduced by Unruh, which replaces the notion of collision resistance when considering quantum attacks. Our contribution is the introduction of a framework that offers significantly simpler proofs for the collapsing property of hash functions. With our framework, we can prove the collapsing property for hash domain extension constructions entirely by means of decomposing the iteration function into suitable elementary composition operations. In particular, given our framework, one can argue purely classically about the quantum-security of hash functions; this is in contrast to previous proofs which are in terms of sophisticated quantum-information-theoretic and quantum-algorithmic reasoning. This is work in progress.

3.5 Towards cryptographic applications of quantum walks

Peter Høyer (University of Calgary, CA)

Joint work of Cătălin Dohotaru, Peter Høyer, Mojtaba Komeili
Main reference Peter Høyer, Mojtaba Komeili: "Efficient Quantum Walk on the Grid with Multiple Marked Elements", in Proc. of the 34th Symposium on Theoretical Aspects of Computer Science, STACS 2017, March 8-11, 2017, Hannover, Germany, LIPIcs, Vol. 66, pp. 42:1–42:14, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.
URL http://dx.doi.org/10.4230/LIPIcs.STACS.2017.42

We give an introduction to quantum walks, including Szegedy's correspondence between random walks and quantum walks. We survey the main known general quantum walk algorithms. We discuss properties of three of the most commonly used classes of graphs in quantum algorithms and protocols: the complete graphs, the Johnson graphs, and the tori graphs. We show how a quantum walk on a Johnson graph is used in an algorithm for the computational problem of element distinctness. We show how a quantum walk on a torus is used in protocols for the 2-party communication problem of disjointness. We discuss and speculate on future potential uses of quantum walks in cryptographic protocols. We end with pointers to literature and surveys.

3.6 Post-quantum security of the sponge construction

Andreas Hülsing (TU Eindhoven, NL)

Joint work of Jan Czajkowski, Leon Groot Bruinderink, Andreas Hülsing, Christian Schaffner, Dominique Unruh
Main reference Jan Czajkowski, Leon Groot Bruinderink, Andreas Hülsing, Christian Schaffner, Dominique Unruh: "Post-quantum security of the sponge construction", IACR Cryptology ePrint Archive, Report 2017/771, 2017.
URL https://eprint.iacr.org/2017/771

We investigate the post-quantum security of hash functions based on the sponge construction. A crucial property for hash functions in the post-quantum setting is the collapsing property (a strengthening of collision-resistance). We show that the sponge construction is collapsing (and in consequence quantum collision-resistant) under suitable assumptions about the underlying block function. In particular, if the block function is a random function or a (non-invertible) random permutation, the sponge construction is collapsing.

3.7 Random self-reducibility for SIDH

David Jao (University of Waterloo, CA)

Joint work of David Jao, David Urbanik

We discuss preliminary results concerning the self-reducibility of Supersingular Isogeny Diffie-Hellman (SIDH) problem instances over the same base field.

3.8 Connections between Learning with Errors and the Dihedral Coset Problem

Elena Kirshanova (ENS - Lyon, FR)

Joint work of Zvika Brakerski, Elena Kirshanova, Damien Stehlé, Weiqiang Wen
Main reference Zvika Brakerski, Elena Kirshanova, Damien Stehlé, Weiqiang Wen: "Learning With Errors and Extrapolated Dihedral Cosets", CoRR, Vol. abs/1710.08223, 2017.
URL http://arxiv.org/abs/1710.08223

In this talk I explained the result that shows that under quantum polynomial time reductions, LWE is equivalent to a relaxed version of the dihedral coset problem (DCP), which is called extrapolated DCP (eDCP). The extent of extrapolation varies with the LWE noise rate. By considering different extents of extrapolation, the result generalizes Regev's famous proof that if DCP is in BQP (quantum poly-time) then so is LWE (FOCS 02). I also discussed a connection between eDCP and Childs and Van Dam's algorithm for generalized hidden shift problems (SODA 07).The result implies that a BQP solution for LWE might not require the full power of solving DCP, but rather only a solution for its relaxed version, eDCP, which could be easier.

3.9 Quantum Cryptanalysis of Block Ciphers: A Case Study

Yi-Kai Liu (NIST - Gaithersburg, US)

Joint work of Brittanney Amento-Adelmann, Markus Grassl, Brandon Langenberg, Yi-Kai Liu, Eddie Schoute, Rainer Steinwandt

Quantum computers can achieve a quadratic speedup over classical computers, when performing an exhaustive key search against a block cipher. This quantum attack uses Grover's algorithm, and it requires the implementation of the target cipher using reversible logic, so that it can be run on a superposition of different inputs.

We report quantum circuits and resource bounds for several well-known block ciphers: MARS, SERPENT, Simon, and Speck. We find that quantum cryptanalysis of Simon and Speck is feasible on relatively small quantum computers, with a few hundred logical qubits, and 10^5 to 10^6 quantum gates per Grover iteration. This is a consequence of the design of those ciphers, using large numbers of simple "ARX" operations (e.g., addition mod 2^n, bit-rotation, and bitwise XOR). SERPENT, which uses small S-boxes, has somewhat different resource requirements. At the other extreme, MARS requires much larger quantum circuits, due to its complex internal structure and its large pseudorandom S-boxes.

3.10 Grover Meets Simon - Quantumly Attacking the FX-construction

Alexander May (Ruhr-Universität Bochum, DE)

Joint work of Gregor Leander, Alexander May
Main reference Gregor Leander, Alexander May: "Grover Meets Simon - Quantumly Attacking the FX-construction", in Proc. of the Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 10625, pp. 161–178, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-70697-9_6

Using whitening keys is a well understood mean of increasing the key-length of any given cipher. Especially as it is known ever since Grover's seminal work that the effective key-length is reduced by a factor of two when considering quantum adversaries, it seems tempting to use this simple and elegant way of extending the key-length of a given cipher to increase the resistance against quantum adversaries. However, as we show in this work, using whitening keys does not increase the security in the quantum-CPA setting significantly. For this we present a quantum algorithm that breaks the construction with whitening keys in essentially the same time complexity as Grover's original algorithm breaks the underlying block cipher. Technically this result is based on the combination of the quantum algorithms of Grover and Simon for the first time in the cryptographic setting.

3.11 New Results on Symmetric Quantum Cryptanalysis

Maria Naya-Plasencia (INRIA - Paris, FR)

Joint work of André Chailloux, María Naya-Plasencia, André Schrottenloher
Main reference André Chailloux, María Naya-Plasencia, André Schrottenloher: "An Efficient Quantum Collision Search Algorithm and Implications on Symmetric Cryptography", in Proc. of the Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 10625, pp. 211–240, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-70697-9_8

In this talk I will present some recent results on symmetric quantum cryptanalysis: a new efficient quantum collision search algorithm (joint work with A. Chailloux and A. Schrottenloher) and an extensive analysis of the use of modular additions on symmetric primitives (joint work with X. Bonnetain).

3.12 Thermodynamic Analysis of Classical and Quantum Algorithms for Preimage and Collision Search Problems

Ray Perlner (NIST - Gaithersburg, US)

We analyze the performance of classical and quantum search algorithms from a thermodynamic perspective, focusing on resources such as time, energy, and memory size. We consider two examples that are relevant to post-quantum cryptography: Grover's search algorithm, and the quantum algorithm for collision-finding. Using Bennett's "Brownia" model of low-power reversible computation, we show classical algorithms that have the same asymptotic energy consumption as these quantum algorithms. Thus, the quantum advantage in query complexity does not imply a reduction in these thermodynamic resource costs. In addition, we present realistic estimates of the resource costs of quantum and classical search, for near-future computing technologies. We find that, if memory is cheap, classical exhaustive search can be surprisingly competitive with Grover's algorithm.

3.13 Quantum resource estimates for computing elliptic curve discrete logarithms

Martin Roetteler (Microsoft Corporation - Redmond, US)

Joint work of Martin Roetteler, Michael Naehrig, Krysta M. Svore, Kristin Lauter
Main reference Martin Roetteler, Michael Naehrig, Krysta M. Svore, Kristin E. Lauter: "Quantum resource estimates for computing elliptic curve discrete logarithms", CoRR, Vol. abs/1706.06752, 2017.
URL http://arxiv.org/abs/1706.06752

We give precise quantum resource estimates for Shor's algorithm to compute discrete logarithms on elliptic curves over prime fields. The estimates are derived from a simulation of a Toffoli gate network for controlled elliptic curve point addition, implemented within the framework of the quantum computing software tool suite Liquid. We determine circuit implementations for reversible modular arithmetic, including modular addition, multiplication and inversion, as well as reversible elliptic curve point addition. We conclude that elliptic curve discrete logarithms on an elliptic curve defined over an n-bit prime field can be computed on a quantum computer with at most $9n + 2\lceil\log_2(n)\rceil + 10$ qubits using a quantum circuit of at most $448n^3 \log_2(n) + 4090n^3$ Toffoli gates. We are able to classically simulate the Toffoli networks corresponding to the controlled elliptic curve point addition as the core piece of Shor's algorithm for the NIST standard curves P-192, P-224, P-256, P-384 and P-521. Our approach allows gate-level comparisons to recent resource estimates for Shor's factoring algorithm. The results also support estimates given earlier by Proos and Zalka and indicate that, for current parameters at comparable classical security levels, the number of qubits required to tackle elliptic curves is less than for attacking RSA, suggesting that indeed ECC is an easier target than RSA.

3.14 Factoring integers by algorithms for lattice reduction

Claus Peter Schnorr (Goethe-Universität - Frankfurt am Main, DE)

URL http://www.math.uni-frankfurt.de/ dmst/research/papers/SVP9.pdf

We factor an integer N by enumeration algorithms that find vectors of the prime number lattice $\mathcal{L}(\mathbf{B}_{n,c})$ close to a specific target vector $\mathbf{N}_c$ representing N. The algorithm NewEnum performs the stages of exhaustive enumeration of close, respectively short lattice vectors in order of decreasing success rate, stages with high success rate are done first. These algorithms generate for the n-th prime p_n triples of p_n-smooth integers $u, v, |u - vN|$ that factorize the integer N. An integer N can be factored by about $n + 1$ p_n-smooth triples $u, v, |u - vN|$. Our CVP-algorithm generates for $n = 90$, $n + 1$ such relations and factors $N \approx 1014$ in 6.2 seconds. We consider extensions to large N.

3.15 Code based cryptography and quantum attacks

Jean-Pierre Tillich (INRIA - Paris, FR)

This talk is a survey of how quantum algorithms can be used to speed up the fundamental problem on which code-based cryptography relies, namely the decoding problem. I will first cover Bernstein's algorithm which uses Grover to get a quantum speed-up over the simplest information set decoding algorithm, namely the Prange algorithm. Then I will cover my work with Kachigar on how using quantum walk techniques with Grover to get a further quantum speed-up. I will also talk about a result of Grilo and Kerenidis using quantum Fourier techniques for solving the LPN/LWE problem in polynomial time if we have access to a quantum superposition of all LPN samples.

3.16 Security of Fiat-Shamir

Dominique Unruh (University of Tartu, EE)

Main reference Dominique Unruh: "Post-quantum Security of Fiat-Shamir", in Proc. of the Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 10624, pp. 65–95, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-70694-8_3

We describe how classical security proofs for Fiat-Shamir go wrong in the quantum setting, and what can be done about it.

3.17 Status of the development of quantum computers

Frank K. Wilhelm (Universität des Saarlandes - Saarbrücken, DE)

I am reviewing three major current routes to quantum correction. For cryptanalysis, fault-tolerant quantum computing with the surface code is currently the only viable way. I am proposing a five-tier evaluation system for the status of quantum computing implementations and describe candidates for the lowest three. I try to speculate on how to scale up to 1 MQubits.

4 Open problems

The discussion started with presentations of various challenge problem web pages.

Multivariate-quadratic challenges. Tsuyoshi Takagi (University of Tokyo, JP) started with an overview of the Fukuoka MQ challenge page, for solving multivariate quadratic polynomial challenges: www.mqchallenge.org. There are three sub-families of challenges, including three encryption schemes and three signature schemes (over finite fields of over 2, 2^8 and 31), and a Hall of Fame for each of the six families of schemes.
The coefficients of the challenges are generated from the digits of Π, however for the encryption schemes the MQ challenge team provide one random answer to guarantee a solution and thus the team does not participate in the three encryption challenges (but can participate in the signature challenges).

Lattice challenges. Johannes Buchmann (TU Darmstadt, DE) presented the lattice challenge web pages at latticechallenge.org. The main page outlines the challenge consisting of lattice bases for which the solution of SVP implies a solution of SVP in all lattices of a certain smaller dimension. There are two ways to enter the Hall of Fame:

- Find a short vector in a new challenge dimension
- Find an even shorter vector in one of the dimensions already listed in the hall of fame.

Wild McEliece challenges. Tanja Lange (TU Eindhoven, NL) presented the wild McEliece challenges at pqcrypto.org. The web page also lists other post-quantum challenge pages.

Other challenges. The NTRU challenges (www.onboardsecurity.com/products/ntru-crypto/ntru-challenge) and R-LWE challenges (web.eecs.umich.edu/~cpeikert/rlwe-challenges/) were also mentioned.

The discussion continued with presentations of other open problems.

- Yi-Kai Liu (NIST – Gaithersburg, US) highlighted a paper by Kimmel, Lin and Lin on “Oracles with costs” as a possible tool for quantum cryptanalysis. arXiv:1502.02174
- Phong Nguyen (University of Tokyo, JP) presented some open questions related to sampling vectors in a lattice.
- John Schanck (University of Waterloo, CA) presented a variant of the subset sum problem where the collection of numbers is a subset chosen from a much larger set of randomly sampled numbers.

Participants

- Gorjan Alagic
University of Maryland – College Park, US
- Shi Bai
Florida Atlantic University – Boca Raton, US
- Gustavo Banegas
TU Eindhoven, NL
- Daniel J. Bernstein
University of Illinois – Chicago, US
- Jean-François Biasse
University of South Florida – Tampa, US
- Alexei Bocharov
Microsoft Corporation – Redmond, US
- Johannes A. Buchmann
TU Darmstadt, DE
- Yfke Dulek
CWI – Amsterdam, NL
- Serge Fehr
CWI – Amsterdam, NL
- Tommaso Gagliardoni
IBM Research Zurich, CH
- Vlad Gheorghiu
University of Waterloo, CA
- Maria Isabel González Vasco
King Juan Carlos University – Madrid, ES
- Sean Hallgren
Pennsylvania State University – University Park, US
- Peter Hoyer
University of Calgary, CA
- Andreas Hülsing
TU Eindhoven, NL
- David Jao
University of Waterloo, CA
- Stacey Jeffery
CWI – Amsterdam, NL
- Elena Kirshanova
ENS – Lyon, FR
- Stavros Kousidis
BSI – Bonn, DE
- Thijs Laarhoven
IBM Research Zurich, CH
- Bradley Lackey
University of Maryland – College Park, US
- Tanja Lange
TU Eindhoven, NL
- Yi-Kai Liu
NIST – Gaithersburg, US
- Alexander May
Ruhr-Universität Bochum, DE
- Michele Mosca
University of Waterloo, CA
- Michael Naehrig
Microsoft Research – Redmond, US
- Anderson Nascimento
University of Washington – Tacoma, US
- Maria Naya-Plasencia
INRIA – Paris, FR
- Phong Q. Nguyen
University of Tokyo, JP
- Ray Perlner
NIST – Gaithersburg, US
- Martin Roetteler
Microsoft Corporation – Redmond, US
- Alexander Russell
University of Connecticut – Storrs, US
- John M. Schanck
University of Waterloo, CA
- Claus Peter Schnorr
Goethe-Universität – Frankfurt am Main, DE
- Nicolas Sendrier
INRIA – Paris, FR
- Daniel Smith-Tone
NIST – Gaithersburg, US
- Rainer Steinwandt
Florida Atlantic University – Boca Raton, US
- Adriana Suárez Corona
University of León, ES
- Tsuyoshi Takagi
University of Tokyo, JP
- Jean-Pierre Tillich
INRIA – Paris, FR
- Dominique Unruh
University of Tartu, EE
- Frank K. Wilhelm
Universität des Saarlandes – Saarbrücken, DE

Report from Dagstuhl Seminar 17411

Hyperspectral, Multispectral, and Multimodal (HMM) Imaging: Acquisition, Algorithms, and Applications

Edited by

Gonzalo R. Arce[1], Richard Bamler[2], Jon Yngve Hardeberg[3], Andreas Kolb[4], and Shida Beigpour[5]

1 **University of Delaware, US,** arce@ece.udel.edu
2 **DLR - Oberpfaffenhofen, DE,** richard.bamler@dlr.de
3 **Norwegian Univ. of Science & Technology, NO,** jon.hardeberg@ntnu.no
4 **Universität Siegen, DE,** andreas.kolb@uni-siegen.de
5 **MPI für Informatik - Saarbrücken, DE,** shida@mpi-inf.mpg.de

Abstract

In the last couple of decades, hyperspectral, multispectral, and multimodal (HMM) imaging has emerged as an essential tool in various fields of science, medicine, and technology. Compared to integrated broad-band information as, e.g., present in RGB images, HMM imaging strives to acquire a multitude of specific narrow bands of the electromagnetic spectrum in order to solve specific detection or analysis tasks. HMM research is interested in studying light-matter interaction in a wide range of wavelengths from the high energy radiation down to Terahertz radiation (sub-millimeter waves). Furthermore, combining spectral data captured using different imaging modalities can unveil additional information of the scene that is not revealed solely by each of the individual imaging modalities.

The workshop intended to connect researchers from different disciplines that involve HMM imaging and analysis. Even though there are very different approaches towards HMM imaging research and application, the main hypothesis of the workshop was that there is a large amount of common goals, approaches and challenges. Thus, these disciplines will benefit from intensifying communication and knowledge transfer and an out-of-the-box thinking and a broader vision of the fundamental concepts regarding common fields of interest, e.g., in the configuration of HMM acquisition systems, data analysis, and improved development techniques by common software bases and validation tools.

The seminar succeeded in bringing together researchers from different scientific communities and fostering open-minded discussions across very different fields of research and application.

Seminar October 8–13, 2017 – http://www.dagstuhl.de/17411

1998 ACM Subject Classification Digitization and Image Capture, Segmentation, Scene analysis, Picture/ image generation, Multidimensional, Earth and atmospheric sciences, Life and medical sciences

Keywords and phrases compressive sensing, computer vision, hyperspectral and multispectral imaging and analysis, multi-modal sensor fusion, remote sensing and geoscience

Digital Object Identifier 10.4230/DagRep.7.10.14

Edited in cooperation with Hilda Deborah

Hyperspectral, Multispectral, and Multimodal (HMM) Imaging: Acquisition, Algorithms, and Applications,
Dagstuhl Reports, Vol. 7, Issue 10, pp. 14–42
Editors: Gonzalo R. Arce, Richard Bamler, Jon Yngve Hardeberg, Andreas Kolb, and Shida Beigpour
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Andreas Kolb
Gonzalo R. Arce
Richard Bamler
Shida Beigpour
Hilda Deborah
Jon Yngve Hardeberg

On the last day of the seminar, the attendees had a very intense discussion about the usefulness of the seminar itself, the grand challenges related to the highly interdisciplinary field of research, and the next steps that should be taken in order to further improve on the cross-fertilizing effects in hyperspectral, multispectral, and multimodal imaging.

Take Home Messages

All attendees agreed on the high quality and open mindedness of the discussions at both, the group level, e.g., in the plenary sessions and the working groups, and also on personal level. All participants assess this Dagstuhl seminar as a great success, especially due to the interdisciplinary discussion and the new insights resulting from this. Despite differences of the individual fields present in the seminar, e.g., remote sensing, color reproduction, and material classification, and despite the wide variety of applications such as medical, environmental monitoring, and arts, a large set of common questions and problems could be identified. All attendees highly appreciated the fact, that unlike in conferences, which usually have a rather narrow perspective on HMM challenges and solutions, as they usually address a single community with a very similar perspective on the field, this seminar brought together people with very different points of view.

This Dagstuhl seminar was a starting point of a number of connections that could be established directly, and several mid-, and maybe even long-term collaborations and joint research actions. There have been several highlights related to the full pipeline from data acquisition, via data processing to applications.

Grand Challenges

On the basis of common and interdisciplinary ground setup in this seminar, several challenges have been identified, which the seminar's attendees see as important to be addressed in further research and engineering work.

Data Acquisition Independent of the specific range of the addressed electromagnetic spectrum, the seminar participants see a severe restriction in the usage of HMM sensors due to their inflexibility, e.g., in selecting spectral bands, bulkiness, high calibration efforts, and acquisition speed (see also the working group report on this topic). Enhancing on these limiting factors has the potential to bring about fundamentally new spins in various application domains. Some approaches presented at the workshop have the potential to push back these limits to some degree. On the other hand, most likely there will be no general purpose HMM acquisition device available in the next decades that covers the majority of application requirements. Still, the seminar attendees agree on the importance of enhancing the applicability of existing and future acquisition devices towards more flexible band selection, fast and efficient, (semi-)automated calibration, and, for some applications, compactness. Ideally, future research provides means for an abstract definition of application specific characteristics from which a specific selection and/ or instantiation of an acquisition device can be deduced.

Data Processing and Validation Regarding data processing and validation, three main topics have been discussed: The usefulness and limitations of machine learning and, especially, deep learning (see working group report), the importance of verified and metric data, and the need for a proper reference and benchmarking data set. Even though there are and have been ongoing activities in spectral normalization and validation, e.g., on the level of CIE or other standardization institutions, or in the field of metrology, there still is the lack of widely existing and accepted methods and data even if restricted to specific fields of application.

The seminar participants see a huge potential in all three areas. Still, major obstacles have to be overcome in order to leverage these potentials. In machine learning/ deep learning, one main issue is the lack of guarantees that the results obey specific constraints to, e.g., physical limits or relations. The lack of verified, metric data, and proper reference and benchmarking data, on the other hand, can only be overcome if there is a stronger common basis for best practice within and, even more important, between the HMM sub-disciplines.

Information Exchange The existence of common information bases is tightly linked to the prior point regarding data processing and validation. So far, there are only few options and pseudo-standard for sharing data and algorithms. While there are good examples, e.g., Open CV library in computer vision, setting up this kind of “standard” is, and will be, much harder in the diverse and partially fragmented HMM research domain. Apparently, this chicken-egg problem can only be solved from within the involved research domains themselves by the normative power of fact of the actions taken by the researchers themselves.

Next Steps

Participants discussed various options for future activity as a follow-up on this Dagstuhl seminar. As one essential restriction of the discussion, attendees became aware of their own limitations in knowing all relevant work and requirements existing in the HMM research subfields. Therefore, the obvious approach to enhance the fields’ convergence by publications, e.g., a special issue or book, and/ or workshops has not the highest priority, even though an introductory workshop or piece of literature for 1st year PhD students would be highly appreciated.

However, participants of Dagstuhl seminar see two main options to proceed in order to keep the initiated process of convergence going and to improve on at least two of the three main challenges identified, i.e., regarding data processing and validation and the exchange of information.

HMM Webpage: As there is a severe lack in common information widely used and recognized, the group of researchers who attended Dagstuhl seminar see the potential of a common, web-based information platform.

In this respect, Masahiro Yamaguchi is open to provide the already established web-link multispectral.org and Andreas Kolb will investigate options for setting up and hosting this kind of platform. In any case, this kind of activity needs to rest on several shoulders, thus the attendees are called to follow through with the activities, on the operative level.

Follow-up Dagstuhl Seminar: As Dagstuhl supports follow-up seminars, the attendees agree on the usefulness of having this kind of seminar in order to evaluate the common, interdisciplinary activities that arose from the first Dagstuhl seminar. In case of a new edition of the workshop, participants agree on having more industrial partners involved.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Imaging Spectroscopy in Earth Observation – Sensors, Tasks, and Challenges

Richard Bamler (DLR - Oberpfaffenhofen, DE)

Imaging spectroscopy in Earth observation is used from satellites, airplanes or drones to map land cover/ land use, materials, water quality, soil properties, hazards and risks, etc. Data processing results in maps of either detection or classification or of continuous-valued variables like plant water content or chlorophyll concentration. Particularly stringent requirements are posed from remote sensing of inland waters.

DLR is responsible for two spaceborne hyperspectral instruments and missions, DESIS (launch 2018 onboard the ISS) and the satellite EnMAP (launch 2020). While DESIS operates in the VNIR domain, EnMAP features about 200 channels in both VNIR and SWIR.

Spaceborne multi-/ hyperspectral instruments suffer from the spatial/spectral resolution trade-off. Therefore, data fusion for sharpening is often required. This is only one challenge of hyperspectral data processing among many others, like atmospheric correction, spectral unmixing, possible low SNR, spectral variability, clouds and the lack of sufficient ground truth data for training and validation.

Finally the question is raised how deep learning can solve some of these challenges. This question has been discussed in a subsequent workshop.

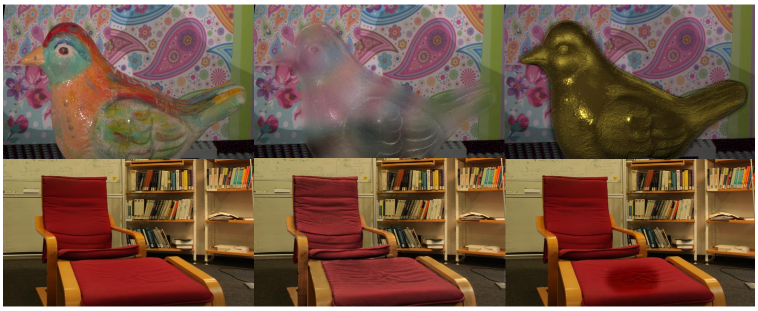
3.2 Hyperspectral Imaging for Image-based Rendering?

Shida Beigpour (MPI für Informatik - Saarbrücken, DE)

Joint work of Beigpour, Shida; Shekhar, Sumit; Myszkowski, Karol; Seidel, Hans-Peter

Photographs and videos are 2D projections of the three-dimensional scene that encode the characteristics of that scene. While such medium is not able to preserve all the aspects of the scene, it still contains valuable information which enables e.g., human subjects to understand the scene. Human perception plays and important role in this context. It has been shown that human perception relies on certain heuristics rather than performing inverse optics.

Image-based inverse rendering techniques use such cues to infer a mid-level representation of the scene known as “intrinsic layers” (i.e., reflectance, shading, and specularity) in order to then be able to modify certain aspects such as materials, illumination, and texture of the objects in the scene. Each intrinsic layer encapsulates certain characteristics of the image allowing for more control over the quality of the results. For example, as specular layer encapsulates the surface gloss, metallic appearance can be rendered by filtering this layer.



We introduce complex perceptual appearance effect (e.g., translucent, gold-plated, weathered, etc.) achieved by signal-based filtering of intrinsic layers. This enables users to interact with the scene in Virtual Reality (VR) and Augmented Reality (AR) applications. So far, spectral imaging has hardly been considered for this task. We show the importance of such data in correctness of our methods as well as datasets and benchmarks.

References

1 Beigpour S., Shekhar S., Myszkowski K., Seidel H.: Light-field Appearance Editing based on Intrinsic Decomposition (2018)

2 Boyadzhiev I., Bala K., Paris S., Adelson E.: Bandsifting decomposition for image-based material editing. ACM Trans. on Graphics 34, 5 (2015).

3 Beigpour S., Kolb A., Kunz S.: A comprehensive multi-illuminant dataset for benchmarking of intrinsic image algorithms. In Proc. IEEE International Conference on Computer Vision (ICCV) (December 2015), pp. 172–180.

3.3 Quantifying Composition of Human Tissues from Multispectral Images using a Physics-based Model of Image Formation

Ela Claridge (University of Birmingham, GB)

Joint work of Cotton, Symon; Preece, Steven; Styles, Iain
Main reference Preece S and Claridge E (2004) Spectral filter optimisation for the recovery of parameters which describe human skin. IEEE Trans. PAMI 26(7), 913-922.
URL http://dx.doi.org/10.1109/TPAMI.2004.36
Main reference Styles IB, Calcagni A, Claridge E, Orihuela Espina F, Gibson JM (2006) Quantitative analysis of multispectral fundus images. Medical Image Analysis 10(4), 578-597.
URL http://dx.doi.org/10.1016/j.media.2006.05.007

Through an understanding of the image formation process, diagnostically important facts about the internal structure and composition of tissues can be derived from their multispectral images. A physics-based model provides a cross-reference between image values and the underlying histological parameters. It is constructed by computing the spectral composition of light remitted from the tissue given parameters specifying its structure and optical properties. Once the model is constructed, for each pixel in a multispectral image its histological parameters can be computed by model inversion. Represented as images, these 'parametric maps' show the concentration of relevant absorbers and volumetric density of scattering structural tissue components. Skin parameters can be recovered with accuracy sufficient for diagnostic use. Retinal imaging poses many challenges, including limited incident illumination, eye movement during multispectral image acquisition, lack of spatial calibration of illumination, lack of flexible tuneable multispectral filters, mathematically and computationally complex inversion and difficulties with validation against histology.

3.4 Is it Possible to Design Optimal Cameras for Robotic Vision?

Donald G. Dansereau (Stanford University, US)

Joint work of Donald G. Dansereau, Daniel Bongiorno, Tom Bridge, Mitch Bryson, Oscar Pizarro, Stefan B. Williams

I give two examples of hyper/multi-spectral imaging in robotics: shark detection from flying robots, and seabed classification from underwater robots. I demonstrate that as few as four colour bands are required to offer increased performance in aerial imaging through water, and even single-pixel spectrometers can improve classification results in autonomous underwater vehicle (AUV)-based seabed survey.

I use these examples to introduce an unresolved problem in robotic vision: How does one design the optimal camera for a given task? Approaches from the robotics and computational imaging communities generally maximize colour fidelity rather than system-level performance, or rely on large training sets and lack generality. Meanwhile, information theory addresses optimal sensing in simple scenarios, but does not fully address the statistics of visual sensing as shown by the dramatic results recently demonstrated in deep learning, which leverages complex learned visual priors.

I raise the question of how one might go about designing optimal cameras for robotic vision, with the hope of uncovering relevant tools and principles from within the HMM community.

3.5 Quality Assessment of Spectral Image Processing Algorithms

Hilda Deborah (Norwegian Univ. of Science & Technology - Gjøvik, NO)

Joint work of Noël Richard, Jon Yngve Hardeberg, Christine Fernandez-Maloigne
Main reference H. Deborah, "Towards spectral mathematical morphology," PhD thesis: Norwegian University of Science & Technology (NTNU) and University of Poitiers, 2016.
URL https://brage.bibsys.no/xmlui/handle/11250/2433012

In the past decades, hyperspectral imaging has been increasingly exploited as it offers a significant gain of accuracy. However, accurate measurements do not entail accurate final processing results. Accuracy can only be obtained when bias, uncertainty, etc., are managed at each level of the processing. Hence the need to enforce metrology to spectral image processing. In my talk, I showed several quality assessment protocols that were designed to metrologically validate spectral difference functions, spectral ordering relations, and morphological crack detection algorithm. Nevertheless, the design of the protocol is generic and can be adapted to other spectral processing algorithms.

3.6 Trends, Issues, and Opportunities in Fusion-related Problems

Nicolas Dobigeon (University of Toulouse, FR)

This talk discusses some open issues related to the problem of fusing multiple images of different spatial and spectral resolutions.

First, it discusses the inverse problem framework, generally considered to conduct this task. In particular, the choice of the regularization is still a challenging question and can be motivated from different points of view: to ease the computations, to promote spatial or spectral features of the fused image, to ensure physically motivated modeling, to exploit outputs from machine learning techniques. Besides, the question of the need for regularization is also discussed since, from a Bayesian perspective, the maximum a posteriori estimation always leads to a trade-off that might lead to unacceptable solution.

Most of the fusion techniques rely on the prior availability of registered, corrected pairs of images to be fused, and possibly on the technical specification of the sensors. In practice, this availability can be limited.

Some opportunities are also discussed. The traditional use case for fusion consists in fusing a pair of optical images of different spatial and/or spectral resolutions acquired at the same date. However, over applicative scenario of interest can appear. For instance, is there any interest to deal with a pair of images without complementarity in terms of spatial and spectral information? How can we fuse more than two images? Moreover, when the images have been acquired at different time instants, detecting changes between these images can be envisioned as a change detection problem. Another open question is to process non-optical data (e.g., SAR images, LiDAR, database). Finally, the main interest of the fused product is questioned, besides visualization perspectives. For a particular task (e.g., classification, detection, unmixing), is there any interest to fuse before this task? Should we design task-driven fusing schemes?

References

1 L. Loncan, L. B. Almeida, et al, "Hyperspectral pansharpening: a review," IEEE Geosci. Remote Sens. Mag., 3(3): 27–46, 2015

2 N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," IEEE Geosci. Remote Sens. Mag., 5(2): 29–56, 2017

3 M. Nikolova, "Model distortions in Bayesian MAP reconstruction," Inverse Problems and Imaging, 1(2): 399–422, 2007

4 R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?" IEEE Trans. Signal Process., 59(5): 2405–2410, 2011

5 V. Ferraris, N. Dobigeon, Q. Wei, and M. Chabert, "Detecting changes between optical images of different spatial and spectral resolutions: a fusion-based approach," IEEE Trans. Geosci. Remote Sens., 2018

6 V. Ferraris, N. Dobigeon, Q. Wei, and M. Chabert, "Robust fusion of multi-band images with different spatial and spectral resolutions for change detection," IEEE Trans. Comput. Imag., 3(2): 175–186, 2017.

3.7 Trends, Issues, and Opportunities in Unmixing-related Problems

Nicolas Dobigeon (University of Toulouse, FR)

This 5-minute talk provides some insights to unmixing-related problems, even in the conventional linear mixing framework.

First, it discusses the choice of the mixing models to be used when conducting unmixing. There are plenty of nonlinear and robust models. An open questions is: How and when should we choose a particular model? A tentative response has been brought for vegetated areas. Moreover, overcoming the inherent spectral variability is also a challenging question. To validate these models and the associated unmixing algorithms, no standard benchmark has been proposed. Moreover, this validation requires the availability of ground-truth data, which is not common.

Then this talk discusses non-standard algorithmic schemes and implementations. Such strategies are generally necessary to face with huge data volume. Multi-temporal image unmixing and hyperspectral video unmixing are promising research issues, which can be tacked off-line, on-line or in a distributed manner.

The question of the supervision of unmixing procedures is also discussed. Most of the research works tend to propose fully unsupervised unmixing procedure. However is it really beneficial? Indeed, in most applicative contexts, some external information is available and can be incorporated.

Finally, one wonders if there is any real interest to unmix, from an end-users point-of-view. For instance, for the mapping of a particular single material, could we design some partial unmixing techniques?

References

1 J. M. Bioucas-Dias, A. Plaza, et al, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," IEEE J. Sel. Topics Appl. Earth Observations Remote Sens., 5(2): 354–379, 2012

2 N. Dobigeon, J.-Y. Tourneret, et al, "Nonlinear unmixing of hyperspectral images: Models and algorithms," IEEE Signal Process. Mag., 31(1): 89–94, 2014

3 N. Dobigeon, L. Tits, et al, "A comparison of nonlinear mixing models for vegetated areas using simulated and real hyperspectral data," IEEE J. Sel. Topics Appl. Earth Observations Remote Sens., 7(6): 1869–1878, 2014

4 A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," IEEE Signal Process. Mag., vol. 31, no. 1, pp. 95–104, Jan. 2014.

5 P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Online unmixing of multitemporal hyperspectral images accounting for spectral variability," IEEE Trans. Image Process., 25(9): 3979–3990, 2016

6 P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Partially asynchronous distributed unmixing of hyperspectral images," Submitted

7 A. Lagrange, M. Fauvel, S. May, and N. Dobigeon, "Hierarchical Bayesian image analysis: from low-level modeling to robust supervised learning," Submitted.

3.8 HMM Imaging of the Aquatic Ecosystem

Peter Gege (DLR - Oberpfaffenhofen, DE)

Main reference P. Gege (2017): Radiative transfer theory for inland waters. In: Mishra D.R., Ogashawara I., Gitelson A.A. (Eds.), Bio-Optical Modelling and Remote Sensing of Inland Waters. Elsevier, p. 27-69. ISBN: 978-0-12-804644-9

URL http://dx.doi.org/10.1016/B978-0-12-804644-9.00002-1

The oceans are monitored since decades using multispectral sensors on satellite, but very little information is available for the majority of inland waters since they are too small for ocean colour satellites, optically too complex for most multispectral sensors and too numerous (around 120 million lakes $>$ 15 m) for traditional sampling. Since inland waters cover less than 4% of Earth's surface, their importance for global processes has long been overlooked, but new data indicate that they may be more important than the oceans in some aspects, e.g. they bury twice as much carbon from the atmosphere by sedimentation. A number of hyperspectral space sensors will be launched in the next years whose resolution of 30 m is suited to monitor water quality of nearly 90% of the lake areas. I present the principles of the models that are used to analyse hyperspectral data over water and discuss the potential and challenges of hyperspectral imaging for optically complex water types.

3.9 Assessing the Need for Fundamental Biophysical Data

Gladimir V. G. Baranoski

Predictive models of light and matter interactions are employed in a wide range of applications in several fields such as computer graphics, remote sensing and biomedical optics, just to name a few. It is a well-known fact that a well-designed model is of little use without reliable specimen characterization data (e.g., thickness and pigment concentrations) to be used as input, and reliable evaluation data (e.g., spectral reflectance and transmittance) to be used in the assessment of its predictive capabilities. Ideally, the specimen's characterization data to be incorporated into a model should correspond to the specimen used to obtain the measured data employed in its evaluation. However, the few spectral datasets available in the literature rarely provide a comprehensive description of the target specimens. Data is even more scarce for materials in their pure form, such as natural pigments, whose absorption profile is often obtained either through inversion procedures, which may be biased by the inaccuracies of the inverted model, or does not take into account in vivo and in vitro discrepancies. In this talk, we discuss these issues and their practical implications for the development of robust hyperspectral technologies relying on light interaction models.

3.10 Spectral to the People: Towards Affordable and Easy-to-use Spectral Imaging

Jon Y. Hardeberg (Norwegian Univ. of Science & Technology - Gjøvik, NO)

In recent decades there has been a significant volume of research carried out in the field of spectral imaging, that is, imaging systems and methodologies in which the spectral radiance or reflectance of the imaged scene or objects is captured and processed. Such systems have shown their usefulness in many application domains such as cultural heritage, medical imaging, biometrics, remote sensing, food quality, etc.

High spectral accuracy generally comes at a high cost, for instance using so-called push-broom line-scanning hyperspectral imaging technology. On the other hand, multispectral imaging systems with a lower number of spectral channels have been developed, in which typically multiple subsequent image capture operations with a 2D panchromatic image sensor is needed, together with optical filters mounted on a filter wheel or liquid crystal tunable filters. While such systems are generally cheaper, their cost is still prohibitive for many applications. Furthermore both approaches face obvious challenges when applied to real-world non-stationary scenes. And finally users often find the technologies to be lacking in user friendliness, for instance due to the need for complicated calibration procedures and limited software for analysing the data. In summary, key obstacles to broader acceptance of spectral imaging for new applications are cost, user friendliness, and speed.

Recently, new approaches for faster and more practical spectral image acquisition have been proposed, including the three promising ideas of using spectral filter arrays, using two color cameras with additional optical filters in a stereoscopic configuration, and using active LED illumination in conjunction with RGB or panchromatic area image sensors. In this presentation we gave a brief overview our recent research using these three approaches, discuss their advantages and disadvantages, as well as directions for further research aiming for faster, cheaper, and more user friendly solutions for spectral imaging.

3.11 Visualization and Visual Analysis of Hyperspectral, Multispectral, and Multimodal Data

Andreas Kolb (Universität Siegen, DE)

Compared to fully automated techniques for multi- and hyperspectral image analysis, interactive visual analysis approaches allow the ad hoc incorporation of expert knowledge, thus making the exploration of the unknown possible. This, furthermore, gives option to understand the retrieved results in their contexts. However, there are several challenges to any interactive visual analysis approach. In order to prevent the human involvement to be too time consuming, efficient and flexible to use analysis components are needed. Furthermore, the dimensionality of both, the visual feedback as well as the interaction mode is very limited, and the visual analysis results are often only qualitative.

In this presentation, two examples are given, incremental spectral unmixing and error guided endmember selection. Incremental spectral unmixing addresses the need of providing interactive unmixing functionalities, which are able to give direct feedback to the user is he/ she changes the set of endmembers and/or the unmixing conditions. These kind of approaches can predict the unmixing result as user changes the endmember set incrementally, i.e., by adding or removing a single endmember. Error guided endmember selection is a concept in which the user get more sophisticated feedback regarding the quality of unmixing. Common approaches use a pre-defined, scalar error metric, which solely given information about the fitting quality of the spectral reconstruction with respect to the given raw spectrum. The presented approach color-codes the spectral signature of the reconstruction error on a coarse level, thus supporting the identification of regions in the multispectral image that have similar spectral error characteristics.

Future visual analysis might incorporate prominent techniques used for automated classification and detection tasks, such as machine learning and more complex optimization methods.

References

1 B. Labitzke, S. Bayraktar, A. Kolb, "Generic Visual Analysis for Multi- and Hyperspectral Data", In Data Mining and Knowledge Discovery, Special Issue: Intelligent Data Visualization, 2012, pages 117-145

2 B. Labitzke, F. Urrigshardt, A. Kolb, "Expressive Spectral Error Visualization for Enhanced Spectral Unmixing", In Proceedings of International Workshop on Vision, Modeling and Visualization, 2013, pages 9-16

3 B. Labitzke, A. Kolb - Efficient and Accurate Linear Spectral Unmixing In IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2013

3.12 Hyperspectral Data for Terrain Classification

Dietrich Paulus (Universität Koblenz-Landau, DE)

Joint work of Christian Winkens
Main reference Christian Winkens, Volkmar Kobelt, Dietrich Paulus: Robust Features for Snapshot Hyperspectral Terrain-Classification. CAIP (1) 2017: 16-27
URL http://dblp.org/rec/conf/caip/WinkensKP17
Main reference Christian Winkens, Florian Sattler, Dietrich Paulus: Hyperspectral Terrain Classification for Ground Vehicles. VISIGRAPP (5: VISAPP) 2017: 417-424

Hyperspectral snapshot cameras in NIR and VIS can be used to navigate autonomous vehicles. We show how data acquired from these devices can be used to classify terrain into drivable and non-drivable areas. This is of particular importance for unstructured outdoor areas. One specific problem is to identify shadowed areas as they may lead to different features and as they are a source of possible misclassification. We show some approaches to detect shadows and to extend the terrain classification at this point. We provide a database of annotated videos from a multispectral stereo camera for evaluation of experiments.

3.13 Classification of Multimodal Data in Raman- and IR Microspectroscopy

Christoph Pomrehn (Hochschule Bonn-Rhein-Sieg - St. Augustin, DE)

Joint work of Daniel Klein

The field of vibrational microspectroscopy can generally be subdivided into the concepts of Raman- and IR microspectoscopy. Both approaches aim to the identification and localization of specific molecular vibrations in matter and thus, to the identification of the substance. Due to the rule of mutual exclusion, basic molecular vibrations that can be detected by one of these modalities, can not be detected by the other and vice versa. Consequently, a multimodal approach might provide complementary information about the object under test. In this talk, we presented results of a classification study determined by comparing monomodal and multimodal classification rates and class error distributions of data, derived from a polymer sample. It turned out, that a clear majority of feature-classifier constellations show a numerical improvement in classification rates.

3.14 Metrological Hyperspectral Image Analysis and Processing

Noël Richard (University of Poitiers, FR)

Joint work of Hilda Deborah and Audrey Ledoux

Are spectral acquired values measures or just data?

Considering an acquired value as a measure allows to take care about some associated properties, like accuracy, bias and uncertainty. Using it as a simple data induces to consider at the same level of information an inaccurate value and a value with a reduced uncertainty, with the associated consequences to the final decision. All the objectives of the metrological processing or analysis are to preserve the spectral and spatial accuracy of the acquired measures in the computation stages.

To preserve the metrological properties of the measures, adapted mathematical definitions of the spectral acquisitions must be adopted. Consequently, hyperspectral measures must be defined as functions over the wavelengths and not as vector or probability density functions (proof are in the proposed bibliography). Inside this definition, acquired spectra are defined as sampling of a physical/ optical continuous function (radiance, reflectance, irradiance). Consequently, adapted spectral distance/ difference functions must be defined and validated under metrological constraints. A first solution is provided (Kullback-Leibler pseudo-divergence or KLPD) respecting all the expected properties expected from a similarity or distance function. In addition, the KLPD naturally splits the spectral similarity as a sum of a shape and an energy difference. Thanks to this construction, analysis tools based on histograms of differences are defined, allowing to process statistical statistics and models of mixing (GMM).

In the context of multispectral images, the spectrum is considered as being expressed inside a non-orthogonal basis of functions (the spectral sensitivity curves of the channel sensor). Thanks to this definition, the metrological processing tools takes into consideration this inter-dependency using a scalar product of functions. In addition, the Di-Zenzo expression allowing to define a gradient inside an orthogonal basis is extended to the non-orthogonal case using a Gram matrix producing so a generic expression for any case of multi-spectral images. Results are provided for colour and multispectral (Silios sensor, 8 spectral channels + 1 pan-chromatic channel).

We conclude about some challenges that need to be addressed collectively in the following months/ years.

References

1 H. Deborah, N. Richard, and J. Y. Hardeberg. "A Comprehensive Evaluation of Spectral Distance Functions and Metrics for Hyperspectral Image Processing". In IEEE J Sel Topics Appl Earth Observ Remote Sens, 8.6 (2015), p. 3224-3234

2 H. Deborah et al. "Assessment protocols and comparison of ordering relations for spectral image processing". In IEEE J Sel Topics Appl Earth Observ Remote Sens. In press

3 H. Deborah. "Towards Spectral Mathematical Morphology". PhD thesis: Norwegian University of Science & Technology, University of Poitiers, 2016

4 Ledoux et al. "Toward a full-band texture features for spectral images". In Image Processing (ICIP), IEEE Intl Conf. 2014, p. 708-712

5 N. Richard et al. "Full Vector Gradient for Colour and Multivariate Images". In IEEE Transactions on Image Processing". Submitted

6 N. Richard et al. "Pseudo-Divergence and Bidimensional Histogram of Spectral Differences for Hyperspectral Image Processing". In J Imaging Sci Technol 60.5 (2016), p. 1-13

7 N. Richard et al. "Statistical moments for hyperspectral data analysis: Application to remote sensing". Submitted.

3.15 Visual Analysis of Fine Details in Raman Microscopy

Christoph Markus Schikora (Universität Siegen, DE)

Joint work of Plack Markus, Bornemann Rainer, Haring Bolivar Peter, Kolb Andreas
Main reference Christoph M. Schikora, Markus Plack, Rainer Bornemann, Peter Haring Bolivar, Andreas Kolb, "Visual Analysis of Confocal Raman Spectroscopy Data using Cascaded Transfer Function Design", In Computer Graphics Forum, 36(3), 2017
URL http://dx.doi.org/10.1111/cgf.13183

Analysis of fine details like grain boundaries in mono-layer graphene is a directly impossible task in confocal Raman microscopy CRM with usual confocal Raman imaging. Current experiments show the possibility to overcome this limitation caused by the resolution limits, by the high amount of data and by to CRM adapted interactive exploitative visual analysis concepts, in which the usage of spatial features and oversampling or super resolution is the key. This opens not only new applications in quality control of graphene and other carbon polymers, but also generally an approach for imaging of fine details below optical limits in confocal Raman microscopy.

3.16 Spectral Filter Array Cameras

Jean-Baptiste Thomas (Norwegian Univ. of Science & Technology - Gjøvik, NO)

Spectral Filter Arrays camera provides a snapshot imaging technology to acquire multispectral images. Advantage of this technology is that it can be embedded into a very standard imaging pipeline with minor modification. Recent advances and commercial availability rise the question on the uses and actual limitations of this technology. I define the imaging procedure and the pipeline, and illustrate identified issues.

References

1 Lapray, P.-J.; Wang, X.; Thomas, J.-B.; Gouton, P. Multispectral Filter Arrays: Recent Advances and Practical Implementation. Sensors 2014, 14, 21626-21659.

2 Thomas, J.-B.; Lapray, P.-J.; Gouton, P.; Clerc, C. Spectral Characterization of a Prototype SFA Camera for Joint Visible and NIR Acquisition. Sensors 2016, 16, 993.

3 Lapray, P.-J.; Thomas, J.-B.; Gouton, P.; Ruichek, Y. Energy balance in Spectral Filter Array camera design. Journal of the European Optical Society-Rapid Publications 2017, 13:1.

4 Lapray, P.-J.; Thomas, J.-B.; Gouton, P. High Dynamic Range Spectral Imaging Pipeline For Multispectral Filter Array Cameras. Sensors 2017, 17, 1281.

3.17 Spectral Imaging for Fluorescent Objects

Shoji Tominaga (Chiba University, JP)

Main reference S. Tominaga, K. Hirai, and T. Horiuchi, Estimation of bispectral Donaldson matrices of fluorescent objects by using two illuminant projections, Journal of the Optical Society of America A, Vol. 32, No. 6, pp.1068-1078, 2015.
Main reference S. Tominaga, K. Kato, K. Hirai, and T. Horiuchi, Spectral image analysis of mutual illumination between florescent objects, Journal of the Optical Society of America A, Vol. 33, No. 8, pp.1476-1487, 2016.
URL http://www.ntnu.edu/employees/shojit

I gave a talk about Spectral Imaging for Fluorescent Objects. Use of fluorescent materials has increased in our daily lives. All sorts of objects we see in each day often include florescence. First, I presented the visual effects of fluorescence. In fact, because of fluorescent emission, many fluorescent surfaces appear brighter and more vivid than the original object color surface. Next, I described the principle of fluorescence. The fluorescent characteristics are well described in terms of the bispectral radiance factor, which can be summarized as a Donaldson matrix. The 2D matrix is an illuminant independent representation of the bi-spectral radiance factor. I introduced several measurement methods of the bi-spectral radiance factor. A two-illuminant projection method is useful in an ordinary scene using spectral imaging system. Finally, I showed the spectral imaging application to the appearance reconstruction problem.

3.18 Exploiting Human Eyes for Remote Sensing

Devis Tuia (Wageningen University Research WUR, NL)

Joint work of D. Tuia, M. Volpi, B. Kellenberger, N. Rey, S. Joost
Main reference Rey, N.; Volpi, M.; Joost, S. & Tuia, D. "Detecting animals in African Savanna with UAVs and the crowds," Remote Sens. Environ., 2017, 200C, 341-351
Main reference Kellenberger, B.; Volpi, M. & Tuia, D.. Fast animal detection in UAV images using CNNs IEEE International Geoscience and Remote Sensing Symposium, IGARSS, 2017
URL https://doi.org/10.1016/j.rse.2017.08.026

Managing wildlife reserves is a complex task: rangers are confronted to poaching, livestock control and grazing needs estimation and often can use only manual counting requiring costly overflight or inaccurate land counts, followed by some extrapolation. In this talk, I present how we approached semi-automatic animals detection with computer vision technologies and images acquired by Unmanned Aerial Vehicles (UAV). I will discuss problems related to i) the acquisition of annotations using a crowd of volunteers; ii) the development of a wildlife detection system based on such annotation and a deep neural network and iii) the improvement of the databased using active learning approaches, also known as human-in-the-loop systems. The project is part of the SAVMAP initiative: http://lasig.epfl.ch/savmap.

3.19 Squeezing Spectra: Hot Topics at the Color Imaging Lab of the University of Granada

Eva M. Valero Benito (University of Granada, ES)

Joint work of Miguel A. Martínez Domingo, Sergi Etxerebere, Juan L. Nieves, Javier Hernández Andrés, Javier Romero, N. Tello Burgos, A. López Montes
Main reference Miguel ángel Martínez-Domingo, Eva M. Valero, Javier Hernández-Andrés, Shoji Tominaga, Takahiko Horiuchi, and Keita Hirai. Image processing pipeline for segmentation and material classification based on multispectral high dynamic range polarimetric images. Opt. Express 25(24), 30073-30090 (2017)
Main reference N. Tello Burgos, A. López Montes, X. Shao, E.M. Valero, J.L. Nieves, M.R. Blanc. "Identification of the evolution of indigo blue by multispectral imaging". Technoheritage 2017, 21-24 May, Cadiz (Spain). Abstract book, pp. 72.

Some of the topics that we have been working on at the Color Imaging Lab of the University of Granada are briefly introduced: detection of aging phase of indigo samples for an art-preservation related application; complete image pipeline for spectral High-Dynamic-Range Polarimetric images, including segmentation and material classification; and the improvement of saliency detection models when spectral features are used as input. These topics give rise to open questions related to bridging the gap between the art-preservation experts and ourselves, the use of spectral imaging as a seed bank for features that allow for simplification of the capture devices, and the possibility of using visual attention to improve the detection of regions of interest in the images to be further processed.

3.20 Optical Imaging Techniques for Non-contact Measurements of Vital Functions and Diagnosis of Tissues in Medicine

Rudolf Verdaasdonk (VU Medical Center - Amsterdam, NL), John Klaessens, and Herke Jan Noordmans

Joint work of VU University Medical Center, Medical Center Utrecht, Norwegian University of Science and Technology

In the recent years, CCD and CMOS camera technologies in the visible and near IR has opened new methods of diagnostics in medicine. For the mid-IR, thermal cameras have become small and practical with high spatial and temperature resolutions. With the introduction of practical 3D scanners, medical images have become quantifiable. New clinical applications were investigated, imaging dynamic changes in tissue perfusion, oxygenation and physiological processes to differentiate between healthy and abnormal tissues using combinations of narrow band spectral images. Near IR cameras were used to measure the heart and respiration rate in patients independent of skin tone and in dark conditions within 3% accuracy. Blood vessel puncture procedures were significantly improved visualizing the vessel structures on a screen like car navigation system. IR thermal imaging was applied successfully in cardiology (predict arterial spasm), urology (cause of impotence), anesthesiology (anesthetic block and pain treatment), aesthetic surgery (transplantation, burn wounds) and dermatology (allergic reactions) some in combination with 3D scanners. None-contact imaging techniques proved to be successful as new diagnostics tool that can easily be introduced in the clinic with minimal risk for the patient with great potential for general practitioners and even at home.

3.21 Introduction to THz Imaging and Spectroscopy

Anna Katharina Wigger (Universität Siegen, DE)

Joint work of Gunnar Spickermann, Matthias Kahl, Christian Weisenstein, Daniel Stock, Peter Haring-Bolivar

Technology in the THz range is manifold as it is adapted from the neighboring domains in the electromagnetic spectrum. Many THz imaging systems have very low bandwidth and are to date not capable of materials recognition. THz spectroscopy is a very broadband technique to analyze material properties as the complex refractive index. The vision is a 3D THz imaging system, that covers multiple bands in real-time in order to recognize not only objects, but also can recognize materials directly.

3.22 Toward Practical Applications of High-Resolution Multispectral/ Hyperspectral Video

Masahiro Yamaguchi (Tokyo Inst. of Technology - Yokohama, JP)

Joint work of Yuri Murakami
URL http://www-oid.ip.titech.ac.jp/

What are the issues toward widespread use of multispectral and hyperspectral imaging? Multispectral and hyperspectral imaging technology has been investigated in remote sensing, color imaging, and machine vision almost independently up to know and recently computer vision and machine learning field is also approaching to multispectral and hyperspectral technology. Gathering knowledge in different application fields is quite beneficial for promoting further commodification of the technology.

Firstly, it is important to explore the advantages of spectral imaging in a variety of fields, and we have demonstrated experimental results in color reproduction, medical image analysis, and human detection from airborne observation. Although the advantages of spectral imaging have been verified in various fields, there still exist serious issues that should be solved for practical use of spectral imaging. Since objects are moving in many cases, single-shot or video spectral imaging is crucial. However, it is still difficult to implement high-resolution spectral imaging with single-shot or video. As a solution to such issue, the approach of hybrid-resolution spectral imaging is illustrated. Finally, the significance of standardization-related activity for promoting practical applications is discussed. CIE (International Commission on Illumination) recently published a technical report "multispectral image format," which has been prepared by TC8-07. CIE also establishes a new research forum "spectral imaging" and work items for promoting spectral imaging technology are now being discussed.

References

1 M. Yamaguchi, H. Haneishi, and N. Ohyama, "Beyond Red-Green-Blue(RGB): Spectrum-Based Color Imaging Technology," Journal of Imaging Science and Technology, 52(1): 010201-1-15, 2008

2 N. Hashimoto, Y. Murakami, et al, "Multispectral image enhancement for effective visualization," Optics Express, 19(10): 9315–9329, 2011

3 L. Yan, M. Yamaguchi, et al, "Using hyperspectral image enhancement method for small size object detection on the sea surface," Proc. SPIE, v. 9643, 9 pages, 2015

4 Y. Murakami, M. Yamaguchi, and N. Ohyama, "Piecewise Wiener estimation for reconstruction of spectral reflectance image by multipoint spectral measurements," Applied Optics, 48(11): 2188-2202, 2009

5 Y. Murakami, K. Nakazaki, and M. Yamaguchi, "Hybrid-resolution spectral video system using low-resolution spectral sensor," Optics Express, 22(17): 20311-20325, 2014

6 CIE 223:2017 Multispectral Image Formats, International Commission of Illumination, 2017.

3.23 Deep Learning in Remote Sensing

Xiaoxiang Zhu (DLR Oberpfaffenhofen & TU München)

Joint work of Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, Friedrich Fraundorfer
Main reference Zhu X., Tuia D., Mou L., Xia G., Zhang L., Xu F., Fraundorfer F. 2017, Deep Learning in Remote Sensing: An Review, IEEE Geoscience and Remote Sensing Magazine, in press
URL https://arxiv.org/abs/1710.03959

In this talk, I intended to answer three questions:

1. What makes deep learning special in remote sensing?
   Keywords: five dimensional data, multi-modal data, big data, physical models etc.
2. Where we are today?
   Showcasing classification, change detection, data fusion, time series data analysis, and geo-info extraction from social media data
3. What are the open issues?
   Keywords: novel applications, transferability, very limited annotated data, benchmark, and combing deep nets with domain expertise.

References

1 Mou L., Ghamisi P., Zhu X., 2017, Unsupervised Spectral-Spatial Feature Learning via Deep Residual Conv-Deconv Network for Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing, in press

2 Mou L., Ghamisi P., Zhu X., 2017. Deep Recurrent Neural Networks for Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing, 55(7), 3639-3655

3 Kang J., Wang Y., Körner M., Taubenböck H., Zhu X. (2017), Building Instance Classification Using Street View Images, ISPRS Journal of Photogrammetry and Remote Sensing, accepted subject to minor revision

4 Hu J., Mou L., Schmitt A., Zhu X. (2017), FusioNet: A Two-Stream Convolutional Neural Network for Urban Scene Classification using PolSAR and Hyperspectral Data, Proceedings of 2017 Urban Remote Sensing Joint event (JURSE), Dubai, United Arab Emirates.

4 Working groups

4.1 Hyperspectral, Multispectral and Multimodal Image Acquisition

Donald G. Dansereau (Stanford University, US)

Joint work of 10 attendees

Attendees participated in a workshop on hyperspectral image acquisition on Tuesday 10 Oct, 2017. The workshop was structured in two stages: first, a brainstorm establishing major topics, then a detailed discussion touching on the most important points. The following is a summary of the major points touched on during the discussion, followed by concrete recommendations.

Manufacturers

A major challenge in effective acquisition is working with camera manufacturers. There are two key hurdles: first, communication from the manufacturers is often limited. They do not always clearly communicate the characteristics and limitations of their products. They also do not provide clear road maps outlining their intended future product developments, or a schedule for the release of calibration data and fixes to existing products.

The second issue comes in how we as a community can clearly communicate our needs to manufacturers. Ideally, we would dictate custom bands and resolutions on a per-camera basis, allowing applications to drive specific camera developments. Alternatively, it would be desirable to specify the needs of an application in terms of discriminative power, and allow the manufacturer to address these needs.

However, custom camera design is expensive and manufacturers are unlikely to address every request. The group concluded that there is an important market gap that would be addressed by an easily customised multispectral camera. We acknowledged that Pixeltek offer custom, possibly costly colour filter arrays (CFAs), and that colour filter wheels can go some way to addressing this challenge, but there seems to be a remaining market gap in providing affordable custom snapshot multispectral cameras.

Calibration

Calibration was identified as a major common concern. This includes radiometric, geometric, and spectral characterisation. There are open questions as to how good a calibration needs to be, and indeed, what the limits are on how good a calibration can be. There are some applications, e.g. satellite and retinal imaging, for which complete calibration seems impossible as it would require direct knowledge of the medium (air, eye lens and vitreous).

In some instances blind calibration may be possible based on prior knowledge of the manifold of physically viable media. The motivating example arose of obtaining a calibrated colour measurement of the back of the eye. There is an unknown spectral impact from the eye's lens, which is variable between individuals and yellows as we age. Suggestions arose around measuring scattered light as a hint of the lens' impact, and characterising the manifold of spectral characteristics typical of human eyes. This could be driven by physically based modelling given knowledge of the source of yellowing in the eye's optics. Then the inverse problem of separating the lens colour from the colour of the back of the eye can be carried out as an optimisation subject to the constraint of physically feasible lens and eye colours. The possibility was also raised that structured light or other coded illumination may be able to disambiguate the colour of the lens from that the back of the eye.

The group discussed whether calibration procedures are sufficiently well defined and universally understood. It was generally held that that spectral calibration is fairly well defined and understood, if tedious, but that radiometric (gain) calibration is not as well defined, and often overlooked. This kind of calibration is time-consuming and not every lab has the capability. Sharing of calibration capability is difficult, expensive, and time-consuming, taking up to a week to calibrate a single sensor.

One concrete recommendation is that calibrations for commercial products be openly shared. Present datasets are spread out, and there is a call to collect these and provide a centralised location for the community to share calibration information and procedures.

Novel Optical Setups

We discussed at some length a set of emerging technologies in hyper- and multi-spectral capture. One set of techniques split the camera's field of view into multiple sub-images, employing filters to turn each sub-image into a separate colour band. This is essentially a light field camera with per-subaperture colour bands, and it might benefit from combining concepts from light field imaging and hyperspectral imaging.

Further ideas on the camera side included coded aperture snapshot spectral imaging (CASSI), custom CFAs including masks that combine colour and polarisation, a NASA-developed holographic multispectral camera, tuneable filters, and microprism arrays for on the order of 100 colour bands with low spatial resolution (order 100×100 pixels).

Further ideas concerned the use of novel lighting arrangements, employing multiple LEDs with diffusers. A concern with these configurations is obtaining sensitivity outside the visual spectrum, and getting IR and other bands to cooperate on a camera with no IR cut filter. Tuneable lighting came up as another option, with the caveat that calibration can be difficult.

We discussed the chicken-and-egg problem of camera design and applications: it is difficult to motivate fabrication of custom cameras before having a strong idea of their application, and conversely it is difficult to demonstrate applications of novel cameras without building them. The question of optimal sensor design came up, and discussion touched briefly on the use of information-based metrics vs. deep learning to address the complex statistics of visual perception.

Lossless Low-Level Processing

Conversation touched on the line lies between raw data, pre-processing, and processing. Often in conventional imaging the impact of demosaicing is neglected, and not considered part of the processing at all. However, without storing additional information this step is destructive, even in the simple case of an RGB camera. In the multi- and hyper-spectral cases things are more complex and the potential for loss of information between steps is strong. There was a consensus that there is a need for more standardised interfacing between levels of processing, and that there is a need for stronger intermediary image representations and file formats. In particular, images should include whatever information is available on the CFA bands, noise levels, exposure, channel alignment, radiometric calibration, and ultimately covariance of demosaiced and processed colour. Ideally, one should be able to reconstruct the raw imagery using the pre-processed imagery and accompanying metadata. We presently lack a universal format with which to allow this level of lossless processing.

Recommendations

The discussion yielded a set of concrete recommendations. For calibration data, there is a need to collect existing calibrations into one place, and to work towards standardising calibration procedures to increase the value of shared calibrations. When dealing with manufacturers be specific and aggressive when asking for calibration information. Manufacturers should provide (and be encouraged to provide) roadmaps of what they plan to calibrate and fix in future. A survey paper covering existing calibration methodologies would be well received by the community.

There is a market gap for an easily customized multispectral camera, allowing application-specific cameras to be quickly customised, evaluated, and deployed.

There is a need for a common file format including metadata at the interface between pre-processing and processing. This should provide, for example, covariance of demosaiced colour, CFA bands, noise levels, exposure, and any available calibration information. There is ideally sufficient information in any pre-processed image to infer the RAW imagery that yielded it.

While presenting an overview of this discussion to the larger group, we identified as an action item follow-up discussion on the potential of CIE file formats for hyperspectral imagery.

4.2 Deep Learning for Analysis of Multispectral Data

Dietrich Paulus (Universität Koblenz-Landau, DE) and Richard Bamler (DLR - Oberpfaffenhofen, DE)

Main reference J. Johnson, A. Karpathy, and L. Fei Fei. "Densecap: Fully convolutional localization networks for dense captioning." In CVPR16, pages 4565–4574, 2016

Main reference R. Kemker and C. Kanan. "Deep neural networks for semantic segmentation of multispectral remote sensing imagery." In CoRR, abs/1703.06452, 2017

Main reference W. Shen, X.G. Wang, et al. "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection." In CVPR15.

The discussion mainly focused on deep learning (DL) for remote sensing.[1]

Other aspects, such as other machine learning (ML) methods and other applications for multispectral image analysis (such as multispectral cameras on vehicles) have been only partially addressed.

General Statements on Deep Learning

Many papers are published in remote sensing journals making use of deep learning - but not really doing research in deep learning.[2] The reason might be, that ML people were often not interested in applications. Researchers in remote sensing had to do research combining ML methods and the application.

In general, there seems not to be enough labeled or ground-truth data – at least in remote sensing – to train large neural networks. This is a big difference to computer vision tasks, where large sets of annotated images exist.[3]

In the following are statements pointed out during the working group session:

1. DL has proven to be superior to other ML methods in detection and classification (labeling).
2. It is not clear yet how DL performs in quantitative parameter estimation from spectral measurement data.
3. The understanding of what really happens in a deep neural network (DNN) is underdeveloped. A lot of progress is achieved by "trying out". If trying replaces understanding, is this still in agreement with our understanding of science?
4. DL can potentially help for building up efficient dictionaries, e.g., for sparse reconstruction.
5. We do not know at the moment how prior knowledge can be incorporated into DL in a systematic way.

[1] A survey on these methods was also given in the presentation of X. Zhu.

[2] Of course, studying the structure and performance of deep neural nets can also be considered research in this field.

[3] This has also been claimed in https://arxiv.org/abs/1703.06452, see also (Kemker and Kanan).

6. It is difficult to obtain a quality measure (error bar, covariance matrix, etc.) of results obtained by DL.
7. To the knowledge of the group DL is currently mostly restricted to "image-type data in – image-type data out". One exception are applications, that generate figure captions from images, see (Johnson, et al). This is an example, where structured symbolic data is computed from an input image.

Nevertheless, in the following are several questions that were raised and remained without answers:

1. Are there e.g. successful examples for vectorization of input images by DL?[4]
2. Do we have to throw away human expertise when we use DL?

Big Problems Solved by Machine Learning

The following areas have been identified where solutions by ML (in the context of analysis of multispectral data) have been published: Detection, classification, super-resolution, enhancement, fusion, change detection, and image restoration. Whether or not these areas can be regarded as "solved" needs to be discussed.

What has not been possible up to now with ML

Find something new in visualization.

4.3 User Involvement, Validation Tools, Data, and Software Sharing

Rudolf Verdaasdonk (VU Medical Center - Amsterdam, NL)

Joint work of 10 attendees

User involvement

The workshop attendees represent mostly research institutes from the HMM field. However, it is important to identify the end users of HMM technology and get them involved in the development of hardware and application software. The following are the fields or end users of HMM technology identified during the working group session: Physician in the medical field, heritage of artifacts and art, military, ecologists, earth observation: drones and satellites, metrology, garbage recycling, food industry, cosmetics, security, autonomous driving, agriculture, and emergency services.

The end users are usually less interested in the HMM technology but just the results it can provide for their needs/ goal. The results/ data should be presented to the user in an intuitive way with similarity to regular presentations in their workflow. The success of acceptance of new HMM technology will depend on the user friendliness of the systems and presentation of results. The needs of end users of HMM technology themselves are, e.g., quantification, segmentation of materials, tissues, and structures, specifications of materials, and detection.

Methods of presenting HMM data to the user:

[4] For computer vision problems, results have been published in (Shen, et al).

- Data fusion in existing visualization methods
- Presenting data in color palettes which are intuitive for the user
- Blinking or blending
- Superposing/ mixing on visual/ original image with adjustable transparency (see example at iipimage.sourceforge.net)
- In contrast with background
- Selection of preset filters like highest and lowest values

Important aspects regarding the acceptance of new HMM technology:

- Communication with user is important
- Cost effective
- Fit in the normal workflow
- System should be mobile and flexible

Regarding the opportunities to open consumers market for HMM technology, if there would be consumer driven need for HMM technology, sensors would be made in mass production and the price would drop enormously, e.g., HMM sensor in a smart phone. In the following are several ideas for potential applications:

- Detection of mushrooms eatable or poisonous
- Food/ meat/ fruit freshness detection in supermarket
- Presences of potential toxic substances on food
- Cosmetics: Color of foundation and make-up
- Color of cloths under different lighting conditions
- Dentist: color of restorations or crowns, check in other light conditions
- Check on health of pets/ animal
- Combinations with thermal camera
- Combination with time of flight sensor (from automotive industry)
- Hair salons: Color prediction before hair dye
- Recognition of materials
- Characteristics of light sources at home in relation to the perception of paint on the walls
- Automotive industry

As a remark, consumer market is different from the needs for accurate measurements in scientific community.

Sharing HMM data sets and software

Within the HMM research community, there is a need for benchmark data/ images for validation and testing software algorithms. It is important to set the rules by the leading people how the dataset can be used and the proper way to refer to data set in publications. Especially, the users need to pay attention to the ground truth when they analyze the data and always make a comparison. The data should be made available through a website. Which organization/ institute could host a website for this? In the following are several of the considerations:

- One organization/ person should be made in charge/ responsible for website. CIE could potentially take the initiative but does not have resources.
- As other options: To join with websites already hosting other datasets, create own website with a wiki page like layout to allow relatively easier maintenance. Another option is also to have a website with links/ torrent to source of original data
- Regarding the data sharing, the ones that are easily available will be small samples while the larger while could be downloaded later

- Example website: Data fusion contest (IEEE) with open data sets available, or the color constancy website

Phantom/reference for testing and validation

There is a need for a test target/reference/phantom to be able to compare and/or validate data from various HMM systems like a 3D physical 'color checker' with various properties like translucency, texture, angle of illumination etc. Example Round Robin Test: reference objects (color checker, paintings, etc.) for MSI where distributed along universities over the world for testing there MSI setup. The result showed that there was a large range in outcome in spectra, not reproducible without a protocol describing the conditions to perform the test. There is much to improve on this topic.

Synthesis and Analysis

Regarding the software or standard algorithms, there is a need for a software platform with algorithms for sharing, comparable to the computer graphics field where standards have been implemented. As potential software platforms are 'Vision' and 'PKLF'. The latter have been proposed as a standard in the past.

Besides sharing the software and algorithms, it is important that the data file should be readable by the software. Thus, the need for standard file format for HMM data. Points discussed during the session were: Which institution should be responsible for the task and which software platform. Matlab is generally not preferred because of non-compatible versions and as it is commercial. Python could be chosen as an alternative.

A website could be the solution, combined with data sharing as discussed in the previous sections. But then again, there is still the question of which institution could host such a website. Although one from the computer science community is more preferable. A remaining discussion point was related to the role/ task the stakeholders such as manufacturers should take.

Finally, aside from all the tasks and challenges that have been identified above, there are already many of the success stories. Moreover, what can be learned from those areas/ users where HMM technology is already accepted or proven successful are, i.e., it is important to be close with the users, to understand their environment and talk with the same language. Such success stories can be found in the areas of dentistry, automotive industry, several medical applications, checker for the health of seeds, earth remote sensing, security such as in airports (milli-/ terahertz), and biometrics.

Participants

- Gonzalo R. Arce
University of Delaware, US
- Richard Bamler
DLR – Oberpfaffenhofen, DE
- Gladimir V. G. Baranoski
University of Waterloo, CA
- Shida Beigpour
MPI für Informatik –
Saarbrücken, DE
- Ela Claridge
University of Birmingham, GB
- Donald G. Dansereau
Stanford University, US
- Hilda Deborah
Norwegian Univ. of Science &
Technology – Gjøvik, NO
- Nicolas Dobigeon
University of Toulouse, FR
- Paul D. Gader
University of Florida –
Gainesville, US
- Peter Gege
DLR – Oberpfaffenhofen, DE
- Sony George
Norwegian Univ. of Science &
Technology – Gjøvik, NO
- Jon Yngve Hardeberg
Norwegian Univ. of Science &
Technology – Gjøvik, NO
- Peter Haring Bolivar
Universität Siegen, DE
- Andreas Kolb
Universität Siegen, DE
- Dietrich Paulus
Universität Koblenz-Landau, DE
- Christoph Pomrehn
Hochschule Bonn-Rhein-Sieg –
St. Augustin, DE
- Noël Richard
University of Poitiers, FR
- Antonio Robles-Kelly
CSIRO – Canberra, AU
- Christoph Markus Schikora
Universität Siegen, DE
- Jean-Baptiste Thomas
Norwegian Univ. of Science &
Technology – Gjøvik, NO
- Shoji Tominaga
Chiba University, JP
- Devis Tuia
Wageningen University Research
WUR, NL
- Eva M. Valero Benito
University of Granada, ES
- Rudolf Verdaasdonk
VU Medical Center –
Amsterdam, NL
- Anna Katharina Wigger
Universität Siegen, DE
- Masahiro Yamaguchi
Tokyo Inst. of Technology –
Yokohama, JP
- Naoto Yokoya
DLR – Oberpfaffenhofen, DE
- Xiaoxiang Zhu
DLR Oberpfaffenhofen &
TU München

Report from Dagstuhl Seminar 17412

Internet of People

Edited by

Elizabeth M. Belding[1], Jörg Ott[2], Andrea Passarella[3], and Peter Reichl[4]

1 **University of California – Santa Barbara, US,** ebelding@cs.ucsb.edu
2 **TU München, DE,** jo@in.tum.de
3 **CNR – Pisa, IT,** a.passarella@iit.cnr.it
4 **Universität Wien, AT,** peter.reichl@univie.ac.at

Abstract

This report provides a summary of the organization, program, and outcome of the Dagstuhl Seminar titled "Internet of People". We first provide the main motivations for organising the seminar. Then, we briefly describe the organisation goals of the seminar, and summarise the rationale for the set of researchers involved. We then report on the activities carried out during the sessions, consisting of talks and group works. Specifically, we provide the abstracts of the talks and extended reports on the output of the groups work. Finally, we draw the main conclusions of the seminar.

Seminar October 8 – 11 , 2017 – http://www.dagstuhl.de/17412
1998 ACM Subject Classification C.2.1 Network Architecture and Design, H.1.2 User/Machine Systems, H.3 Information Storage and Retrieval, J.4 Social and Behavioral Sciences
Keywords and phrases Internet design; Next Generation Internet; human-centric Internet; social-aware Internet
Digital Object Identifier 10.4230/DagRep.7.10.42

1 Executive Summary

Elizabeth M. Belding
Jörg Ott
Andrea Passarella
Peter Reichl

The key objective of the seminar was to bring together a diverse group of researchers and practitioners to reflect on technological and social issues related to the emerging concept of Internet of People (IoP). The group of attendees was composed of 28 people with diverse expertise on the various areas of Internet, coming from Europe, US, Asia and Australia.

The group worked for two and a half days, and the work was organised on (i) seed talks, (ii) snippet talks on selected research topics related to IoP, and (iii) parallel group work. The group sessions were particularly productive, and attendees worked on many topics. Specifically, they covered the following topics: (i) IoP definition, (ii) IoP use cases, (iii) IoP and people; (iv) Privacy, security and trust; (v) IoP architecture, and (vi) transition towards IoP. Over the last day, the group again split in three sub-groups, to focus on conclusions and follow-up activities. Specifically, the three groups produced (i) guidelines for IoP toolkits, (ii) a possible IoP research agenda, and (iii) an IoP manifesto.

Internet of People, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 42–68
Editors: Elizabeth M. Belding, Jörg Ott, Andrea Passarella, and Peter Reichl
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

We managed to bring together a quite balanced group of 32 people with expertise in the design and implementation of wireless ad hoc networks of various types, human-computer interaction, community informatics, urban interaction design, ethnography, media studies, arts and design.

The main take-home message from the seminar is that IoP is an emerging research topic with a lot of potential. It spans many aspects, including but not limited to the set of topics identified for the group work. Each of the group works provided concrete guidelines on the selected topics, possibly providing focused research agenda for the future.

Most of all, we are very happy that attendees greatly enjoyed the seminar, including those attending for the first time a Dagstuhl event (about one third). We do believe that the seminar laid the grounds for future fruitful collaborations, and helped a lot in shaping the key ideas of the emerging research topic of IoP.

2 Table of Contents

3 Background and motivation

The diffusion of personal (mobile) devices and pervasive communication technologies is expected to exponentially increase in the next few years (for example, Cisco foresees an eightfold increase of mobile data traffic between 2016 and 2020, with a compound annual growth rate (CAGR) of 53% [1]). This is pushing more and more the Cyber-Physical Convergence vision, discussed, among others, in [2]. According to this vision, the physical world of the users and the cyber world of Internet applications and services are more and more integrated and converging. Data generated in the physical world (e.g., by sensors embedded in personal users' devices and physical infrastructures) flows to the cyber world, where it is elaborated and exchanged. On the other hand, interactions in the cyber world result in actions in the physical world (e.g., because users modify their behaviour based on information received through Internet applications, or because physical infrastructures are configured through actuators).

One of the key effects of this convergence is that humans are more and more at the centre of the technical systems they use. Humans and the cyber systems through which they communicate become actors of a complex socio-technical ecosystem, and designing effective communication systems needs to take into consideration human behaviours as a structural paradigm, rather than as an afterthought. Moreover, in this view humans are not anymore passive objects of Internet technologies, but they play an active role in the design and even operation in the network, to the point of becoming one of the components of the complex Internet socio-technical system – crowdsourcing being a very primitive example of this new perspective. In [3], this paradigm change is named the "Anti-Copernican Revolution", as it puts (back) the human at the centre of the stage in the design and evaluation of Internet communication systems.

According to this communication ecosystem view, we see future research on Internet-based communication systems as a truly inter-disciplinary field, shaped by at least five main interacting dimensions and linking the technological perspective closely to the social, economic and cognitive sciences (describing the behaviour of humans) for designing the communication and data exchange mechanisms of future communication systems.

1. **ICT** provides the basic enabling solutions for communication to occur. However, the algorithms and protocols for communication and data exchange are not driven exclusively by the need to optimise network resource usage (as in the design of legacy Internet systems). In the converged cyber-physical environment, user devices become proxies of their users in the cyber world: They communicate, exchange and manage data by "emulating" the way their human users would do if interacting with each other in the physical world.
2. **Social sciences** model the way users establish social relationships, how they trust each other, and how they are prepared to share resources with each other. Communication systems exploiting these models ("social-aware networking protocols") have proved to be very efficient in supporting communication in human-centred mobile networks [4, 5].
3. **Cognitive psychology** describes, among others, how human beings perceive and interact with data, how they assess relevance of information, how they exchange it when interacting, and how they extract knowledge out of it. Data-centric communication systems for mobile networks have already been proposed, where these models are exploited to efficiently guide information diffusion among users [6].
4. **Micro-economics** is modelling how humans negotiate the use of infrastructure and content resources, trade and share them. This is also fundamental knowledge to predict

how they can interact with each other through communication systems, how they are prepared to share material and intellectual resources in a complex socio-technical system, and to embed such knowledge in the systems' design and operation.

5. Finally, very useful models have been derived in the area of **complex network analysis**, describing, for example, human social relationships with compact graph descriptions, amenable to characterise human behavioural properties and exploit them in the design of networking solutions.

We stress the fact that the proposed human-centric approach to the design of Internet communication systems is not yet another bio-inspired networking design wave. Because of the fact that user devices act as proxies of their users, embedding efficient models of human behaviour in the core design of communication systems is a natural way to make devices behave as their human users would do if faced with the same choices and decisions. Moreover, this approach is not confined to designing human-centred applications. The inter-disciplinary approach impacts all conventional layers of the communication stack above the enabling communication technologies, and brings advantage at all layers, as shown by the mentioned examples.

This approach can be the basis for a seamless communication ecosystem for Cyber-Physical Convergence, where communication entities can be humans, their personal devices, as well as other "machines" communicating in the cyber world. Specifically, we can foresee at least three classes of communication paradigms:

1. **The "Human proxy" paradigm.** This is based primarily on communications between devices, whereby user personal devices communicate with each other acting as proxies of their human users.
2. **The "Crowdsourcing" paradigm.** This is based both on device-to-device as well as on device-to-human and human-to-human interactions. The human user is seen as another entity of the communication ecosystem, and its behaviour can be modelled and predicted (clearly, up to a certain extent), and the resources it brings exploited to optimise the operations of the system (think, for example, of crowdsourcing systems, where humans are used to solve complex problems in a synergic way together with computers).
3. **The "User experience" paradigm [7].** This is based on interactions between users and devices, and the behaviour of the devices is designed taking into consideration the reactions of the human users to the service offered by the communication system, and the resulting quality of the users' experience.

In this view, another cornerstone for the design of Cyber-Physical Converging communication systems is Quality of Experience. Quality of Experience models interactions between humans and ICT services through a human-centric approach, by taking into consideration human expectations on the quality to be obtained, and reactions to varying level of quality. QoE models can thus be fruitfully integrated in the communication systems design, for example to anticipate the effect of devices behaviour on the human users, or to understand how users could react and behave when exposed to certain tasks to be carried out in collaboration with devices.

For further information about the concept of "Internet of People", we also refer to [8, 9].

4 Organization

The main goal in organsing the seminar was to bring together a diverse set of people with expertise relevant for the Internet of People concept. Specifically, we wanted to involve researchers with complementary backgrounds in the various areas that touch upon IoP, such as:

- Internet architectures
- Mobile networking
- Self-organising networking
- Internet standardisation
- Quality of Experience
- Community Networks and Engagement
- Internet for Development
- Internet Application and Service design
- Internet Governance

This was required, as one of the goals of the seminar was to elaborate the main IoP concept, and exploit the seminar as a seminal event to spread knowledge about this new research area. Therefore, we needed to involve relevant researchers in the various communities possibly interested in the IoP concept. In addition, geographical diversity was also sought, trying to bring to the seminar a good mix of people from Europe, US, Asia and Australia.

The initial set of invitees was shaped based on these guidelines. Also thanks to the reputation of the Dagstuhl seminars, we have been very happy to receive a significantly positive feedback from the invitees. Although some could not attend due to clashing commitments, many of the invitee were able to join. Specifically (besides Europe), we had a significant participation from the US, two researchers from Asia, and one from Australia. It is worth mentioning that we also invited the Next-Generation Unit of the European Commission to join the seminar, as we thought that IoP is very much aligned with the spirit of this new H2020 initiative. We have been very happy to receive a very positive feedback from the Unit, confirmed by the participation of its Acting Head.

All in all, 28 researchers attended the seminar. About one third were newcomers in Dagstuhl. It is worth mentioning that, in the survey after the seminar, all respondents stated that they would come back to another Dagstuhl seminar in the future.

5 The seminar

5.1 Breaking the ice: Initial session

As usual, we started the seminar with a round table introduction of all participants, who had been informed beforehand to prepare a 2-slide presentation stating who they are, what are their main research activities, and what they expected from the seminar. The initial round table was a very nice way to break ice and starting to getting to know each other better. While a good share of attendees were already known to each other, some of them were not. We anticipate that they have been productively engaged into the seminar activities, nevertheless.

After the initial roundtable, the organisers delivered a short presentation, stating their view on IoP before the beginning of the seminar, which motivated them to organise it. Specifically, the presentation started by noticing a few facts relate to the current evolution of the Internet. The first one is the emergence of cyber-physical convergence, whereby there is a tighter and tighter correlation and interplay between what happens in the physical and in the cyber world. The second fact is the expansion of the Internet primarily at the edges, much more than in the core, due to the pervasive diffusion of mobile and IoT devices, i.e., devices with a tight link with human users. The third fact is that in this trend, users' devices become more and more proxies of their human users in the cyber world. These facts potentially have a disruptive impact on the Internet as we now it today, such that it may not be possible anymore to think at the Internet according to "business as usual" innovations, but we might be in need of radical rethinking of all the main Internet primitives. In this view, we need to rethink those primitives taking a human-centric approach, i.e., considering the human behaviour as one of the key design concepts of the new Internet. This human-centric perspective is the key concept behing IoP. Finally, the presentation also made the point that Internet research is not only on the ecosystem around the Internet, as the latter is not an immutable technology given for granted now and for all. Rather, IoP calls for radical new research also in the Internet technologies, which are the key technological underpinning of any technological and societal impact related to the Internet.

The initial presentation already stimulated very lively debate and discussions. Among the many others, Max Ott provided quite a strong feedback about the fact that we need to consider the impact of 5G technologies, which are bound to provide a lot of bandwidth and capacity at the edge. Rather, we need to look at the information side of the network, and consider IoP mostly as a information-centric network. While there was not unanimous consensus on the fact that 5G might solve all networking issues in the mid- long-term, all attendees agreed that IoP would be primarily an information-centric network, and this is a correct perspective to use to look at it. Moreover Jörg Ott proposed a more top-down approach, whereby we should (i) think to the services first, which are human-centric, and (ii) then go down and think to the network that one needs, and whether this is local or global.

All in all this initial session proved to be extremely helpful in breaking the ice, start putting onto the table many key concepts related to IoP, and start identifying possibly complementary and sometimes contradicting views.

5.2 Seed Talk 1: Good City Life

Daniele Quercia, Elizabeth M. Belding, Jörg Ott, Andrea Passarella, and Peter Reichl

We invited Daniele Quercia, from Nokia Bells Labs Cambridge, UK, for the first seek talk. Daniele presented the project "Good City Life", as follows.

The corporate smart-city rhetoric is about efficiency, predictability, and security. "You'll get to work on time; no queue when you go shopping, and you are safe because of CCTV cameras around you". Well, all these things make a city acceptable, but they don't make a city great. We are launching goodcitylife.org – a global group of like-minded people who are passionate about building technologies whose focus is not necessarily to create a smart city but to give a good life to city dwellers. The future of the city is, first and foremost, about people, and those people are increasingly networked. We will see how a creative use of network-generated data can tackle hitherto unanswered research questions. Can we rethink existing mapping tools [happy-maps[1]]? Is it possible to capture smellscapes of entire cities and celebrate good odors [smelly-maps[2]]? And soundscapes [chatty-maps[3]]?

Daniele's presentation was very well received, and stimulated also controversial discussions. Among the many points raised, it was questioned the fact that, in general, "better" areas of the cities become more expensive, and therefore making a city "nicer" might lead to excluding vast portions of the population from it. However, there is a middle point to be met between the right to live in a nice environment, and the price of it. More related to Internet design concepts, and to IoP topics, the Good City Life concepts can provide very useful input to design human-centric IoP services and applications, possibly at a global scale. It would be possible to design services to foster interaction between people through urban elements, ultimately exploiting IoP to make services that make people happier.

5.3 Seed Talk 2: IoP – People Centric Designs

Paul Houghton, Elizabeth M. Belding, Jörg Ott, Andrea Passarella, and Peter Reichl

The second seed talk was given by Paul Houghton, from Futurice, Finland. Paul conveyed his experience on human centric services and application designs. Paul took the angle of human-centric IoT (which is a part of IoP), advocating the need to start from a user-centric perspective. He made the case of lego-type IoT (inside IoP), whereby IoT components can be miniaturised and form-factored into lego bricks, that one could mount and compose appropriately. This would also put into the picture gaming-inspired IoT designs. The ultimate goal, would be support extremely cheap IoT systems that any user can build on their own, out of very basic technologies affordable to anyone. An example of a prototype developed along this line is the 3D parametric LEGO-compatible model in OpenSCAD[4], to generate arbitrary size blocks such as electroncs enclosures and mounting panels. First print the calibration blocks, then use those turning parameters to fabricate with a perfect fit using different plastics.

Moreover, Paul also covered an industrial-oriented perspective, envisioning a sort of IoP design kit. He made the point that people want new techonlogies, but most of the time they don't know how to use them. Therefore, we need designing for IoP workshops, i.e., interactive, collaborative tools. The details of such a kit implicitly sets the boundaries and mindset, for better and for worse. The IoT Service Kit is a prototype along these lines presented by Paul during the talk. It is a board game that brings domain experts out of their silos to co-create user-centric IoT experiences and achieve mutual understanding. Clashes and miscommunication between differing perspectives and disciplines can disrupt the workflow. The playful nature of the Kit brings down walls and naturally incites communication. On the other hand, Paul also highlighted that industry adoption needs simple, usable concepts they can map to ideas they already know.

[1] http://www.ted.com/talks/daniele_quercia_happy_maps
[2] http://goodcitylife.org/smellymaps/index.html
[3] http://goodcitylife.org/chattymaps/index.html
[4] https://github.com/paulirotta/parametric_lego

5.4 Panel: IoP around the world

Peter Fatelnig, Pan Hui, Max Ott, Ellen Zegura

After the seed talks, we organised a panel, initially conceived to provide views about IoP-related efforts around the world. For this reasons, we invited in the panel one representative from each continent involved in the attendance. This perspective was taken in the initial presentation from Ellen, who reminded some lessons learned in disruptive Internet designs funded in the US through, e.g., the NSF FIND programme. Moreover, Peter presented the main points of the coming Next Generation Internet H2020 calls, and how they are framed in the more general Internet of People concepts.

Then, the discussion then drifted towards arguments that were already been touched upon during the previous presentations, and expanded the discussion on these points quite a bit. One point that was discussed was the problem of large monopolies, and the typical tendency to for "winner takes it all" phenomena, which happened for Internet in the 70s/80s and now is happening for Facebook. Another important aspect that has been discussed was related to privacy and trust, as a possible perspective to migrate towards a more decentralised, IoP-like paradigm. However, some evidences contradict the typical importance given to these aspects, such as the fact that people uses services like Facebook even if they don't trust them. Again, the information-centric perspective was brought to the table, as one possible incarnation of IoP. An information-centric IoP would be more decentralised than the Internet, in the sense that everyone would "owns" part of the data that make up the network. Therefore, decentralisation would be key for data-centric services, for doing data computation in a privacy-preserving way.

5.5 Group work: IoP definition

Andrea Passarella

After the panel, we split into three separate groups. Before the meeting, we identified some possible topics for discussion, which have been refined before splitting into the groups. The three groups turned out to be homogeneously subscribed, so there was no need to reshuffling or reorganising them. The outcome of all groups were reported at the beginning of the second day of the seminar.

The first group worked on the IoP definition, and came out with a set of features for IoP. The first feature is that IoP would be a network of active Digital Twins. Specifically, there would be one entity per person, representing their identity, also defining the person's profile. Such a digital twin would ideally collect all information about the respective person, which is currently scattered and sometimes inconsistently stored across current Internet services. The digital twin would control access to personal data by external services, thus acting on behalf of "its person", even when s/he is not active in the Internet.

The second IoP feature is that IoP would be a network where the "IP node" is a person. Thus, personal devices would only be incarnations of the person at a lower layer. Devices would communicate seamlessly, exploiting the most appropriate communication means at any given point in time, including global vs local communication, as appropriate. It was found that this could embody even legal frameworks at the "personal IP" layer.

The third IoP feature discussed was that IoP would be a network including human-centric primitives, primarily at the edge. In such a network, personal devices would work with each other based on their users' behaviour, data sharing/management/access being the main focus, according to an information-centric perspective. These novel primitives would be unleashed by more "programmer-friendly" standardised support for local communication. Finally, it was discussed that IoP primitives would complement (and not replace) conventional Internet primitives.

A fourth key feature was considered to be that IoP would be a network bringing value to people, not to Things (or to Big Things). In this sense, there is a huge difference between IoP and IoT, as (i) IoP would be using IoT as a means for people-centric interaction, and (ii) M2M communication per se would be of little value, if not in the context of people interactions. Another key aspect is the relation between IoP and "Big Things", i.e., the fact that data about people behaviour is most of the time of benefit only to the big players in the IoT domain. IoP could provide technical (and non-technical) mechanisms to bring (more?) control by people over their data.

IoP was also seen as an open network for human-centric innovation, thus going back to the roots along which Internet was conceived. IoP is seen as an open ecosystem hosting innovative, unforeseen services, vs. the perception of the Internet as a centralised monopoly in the hands of a few (i.e., Internet = Google + Facebook).

Last, but not least, IoP is seen sas an "organically growing", people-centric Internet, exploiting the analogy of global vs local farming productions (e.g., Monsanto vs organic farming). Along similar lines, IoP would address different needs, between one-size-fits-all managed network and a network that organically grows from personal devices, aggregating and controlling them according to the purpose for their users. This clearly calls for novel ways of decentralised management, governance and control.

5.6 Group work: IoP use cases

Ellen Zegura

The group began with an example provided by Max about a company selling hearing aids that wants to be able to collect data from the devices to improve the product, however the data is considered medical data and cannot cross national boundaries. This is an instance of private data that in aggregate might benefit users of this product (and the company). It is an instance of a clash between Internet boundaries and national policy boundaries. We discussed a re-design where the computations move to the data rather than the data moving to the computation. Will this solve the problem? What if all the data needs to be together? We discussed providing users with greater control over allowing access to their data. We discussed the education challenge that users don't understand their data, and service providers (e.g., Apple) become data gatekeepers.

We discussed the limitations of the Internet in crisis situations, such as those produced by natural and human-caused disasters. We discussed the challenge of enabling collective groups of people to accomplish something immediately and locally, such as citizen volunteers for search and rescue. Current practice is to cobble together digital and non-digital tools in ad hoc and organic ways. We briefly discussed the idea of a human sensor system that would be created and sustained over a long period of time and that could be queried to take the pulse of a community or to track reactions or attitudes over time or to provide large-scale crowd sourced information (e.g., is help needed where you are).

We spent a long time talking about whether and how Facebook is an Internet of People. We discussed why people like Facebook and what they use it for beyond the obvious of staying connected while apart. Examples of uses included humor, cleverness, a pool to tap for commiseration (e.g., Elizabeth's travel woes!), a trusted subgroup to get advice from (e.g., Dagstuhl travel advice), a window into the views of people you don't normally interact with (e.g., political differences in friends of friends). We discussed the risks and value of on-line forums that allow one to shape the presentation of self (see Goffman book), with possible relevance to the digital twin idea. A risk of on-line representation is that social norms will not always carry over. We discussed whether the fact that Facebook is a company making money and gathering data means it cannot be an Internet of People. We forsee a potential tipping point for Facebook as they face pressure to filter certain content, to be more transparent with ads (e.g., Russia buying ads influencing US election).

We talked about the value of fairly immediate feedback from a local crowd, e.g., to learn how to improve a presentation, but also how that feedback is very personal and should not go to the cloud. This capability – private but rapid insight into what people are thinking for personal use– was mentioned as valuable, even vital, for self-development, evolution, and learning. We talked about a tool for gathering feedback from people based on micro-narratives and self-signification, forming a type of self-ethnography. Maybe this is useful for getting anonymous but useful feedback? Tools like this can be used to measure and track culture change.

Our last IoP use case (or perhaps it was an example of an IoP) was community networks. Leonardo shared the scope, history and uses of a number of community networks in Europe and elsewhere. Community networks arise for multiple reasons – in some cases there is nothing else, in others there is a desire to operate without ISP constraints, some favor the philosophical reasons connected to freedom. Many (most?) community networks rely on a sufficient number of tech geeks who have the experience and inclination to manage the nodes. That makes it challenging in communities that lack this expertise. Sometimes community networks serve to create and demonstrate demand that then attracts a commercial ISP to the area and results in the end of the community network. Community-based networks can generate social and cultural capital.

5.7 Group work: IoP and People

Kirsi Louhelainen

The third group discussed about the interplay between IoP and people. What follows is a set of bullet points highlighting the key aspects of the discussion. The discussion was organised around three main themes, (i) a roundtable on establishing an initial perspective on IoP; (ii) reflection on some basic issues, and (iii) discussion aspects.

1. Roundtable: establishing an initial perspective on IoP The key aspects discussed are as follows:
   - IoP as long-term counterperspective to IoT. Internet is for communication between humans, not between things. Sort of philosophical perspective.
   - Internet built by people for people/community networking/democratizing control
   - Embedding social aspects into technology and vice versa.
   - Improve collaboration: enhancing how people work + enable them to form social groups. AI usage for enhancing collaboration.
   - Content is cheap, but you need reputation and trust (not easily duplicated, costly). Tactile Internet (low latency Internet, touch will be transmitted to other person).
   - Milgram-Experiment -> social networks? Networks set up by people, plus trust. Internet of social networks. Services for people. Alexa is not in IoP. Social networking/connections between people.
   - Low latency/AR could be a use case. 2 types of data: information (public interest) vs private information (you want control). IoP pushes both of this to the extreme: 2 silos (public information with low entrance threshold), tied to private information. Allows new digital market place: opening up and generating energy for new service creation. More value in giving up certain control options and instead increase user numbers.
2. Basic issues
   - How do people use the Internet?
   - Social aspects of the Internet
   - People setting up the Internet themselves – (what does setting up mean?)
   - Individuals vs groups vs communities
3. Discussion aspects
   - Is it a local or a global thing? What is the smallest constraint space for sharing public and private information? Storing data, anonymization, history of data.
   - Digital memory: we should not save everything for all time. Audit logs + need to store some information (how many people use service) plus personalized log
   - Mechanisms that allow people to control what happenswith them.
   - Personalization
   - Handling multiple people w.r.t. their needs through some sort of bartering (with automatic convergence)
   - Making users an equally important staktholder as the other stakeholders ("workers' union" type)
   - Charter of Internet rights (e.g. privacy....). Might include the right not to use the Internet!. See IoT manifesto (https://www.iotmanifesto.com/)
   - Fundamental categories: Base rules – physical properties – Internet rights
   - Example: AI checking whether persons are real? One possible IoP principle: NO BOTS. Alternative: autonomous agents who behave like humans = proxies (and share one ecosystem)
   - Tagging suspicious data / means for transparency
   - Is IoP more about traditional Internet sites or about new sites?
   - FAT – fairness/accountability/transparency: use case specific? Limited to contexts?

5.8 The second day

The second day started with a presentation of the three group works of the first day. This was again an opportunity for ample discussion and reasoning about the IoP concepts and topics, with significant participation from all attendees. Afterwards, we organised the day allocating "snippet" presentations by attendees, and a final session with additional group work. As for the previous day, the topics were roughly identified before the seminar, and have then be refined before splitting up. Also in this case, the composition of the groups followed quite naturally attendees interests, and no significant reshuffling was needed. It is worth noticing that the groups composition was quite different from that of day one.

In the following we provide abstracts of the snippet talks, and summary reports from the group work.

5.9 Snippet talk: Finally Closing Up: QoE in IoP

Markus Fiedler and Tobias Hossfeld

So far, Quality of Experience (QoE; "the degree of delight or annoyance") is perceived by users far above the very Internet core layers, aka TCP/IP. The differences in foci have led to a range of unsuccessful QoE modelling and management approaches, and there is still a clear divergence of viewpoints and agendas for QoE and Internet researchers and practitioners, respectively. However, given that the Internet of People (IoP) resides at the top of the communication stack, it comes naturally close to where QoE resides. Thus, there is hope that QoE will be much closer related to IoP principles, provisioning, services, and management, than it has been the case for Internet so far. Thus, any provisioning and control will be much more efficient, creating delight for and reducing annoyance of users, in the best sense of QoE. So we may hope for better user-friendliness (with its many facets) of IoP compared to classic Internet, i.e. for "power to the users" / "power to the people".

In particular, we discuss what is missing in the QoE world like taking into account social interactions between people or the consideration of data and IoT services. For such services, the Quality of Information (e.g. accuracy, timelineness) may be more relevant and contribute to the overall QoE. A major aspect in IoP is privacy which is often mentioned as QoE context factor, but not explicitly addressed in QoE models. Those aspects need to be taken into account in an IoP-aware design in addition to a QoE-aware design. While the primary goal of QoE may be the make the people happy, the question arises at which costs. Machine learning approaches may need to know a lot of private information e.g. user's context, location, preferences to optimize QoE. In IoP, people should be made aware of privacy in an easy way. Internet services and apps should empower the user, i.e., allow the user to decide. This may include the degree of neutrality of recommendation mechanisms or which kind of data to be collected. In summary, IoP requires to allow the implementation of "ethics".

5.10 Snippet talk: IoP for the 99%

Nicki Dell

My goal would be to broaden the conversation to the ~80% of the world's population that lives in the so-called "developing regions". Most of the biggest technological advances have primarily benefitted people who live in western societies. It is crucial that we expand this view. Doing so requires us to think about how to (re)design the Internet and computing experiences to account for cultural, social, linguistic, and socio-economic diversity. People have vastly different value systems and desires. We need designs that support this diversity – combining new technical innovations with human and social aspects of design. A multi-disciplinary approach is essential! HCI, design, networking, law, security, privacy and more. We also need to push beyond the prevalent model of designing for individuals and consider how to design for different groups – families, villages, communities, cities, and so on. How do the design principles change when we expand our view? How do the technical primitives change? How can we come together to design better systems that accommodate people's values, meet their needs, and simultaneously make the world a better place.

5.11 Snippet talk: IoP and Community Networks

Leonardo Maccari

Community networks (CNs) are large-scale wireless mesh networks made of tens, hundreds, or even thousands of nodes that are blooming in many world regions. CNs are organized through a bottom-up, decentralized and participatory process by communities of people, thus they challenge the current for-profit, market-based Internet access model of commercial Internet Service Providers. Today, we know that at a certain scale CNs help to overcome the digital divide where the market fails, however, the degree of innovation of a CN is not only embodied in the number of bits per second it can carry. It resides in the kind of P2P applications that it can enable and that can challenge the current centralized computing model. It resides also in how many people from the currently marginalized groups can get access via the CN. Finally it resides in the extent to which the network can be governed as commons, and not only as a for-profit initiative. Commons-based governance makes it possible to have transparency, participation and to democratize the key decisions on the way networks work.

A CN is an archetypal example of the Internet of People, in which literally every network node is one person. The netCommons project deals with CNs and researches on the way they can scale, be sustainable, offer applications, and interact with society at a broader level.

5.12 Snippet talk: The organic Internet or The Internet of (the) People

Panayotis Antoniadis

The popular Internet platforms that mediate today our everyday communications become more and more efficient in managing vast amounts of information, rendering their users more and more addicted and dependent on them. Alternative, more organic, options like community networks, http://netcommons.eu, and DIY networking, see http://mazizone.eu/, do exist and they can empower citizens to build their own local networks from the bottom-up.

If we wish to facilitate the creation of an "Internet of People" where People are not just extensions of Things, we need to design for diversity, participation, local ownership and governance, and in this sense David Clark's "design for tussle" needs to be redefined in light of the eventual concentration of power over the Internet infrastructure and services, in the hands of very few global corporations. Some of these ideas are included in an upcoming book chapter (submitted draft attached) which will appear in November 2017: http://www.palgrave.com/de/book/9783319665917

5.13 Snippet talk: IoP and Agile wireless network architectures and protocols

Mariya Zheleva

Internet access in rural areas, displaced persons scenarios and cases of political oppression are just a few examples that have demonstrated at scale that the Internet today is far from open, inclusive and equal to all. Emerging agile wireless networks and protocols have the potential to resolve some of these growing limitations and establish the IoP, as defined during this Dagstuhl seminar. Such architectures and protocols will bring innovation both at the network and the client side and will enable local entrepreneurship to foster organic growth and cultivate the diversity of the future Internet.

5.14 Group work: IoP architecture

Andrea Passarella

The group tried to "think architecturally", i.e., to identify key architectural concepts for IoP. The outcome was a set of concepts, summarised as follows.

Firstly, the group discussed about what is an IoP "end-point", particularly, if the IoP node would be a (group of) person(s), or its digital twin. We agreed that we need identity-based routing, which in turns, calls for trust management. But, most likely, we found that data-centric routing should be natively supported as well. A second concept is the fact that we need flexible scoping in routing/data access. Specifically, we don't necessarily need that every-thing, every-one, every-data is accessible globally. So, mechanisms to dynamically define the "visibility" scope of IoP entities are required (moving from local to global visibility). Moreover, there is a need for large-scale measurement studies about locality of data access. We then discussed what should be inside IoP headers. One possibility would be to have "manage-me-like-that" embedded information (e.g., for geo-fencing packets). This might also be a way to support "human-value-centric" forwarding. We also discussed whether this would be essentially similar to active networking.

A significant part of the discussion was related to whether we need a "narrow-waist", and what this would be, in the case. The group agreed that a natural narrow-waist would be the social graph(s) of the IoP users. In this case, the abstractions of nodes would be Persons, Communities, "Legal entities" behind "things". This would be probably a multi-dimensional (or a hyper-) graph, to accommodate for the different roles each person takes at different points in time. But then, how do we account for trust? In a completely centralised manner? But then, we would need a globally trusted entity, which might be quite questionable. Or, should trust be a completely distributed and subjective way, i.e., one of the properties of a link on the social graph? Another related concept is how do we cope with dynamism, such as, e.g., stable social relationships vs "ephemeral" social relations. In this context, we should take into account that the social graph would subjective, and each node would have its own view of it. Scalability issues were also discussed, i.e., whether we should consider one gigantic flat graph vs a hierarchical graph, at the hierarchical levels of persons; communities; groups of communities, ... Finally, we discussed what would be the relationship with the current Internet stack. Most likely, we would use the current stack when appropriate (e.g., for global communication). But the, an issue is how to integrate IoP with "conventional" traffic engineering approaches (e.g., fairness). We should also be open to use other "transports" when more appropriate, e.g., in case of local communications.

5.15 Group work: Privacy vs. Sharing and Knowledge Creation

Panayotis Antoniadis, Nicola Dell, Thorsten Strufe

"Sharing is caring, privacy is theft, secrets are lies" – Dave Eggers (The Circle)

The Internet of People is based on individuals communicating with each other – one-on-one, in small groups, or large forums. These communications may create "leave-behind" artefacts, such as posts, photos, or videos, to facilitate the ongoing conversations, or may be shared wth a wider group (e.g. public blog post), In short, in the Internet of People, information is primarily created, curated, and consumed by the People and for the People. The creation, and especialy curation and aggregation aspects may be supported by services (e.g agents, bots, ...) and in turn leverage social relationships and cross group boundaries with great opportunities towards the commons, benefitting everybody. [...for routing and providing its services, and the aggregation and processing of the collective behavior and data may offer great opportunities towards the commons, benefitting everybody....] This also implies that individuals may be observed and tracked, their data accessed by potentially unintended audiences, with potentially adverse or even dangerous consequences for the IoP participants.

For the purpose of systematizing this spectrum it makes sense to understand notions of privacy, threats, potential benefits, and the factors that may lead to an outcome in which the benefits and drawbacks for all stakeholders can be balanced.

Privacy as an abstract concept essentially has a very different meaning in different cultures. This has been discussed at great length in the context of the difference between the Anglo-Saxon and the European notion of privacy: The former being shaped by the right to be let alone (or: freedom of processing, the regulation of markets, and an intrinsical opt-out notion; trust in companies and distrust in governments), whereas the latter traditionally follows the notion of data sovereignty and informational self-determination (or: intrinsically opt-in and control over the data throughout its lifetime; distrust in both governments and companies). This discussion, while quite prominent, has ignored the profound differences compared to other cultures. Many Asian and African countries, for example, don't only exhibit entirely different utilization of electronic devices and services, but are also characterized by different privacy expectations. The discussion also ignores the discrepancy between the legal, idealistic, and real situation: The European perspective fails to address the aggregation and continued processing of aggregates (which part can you take back? What does the difference between the aggregates before and after disclose about the data that one wants to take back?), and all current notions ignore that personal data often shares dependencies between individuals (Statistics can disclose seemingly hidden attributes that individuals do not want to share. The data of groups may disclose private attributes of its individuals, with the extreme of statistically similar DNA sequences between relatives , where some may want to publish, and others hide parts of this shared information).

Sovereignty and responsible action imply that the individuals and stakeholders actually comprehend the value of data. This raises additional challenges. It may not even be possible for an individual to assess the value of the data it is willing to share or expose in terms of recorded behavior, as the current, advertisement-driven market values the various data of individuals differently: being able to identify and analyse hyper-consumers and influencers of course is much more valuable, than collecting yet more data about the average Jane (both in terms of numbers of average vs peculiar individuals, as well as in terms of expected spending capacity and influence). However, this role in the overall audience is difficult or impossible to judge for the individual itself. It additionally is difficult, may be impossible, to gauge how the exposed data can be mined, what happens with the aggregates, and most probably unfeasible to even guess how these data sets can be linked, correlated and mined in the future [5]. Another observation is that many stakeholders (primarily companies) currently collect and store data about individuals without analysing them, nor having a clear plan or even ideas about how to process them and for which purpose in future. It's just simple to collect and to store just in case. A third observation in this context is that the market valuation and income of companies actually is only indirectly related to the data they collect, but directly related to the type and extent of audiences, whose attention they can sell. The targeting of specific groups requires knowledge about the individuals, but the business model is primarily based on selling attention, selling actual data or aggregates (at least as observed by the public) seems to be a secondary income, if it represents a notable income at all. Changing perspectives could hence be a sensible approach: Privacy should probably not so much be viewed as the value an individual allocates to its data, in the decision of sharing – but rather as the potential risk to the welfare and well-being if "lost" to the public domain, or commercial and institutional parties.

[5] This observation also challenges the initiative of the GDPR to oblige all data collectors and processors to comprehensively explain results and ramifications from data processing

An important factor in this picture is the question of trust: The privacy of different personal attributes of course depends on the audiences that get access to them. Considering an IoP, it may be perfectly acceptable for individuals to share sensitive information, like for instance their location, to their significant other, their family or friends, their colleagues, or neighbours. This could also have a geographic aspect: while it may be undesirable to share this information globally or with a remote, commercial provider, it may be perfectly fine to share the location with people in the direct vicinity, probably even with small local businesses. It is also well conceivable that this trust is based on interest or other properties, and platform collectivism, in which all participants in a system share their respective data with everybody else on the platform, is well conceivable in the IoP.

This raises the question of the architecture and stakeholders of services on the IoP. Considering social media like services, the current architectures comprise of the three immediate stakeholders of users, providers, and (advertising) customers, as well as society as a the general context. The current discussion of privacy has a strong focus on the users, who are expected to understand how their actions and their data could be collected, used, and their exploitation have a potential adverse effect. The common narrative hence claims the responsibility to be with the users, who should know what to share to whom, not to overshare, to use the audience selection mechanisms appropriately, etc. The responsibility of the providers is commonly conveniently avoided, despite the fact that only the providers could even remotely assess the value, make informed decisions about which data is rather common or sensitive, and could potentially provide effective protection of the data. The providers so far, however, have no incentive to protect, avoid, or even minimize data, and hence push towards even more sharing with even larger audiences, going as far as claiming that the post-privacy culture was the future.

This imbalance of responsibility is even more pronounced by the lack of credible information on the current uses of one's data (at least theoretically private data can be used not only for "innocent" advertisement but also for manipulation of behaviour, addition tactics, and more), but most importantly on the future potential uses, ad in the case of a change of political situation (e.g., a dictatorship).

Considering the incentives between the stakeholders it becomes obvious that they are currently not aligned, and that it may make sense to reflect on the prime driving instincts of fear and greed. The optimistic view here would suggest to create markets in which it is beneficial to sell services and devices that allow for privacy-preserving utilization and deployments of the IoP. Businesses could offer such devices that guarantee good services under protection of the sensitive, personal information of the users, and the invisible hand of the market could take care of the remaining, insecure service providers. A more pessimistic view would suggest to focus on the responsibility of the providers, and align their incentives with that of their users. An approach could be regulation and severe penalties for data loss incidents. The GDPR includes first steps in this direction, threatening the providers with fines of up to 4% of their annual turnover in case of the loss or maltreatment of personally identifiable data. This observably has caused several companies to rethink their current practice of collecting everything, just for the potential case of future opportunities (or neglect). A first step in this direction could be the requirement for companies to put the data collections they hold on their annual balance sheet. Depending on the approach this could be seen as either an asset, or a risk. In any case, this would raise the level of attention to the board of directors, and hence become a point for consideration for the CFO's and CEO's.

Sharing data may of course generate knowledge, which represents a value, potentially public, in itself. Keeping everything private may hence in fact affect not only the IoP, but even the data owner adversly. It is quite likely that the participants in an IoP will prefer to enjoy the advantages from functions over everybody's data, which is only possible if a noteworthy fraction of the participants actually do share some information. But it is also likely that if they had the power, they would enjoy to benefit also the additional value that their shared information generate (beyond knowledge, also in economic profit). Despite the fact that it is difficult to judge the sensitivity of data, a solution could be to distinguish between public and private data (or: aggregates), share the data that is less sensitive to the public domain, and allow for the local processing of the complete data under the control of each individual, consolidating the public and their own private data.

Taking the role of an engineer, it becomes apparent that the common tools will play a role in the Internet of People: It will need functionality to generate awareness in the users, and it shall provide transparency of the algorithmic results, and as a basic foundation of its design. An extension is accountability – the repudiation of acts, especially of institutional or commercial parties should be avoided. This, however, is a double-edged sword, as means for accountability can directly play into the hands of populistic or even totalitarian regimes, that may require accountability of even innocuous acts of individuals, thus preventing anonymity. A direction offering solutions to this conflict could be tools for obligations management, encapsulating both data and obligations for the recipients, thus explicitly allowing or prohibiting propagation, aggregation, or analysis. Experience with digital rights management in the past, however, has depicted the natural limitations of this approach.

A direct solution with natural fit to the Internet of People paradigm could be a personal device for secure data storage and processing, the "Decentralized Privacy Box". Sold to or built by the participating individuals, it could offer guaranteed secure computation (for example through the integration of Trusted Execution Environments, like the Software Guard Extensions of Intel, SEV of AMD, or similar extensions; or through implementations of secure multiparty computation or simplified algorithms on homomorphically encrypted data). A typical scenario could be the retrieval of public data and local processing of recommendations or added value services with access to the local, personal data. It would also allow for functionality in which two individuals share their private information with eachother, facilitating functionality leveraging both data sets, but preventing access to the private data of the opposite party. Joining various datasets, and potentially removing the sensitive personal parts, the data aggregates could be shared back to the community, the platform, or even the public domain. Micro-payments could further incentivize the participation in services on public data, with subsequent improvement of the public data after augmentation with the local, personal information.

Another, complementary, research direction to pursue in this context concerns the tools (technical but also legal, social, and political) for the "People" of the IoP to be able to create organizations of different scale (at a neighbourhood, city, or even national level), that will enable to participate in some of the aforementioned decisions and take ownership and control of their data and the value generated by it. Platform cooperativism is a recent term that resonates with such ideas, but the design space is very broad and perhaps the best strategy is to provide options, to redefine David Clark's concept of "design for tussle" in the case of IoP.

In summary, the Internet of People paradigm seems to direct towards decentralization of services, giving higher responsibility and probably less access to large entities and creating a more level playing field between all stakeholders than we see today. Sharing towards trusted audiences, providing de-identified aggregates and augmentations of public data shall further knowledge and provide benefit to all. Acknowledging the privacy implications, this has to be done with care – and a decentralized approach, with privacy boxes implementing proven secure functionality as end-user devices seems a promising vision.

5.16 Group work: From Internet to IoP

Markus Fiedler

The discussion focused on the concept of the "Digital Twin" (DT) as a representative of the user in terms of communication-related needs and preferences, and is summarised through the following set of crystallisation points.

- **Feature list:** The DT is a repository of user-owned data and user-related settings. This mandates support of configurable levels of privacy, dependent on the context of the current usage. Likewise, the DT should take care of the user's communication, choosing the best-suited connectivity (in terms of quality, security and economy) for the user. Thus, the DT needs personalisation and configuration facilities. Furthermore, it DT needs to be reachable and thus be addressable and routable from outside.
- **Architecture:** The DT represents a peer in an overlay concept, with corresponding personalised peer-to-peer communication. Given the plethora of desired features in combination with a step-by-step development path, a modular design appears to be mandatory. More discussions on the architecture can be found in Section **??**.
- **Groups:** The DT should support groups, which entails the needs for dynamic configurations and feature interactions.
- **Governance:** Through its personalisation and configuration features, in particular with regards to granting (and revoking) access to personal information, the DT implements the principle "power to the people".
- **Enemies:** Certain social networks have been identified as having conflicting views and implementations of information ownership and (missing) user control.
- **Business models:** In order to power a system of DTs, the DT peers need to contribute to its operations, *e.g.* through using some micro-currency, as alternative to the contemporary "data milking" by large players on the ICT market.
- **Regulatory issues:** If correctly implemented, the DT concept allows for data minimisation. Furthermore, it is expected that regulatory bodies get more possibilities to act against non-compliant stakeholders (*i.e.* a "bigger stick").
- **Implementation:** In order to allow for a successful growth, parallels to the Internet development can be drawn, with bottom-up (instead of top-down) principles; trial-and-error approaches; and workable instantiations.

The presentation to the plenary had the subtitle "...and what coffee's got to do with it". Indeed, parts of the discussion were inspired by the personas of a South American coffee farmer, who should benefit from the IoP without ending up in any communication, configuration or privacy hassle.

So far, no tangible transition plan could be envisioned; the group foresees the emergence of the DT to happen in an Internet-typical bottom-up fashion. Still, the urgent issues at hand are not technical, but related to the stakeholders' attitudes, in particular regarding to ownership and privacy of user-related data. A transition away from the information ownership models of large social networks to "people in control of their privacy" is badly needed in order to pave the way towards a successful IoP.

5.17 The third day

The third day was devoted to two main aspects, i.e., discussing the outcome of the previous day group work, and identifying next steps. To accomplish the second task, we again split in three groups, one focusing on the IoP toolkits, one on IoP research agenda dn roadmap, and the third one drafting an IoP manifesto.

In the following we provide summaries of the outcome of the three groups.

5.18 Group work: IoP toolkits

Panayotis Antoniadis

Toolkits can play a key role in empowering people over the control and design of "their" Internet. The reason is that technology is not neutral and an "Internet of People" should allow for the customization of local infrastructure and services according to the needs and values of smaller or bigger groups of people that wish to democratically co-create the technologies that affect their lives.

In this context, both the design and implementation processes require significant expertise and for this only with powerful and flexible toolkits one can ensure that the Internet of People is owned, designed, and controlled, actually by the people.

Additional toolkits and guidelines are also needed other enabling and facilitating actors like researchers, community organizers and more.

In this working group we focused on two main type of toolkits needed for the Internet of People, on participatory design and DIY implementation.

First, for the participatory design toolkit of the technology itself, the IoP:

- Example: Paul's IoT toolkit, physical objects, toys, cards, maps
- geographic vs. abstractions
- boundary objects (MAZI's transdisciplinary methodology)
- 3 predefined examples

What is different in the case of IoP compared to those many existing toolkits? Mostly the concrete target technology unique, which is beyond software services but include the network infrastructure itself and most importantly the corresponding governance procedures, legal aspects, and more.

One can build on lessons learned from the participatory design literature like focusing on stories and asking people about their place in the world before going into more details.

Of course, the cost of decision making shound not be neglected and for this the IoP participatory design toolkit should include the visualization of trade-offs regarding different design variables and also comprehensive "translations" between design choices and outcome in terms of key values like privacy, anonymity, degree of individual choice, etc.

Second, the DIY implementation toolkit was quickly summarized with the "IoP in a box" concept. In this context there is related work in the context of Community Networks (CNs) and DIY networking with the toolkits by Commotion, https://commotionwireless.net/docs/cck/, and MAZI, http://mazizone.eu/toolkit/ being the most advanced today. A key requirement for such a toolkit to be effective is to include primitives that already work and at the same provide rich options for customization, configurable elements.

5.19 Group work: Agenda and Future Research Topics

Gareth Tyson

This section covers discussions from the Agenda and Future Research Directions break-out group. It lists key opportunities and research challenges. It is structured in a roughly chronological order, however, many of the tasks are closely interconnected.

Requirements, philosophy and implications of IoP. Before re-architecting any technology, it is first necessary to understand the socio-techno and even philosophical underpinnings. Hence, the first step must be to lay out a series of goals, considerations and implications. This should be embedded within a manifesto that delineates the key goals of IoP, its requirements, its intended outcomes and any desiderata. Embedded within this should be a robust state-of-the-art review to understand past pitfalls and future opportunities within this broad landscape. As part of this, we envisage that transparency will be a key aspect of the IoP, such that people can reason over the wider ecosystem (from design to deployment, and beyond). Building transparency tools (e.g. measurements, visualisations) will therefore be a major part of the manifesto.

Architecturing the Digital twin. A common discussion point within the groups was the concept of a Digital Twin (or cyber-me). This constitutes an always-on digital presence that (1) Stores and mediates access to all online data related to an individual; and (2) Acts on behalf of the individual regarding certain authorised activities, e.g. negotiating exchange of data. Consequently, a major step would be: defining the data structures that would be maintained within a Digital Twin; the ways that such data could be accessed and exchanged; the forms of agency such a Twin could have; the manner in which the Twin would be hosted and managed from a infrastructural/systems perspective; and the ways that the individual and their Twin would interact. This would further raise a number of critical legal, ethical and sociological questions regarding the extent to which the individual would be responsible for actions performed by the Twin.

Micro-level Innovations. If we assume that the Digital Twin will constitute a key primite within the IoP, it will next be necessary to exploit it to fulfil the goals specified within the IoP Manifesto. We do not intend to deviate from the current OSI-layered Internet model. However, we envisage that the Digital Twin, and its related wider social information, will feed into this modelled architecture such that layered decision making is informed by the person-centric insights captured within the Digital Twin (and any other related data structures and agency algorithms). For example, socially-informed congestion control may be introduced at the Transport Layer. These types of per-second transactional innovations are considered micro aspects.

Macro-level Innovations. If we consider micro aspects as per-second transactional activities, macro-level innovations pertain to longer-term strategic factors. Currently, the Internet is a composite of many stakeholders – dominated by a small number of hypergiants, e.g. Google, Facebook, AT&T, Cogent etc. The IoP will promote people to the equivalent power position held by these hypergiants. In othe words, the IoP will allow people to negotiate and drive forward strategy decision making with equal force to any existing hypergiant – it will democratise Internet governance. This would involve people (and their Digital Twin) unionising to exert influence on other stakeholders. On a computational-level, this would require the specification of formal interfaces between stakeholders, allowing the exchange of negotiation-like dialogue. This would, of course, be complementary to offline interactions, whilst allowing real-time decision making to take place. Empowering users via this unionisation is critical to enabling change, and for incentivising existing hypergiants to move towards the principles laid out in the IoP Manifesto. This is particularly relevant in the face of the growing number of “gig economy” platforms, which tend to disempower individuals in favour of global operators.

Transitional Considerations: From IP to IoP. Assuming the above technical innovations are successfully designed and implemented, it would next be necessary to enable deployment. As many past efforts (e.g. IPv6, multicast, QoS) have shown, this is not always trivial. It would therefore be vital that transitional considerations are made both during the design and the deployment of IoP. This would not only raise technical challenges, but also issues of governance, business, regulation and legal factors. This would extent beyond the impact on existing network and service operators to include the needs of existing Internet users, who may not necessary wish to engage in the IoP. To be truly people-centric, such users must be considered and given the freedom to leave (whilst maintaining the benefits of the current Internet). Fundamentally, it must be possible for both IP and the IoP to co-exist -only through this will be successful evolution and transition be attained.

Use cases & Killer Application. A frequent criticism of Future Internet architectures is their lack of a “killer application” to motivate uptake. Thus, the identification of such killer applications should be integrated into the design process from the start. These use case applications would then form the basis for evaluation. Critically, it must be shown that the IoP enabled fundamentally new capabilities that go significantly beyond that offered by IP. Key Performance Indicators might include fairness, privacy, energy efficiency, and traditional measures of Quality of Experience (e.g. MOS). Applications that have been discussed include using the Digital Twin to perform offline negotiation on the individual’s behalf; using social information to fulfil the needs of users, e.g. recommendations, pre-fetching of content; using the Digital Twin to mediate and protect user data. Importantly, the IoP should also underpin an innovative and open ecosystem, where any entrants can contribute and expand on these initial ideas. The IoP should therefore encourage bottom-up innovation, liberating individuals from the barriers of entering new digital markets – such principles would be laid out in the manifesto.

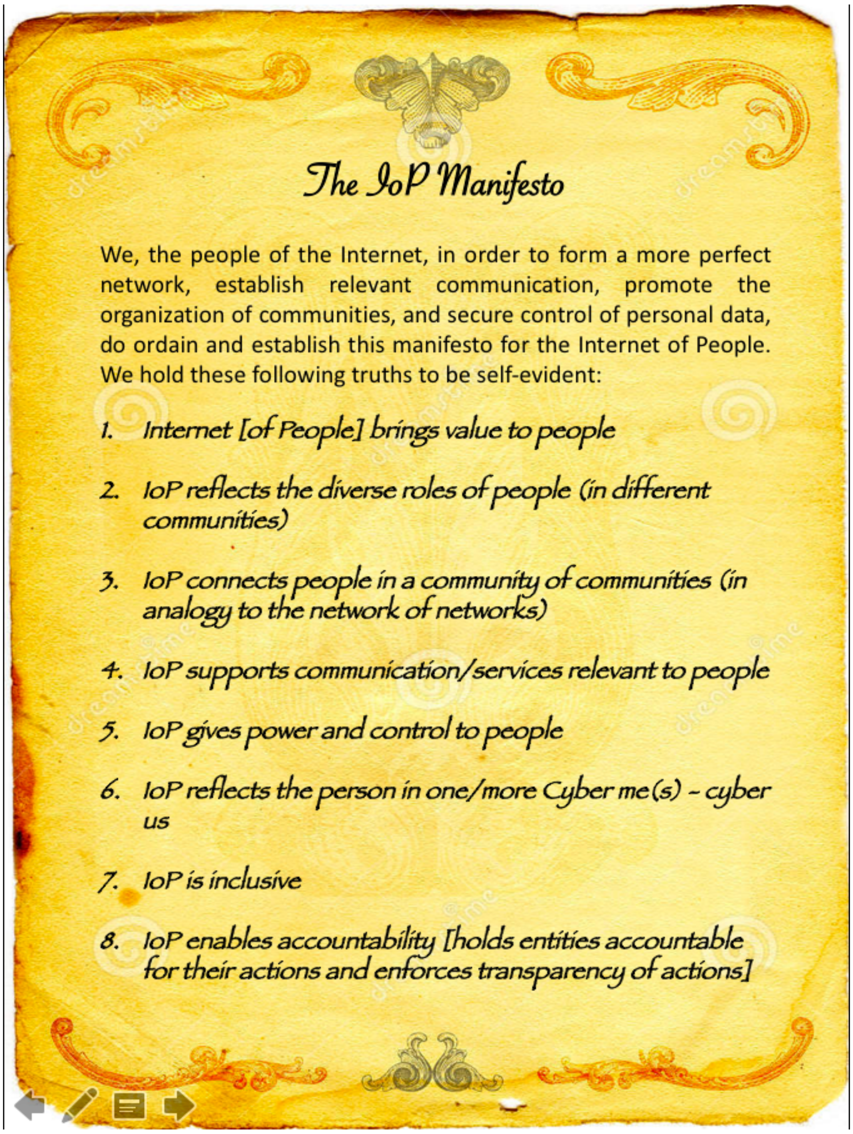
5.20 Group work: IoP manifesto

Anders Lindgren

The best illustration of the outcome is the manifesto itself, as in Figure 1.

6 Conclusions

The seminar was a very good opportunity to bring together a community of researchers interested in the topic of Internet of People, to discuss about this research area during an intense two-and-a-half-day seminar. People arrived to the seminar with different complementary



The IoP Manifesto

We, the people of the Internet, in order to form a more perfect network, establish relevant communication, promote the organization of communities, and secure control of personal data, do ordain and establish this manifesto for the Internet of People. We hold these following truths to be self-evident:

1. Internet [of People] brings value to people
2. IoP reflects the diverse roles of people (in different communities)
3. IoP connects people in a community of communities (in analogy to the network of networks)
4. IoP supports communication/services relevant to people
5. IoP gives power and control to people
6. IoP reflects the person in one/more Cyber me(s) - cyber us
7. IoP is inclusive
8. IoP enables accountability [holds entities accountable for their actions and enforces transparency of actions]

Figure 1 The IoP manifesto.

views, which helped stimulating useful discussions. Overall, we can tell, also looking at the feedback provided by attendees, that the seminar was successful, and attendees have been very happy to take part to it.

The topics discussed ranged from the definition of IoP, to privacy aspects, architectural approaches, security and privacy. We also covered topics such as QoE in IoP, and the need to account for the 80% of the population that is living in developing countries. Thus, the role of people in IoP was largely debated, as well as use cases for this brand-new concept.

Outcomes of the seminar consisted in elaborating a possible research roadmap, outline a set of toolkits, and defining an initial IoP manifesto. Even beyond that, the seminar put together a community of motivated researchers across the world, who had the opportunity to share ideas and initially shape a possibly hot research area for the Next Generation Internet. In the view of the organisers, establishing such a community was one of the primary goals of the seminar, which has been thus fully achieved.

References

1 Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019,” Cisco, Tech. Rep., http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-indexvni/white_paper_c11-520862.html, 2015.

2 M. Conti, S. K. Das, C. Bisdikian, M. Kumar, L. M. Ni, A. Passarella, G. Roussos, G. Troster, G. Tsudik, and F. Zambonelli, “Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence,” Pervasive and Mobile Computing, vol. 8, no. 1, pp. 2–21, 2012

3 P. Reichl, “From QoS to QoE: Buzzword Issue or Anti-Copernican Revolution?”, Keynote abstract, Proc. EuroNF Workshop on Traffic Management and Traffic Engineering for the Future Internet, p. 23, Dec. 2009.

4 E. Daly and M. Haahr, “Social network analysis for information flow in disconnected Delay-Tolerant MANETs,” IEEE Trans. Mobile Comput., vol. 8, no. 5, pp. 606–621, May. 2009.

5 P. Hui, J. Crowcroft, and E. Yoneki, “Bubble rap: Social-based forwarding in delay tolerant networks,” IEEE Trans. Mobile Comput., vol. 10, no. 11, pp. 1576–1589, Nov. 2011.

6 M. Conti, M. Mordacchini, and A. Passarella, “Design and performance evaluation of data dissemination systems for opportunistic networks based on cognitive heuristics,” ACM Trans. Auton. Adapt. Syst., vol. 8, no. 3, pp. 12:1–12:32, 2013.

7 P. Reichl, “Quality of Experience in Convergent Communication Ecosystems”, 2013.

8 P. Reichl, A. Passarella: Back to the Future: Towards an Internet of People (IoP). Invited Paper, Proc. MMBNet 2015, Hamburg, Germany, September 2015.

9 Marco Conti, Andrea Passarella, Sajal K. Das, The Internet of People (IoP): A new wave in pervasive mobile computing, In Pervasive and Mobile Computing, Volume 41, 2017, Pages 1-27, ISSN 1574-1192, https://doi.org/10.1016/j.pmcj.2017.07.009.

Participants

- Panayotis Antoniadis
Nethood – Zürich, CH
- Chiara Boldrini
CNR – Pisa, IT
- Dimitris Chatzopoulos
HKUST – Kowloon, HK
- Nicola Dell
Cornell Tech – New York, US
- Peter Fatelnig
European Commission
Brussels, BE
- Markus Fiedler
Blekinge Institute of Technology –
Karlshamn, SE
- Huber Flores
University of Helsinki, FI
- Heikki Hämmäinen
Aalto University, FI
- Tobias Hoßfeld
Universität Duisburg-Essen, DE
- Paul Houghton
Futurice Oy – Helsinki, DE
- Pan Hui
HKUST – Kowloon, HK
- Teemu Kärkkäinen
TU München, DE
- Eemil Lagerspetz
University of Helsinki, FI
- Anders Lindgren
RISE SICS – Kista, SE
- Pietro Lio
University of Cambridge, GB
- Kirsi Louhelainen
Barona Technologies –
Helsinki, FI
- Leonardo Maccari
Università di Trento, IT
- Jörg Ott
TU München, DE
- Maximilian Ott
CSIRO – Alexandria, AU
- Andrea Passarella
CNR – Pisa, IT
- Daniele Quercia
NOKIA Bell Labs –
Cambridge, GB
- Peter Reichl
Universität Wien, AT
- Jatinder Singh
University of Cambridge, GB
- Thorsten Strufe
TU Dresden, DE
- Gareth Tyson
Queen Mary University of
London, GB
- Ellen Zegura
Georgia Institute of Technology –
Atlanta, US
- Mariya Zheleva
University of Albany –
SUNY, US
- Martina Zitterbart
KIT – Karlsruher Institut für
Technologie, DE

Report from Dagstuhl Seminar 17421

Computational Proteomics

Edited by

Bernhard Küster[1], Kathryn Lilley[2], and Lennart Martens[3]

1 **TU München, DE,** kuster@tum.de
2 **University of Cambridge, GB,** k.s.lilley@bioc.cam.ac.uk
3 **Ghent University, BE,** lennart.martens@ugent.be

Abstract

The Dagstuhl Seminar 17421 "Computational Proteomics" discussed in-depth the current challenges facing the field of computational proteomics, while at the same time reaching out across the field's borders to engage with other computational omics fields at the joint interfaces. The ramifications of these issues, and possible solutions, were first introduced in short but thought-provoking talks, followed by a plenary discussion to delineate the initial discussion sub-topics. Afterwards, working groups addressed these initial considerations in great detail.

Seminar October 15–20, 2017 – http://www.dagstuhl.de/17421
1998 ACM Subject Classification Bioinformatics, Computational proteomics, Metabolomics / metabonomics, Transcriptomics
Keywords and phrases computational biology, computational mass spectrometry, integrative bioinformatics, large scale public data, proteomics
Digital Object Identifier 10.4230/DagRep.7.10.69

1 Executive Summary

Lennart Martens

The Dagstuhl Seminar 17421 "Computational Proteomics" discussed in-depth the current challenges facing the field of computational proteomics, while at the same time reaching out across the field's borders to engage with other computational omics fields at the joint interfaces. The issues that were discussed reflect the emergence of novel applications within the field of proteomics, notably proteogenomics (the identification of proteins based on sequence data obtained from prior genomics and/or transcriptomics analyses), and metaproteomics (the study of the combined proteome across an entire community of (micro-)organisms). These two new proteomics approaches share several challenges, which predominantly revolve around the sensitive identification of proteins from large databases while maintaining an acceptably low false discovery rate (FDR). The ramifications of these issues, and possible solutions, were first introduced in short but thought-provoking talks, followed by a plenary discussion to delineate the initial discussion sub-topics. Afterwards, working groups addressed these initial considerations in great detail.

In addition, both proteogenomics and metaproteomics suffer from coverage issues, as neither is currently capable of providing anywhere near a complete view on the true complexity of the (meta-)proteome. This issue is exacerbated by the fact that the true extent of the proteome remains unknown, and is likely to be time-dependent as well. As a result, a separate

Computational Proteomics, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 69–83
Editors: Bernhard Küster, Kathryn Lilley, and Lennart Martens
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

working group was created to discuss the issues and possible remedies related to proteome coverage.

The field of proteomics has, however, not only extended into novel application areas, but meanwhile also continues to see a strong development of novel technologies. Over the past few years, the most impactful of these is data-independent acquisition (DIA), which comes with its own unique computational challenges. On the one hand, the analysis of DIA data currently relies heavily on spectral libraries, which have so far been a rather niche product in proteomics (as opposed to, for instance, metabolomics, where spectral libraries have a much longer and much more fruitful history), while on the other hand, FDR estimation remains contested in DIA approaches. As a result, two further working groups were established during the seminar, one on the applications for, and methods to create spectral libraries, and the other on the specific challenge of calculating a reliable FDR when performing spectral library searching.

Another key topic of the seminar was the (orthogonal) re-use of public proteomics data, which focused on the provision of metadata for the assembled proteomics data, as this is the key bottleneck facing researchers who wish to perform large-scale re-analysis of public proteomics data, especially when the objective is to obtain biological knowledge. A working group was therefore created to explore the issues with metadata provision, and to explore means to ameliorate the current suboptimal metadata reporting situation.

Throughout the seminar, the topic of visualizing the acquired data and the obtained results cropped up with regularity. A corresponding working group was therefore set up to delineate the state-of-the-art in proteomics data visualization, and to explore the issues with, and opportunities of advanced visualizations in proteomics.

As a last core topic, a short introductory talk and subsequent working group was dedicated to the education of computational proteomics researchers, with special focus on their ability to work at the interfaces with other omics fields (genomics, transcriptomics, and metabolomics). This working group assembled an extensive list of already available materials, along with an overview of the different roles and specializations that can be found across informaticians, bio-informaticians, and biologists, and how each field should evolve in order to bring these more closely together in the future.

In addition to abovementioned topic introduction talks, and the associated working groups, two talks illustrated specific topics of the seminar. Paul Wilmes showed his recent work in bringing metaproteomics together with advanced metatranscriptomics and metagenomics, showing that the flexible use of sequence assembly graphs at the nucleotide level opens up many highly interesting possibilities at the proteome level through enhanced identification. Nevertheless, it was observed that there is strong enrichment for genes with unknown function at the protein identification level, highlighting quite clearly that we have yet to achieve a more complete biochemical understanding of microbial ecosystems. Finally, Magnus Palmblad delighted the participants with a highly original talk on the exploration of mass spectrometry data (of both peptides as well as small molecules) through the five senses (sight, hearing, touch, smell, and taste).

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Exploration of the key computational interfaces between omics domains

Frédérique Lisacek (Swiss Institute of Bioinformatics, CH)

URL https://prezi.com/cfmvkw884qmr/dagstuhl-oct17/

Connecting glycomics with proteomics and interactomics raises many issues. To begin with, protein glycosylation and its impact on structure and function is widely ignored probably due to the lower throughput of glycomics experiments in comparison to other omics. Nonetheless this modification along with many others generates proteoforms and the extent of this repertoire as well as possible cross-talk between modifications remains difficult to study and evaluate. Another obstacle in linking glycomics with other –omics is the independent accumulation of data regarding the constituents of glycoconjugates. Most glycan structures are solved after being cleaved off their natural support and key information on the conjugate is lost. Conversely, protein glycosylation sites are mapped independently of the glycan structure and key information on the attached glycan is lost. Glycoproteomics is on the rise and a promising technology that preserves associated glycan and peptide data though data submission and sharing remains confidential at this stage. Despite the numerous gaps challenging software development, rapid change is expected in this field in the years to come. Furthermore, the role of glycans comes to the forefront in many biomedical applications including for example microbiome studies; glycans mediate specific protein-protein interactions.

3.2 Interpretation of proteomics and transcriptomics to model the dynamics of gene expression

Gerben Menschaert (Ghent University, BE)

The task of integrating proteomics and transcriptomics data faces several challenges. First, there is the complexity of the proteome (due to the wide variety of proteoforms), which confounds the specifics of the (re)-annotation mechanism to employ, not in the least because of issues with the mapping of sequence information between RNA and protein. This also impacts the correlation of quantification between the different ‘omics data types. Specific issues are encountered in the context of immunopeptides, where the sequences are possibly not directly genomically encoded. Because of these issues, it is clear that specific tools and algorithms need to be developed for proteogenomics. The following topics were described in more detail as indicators of where the field is moving. Novel sequencing technologies that are currently developed (PacBio, SMRT-seq and MinION, Oxford Nanopore) will shed new light on the identification of novel translation products from lncRNAs, sORFs, uORFs. But ot make the most of these, cell-type or tissue-type matching across-omics datasets will be needed. Moreover, these novel findings will also have to be deposited into public repositories, as these can then be used for genome re-annotation. In that context, it is necessary to develop a better integration between PRIDE (proteomics) and Ensembl (genomics/transcriptomics). The issues surrounding such integration are missing metadata, stringent filtering for false positives and a need for robust workflows. From a quantitative correlation perspective, robust implementations are needed to compare sequencing-based semi-quantitative measures with mass spectrometry derived quantitative metrics. FInally, there is a need to further improve upon, or develop de novo, tools for the integration with genomics, for the elaboration of standards such as proBED and proBAM to map proteomics data to genome browsers, and for establishng tighter connectoins with platform interfaces like BioConda/Galaxy.

3.3 Assessing and addressing the specific computational challenges of metaproteomics

Thilo Muth (Robert Koch Institut – Berlin, DE)

In recent years, the impact of microbial consortia on human health has gained increasing attention due to the acquired knowledge regarding the important role of the intestinal gut microbiome. Metaproteomics, the mass spectrometry based proteomic analysis of an entire microbial community, helps to elucidate enzymatic capabilities and taxonomic origin of key species. Accordingly, this method can also be used for detecting pathogens in samples from which the exact microbial origin is not known. However, several computational challenges exist of which the most severe ones are highlighted here. First, there is the protein inference issue, which is worsened in metaproteomics because of the occurrence of multiple homologous species, with many homologous protein sequences, in complex and heterogeneous samples. Second is the need to select an appropriate database that covers a sufficient amount of relevant species without too strong a bias (e.g., towards clinical relevant strains). Third is the large and constantly growing size of the resulting sequence databases, which affects FDR estimation and/or sensitivity of the searches in target-decoy approaches. Fourth, biologists typically want answers to very specific questions, which require much more than a protein identification list; for instance, identifying a certain pathway, its functional proteins, and its related microbial species. On the computational level, the combination of multiple database search engines is shown as a reasonably straightforward means to increase the detection rate at the taxonomic level. Finally, the current status and performance of de novo sequencing algorithms is demonstrated, along with their potential and limitations when used as alternative approaches to database driven peptide identification.

3.4 Exploring mass spectrometry data with the five senses

Magnus Palmblad (Leiden University Medical Center, NL)

Another view on mass spectrometry data, exploring the possible applications of sight, hearing, touch, smell, and even taste in the sensory perception and analysis of mass spectrometry data. This is illustrated with anaglyphs, audio of mass spectrometry transients and spectra, and 3D printed mass spectra, complete with chromatographic and isotopic dimensions. A small contest will be held to guess the mass of a compound for which the isotope distribution is given as an image, and another which is provided as a 3D printed model in a closed box. The sense of smell coupled to liquid chromatography and even mass spectrometry is also discussed.

3.5 Training of integrative bioinformatics experts

Hannes Röst (University of Toronto, CA) and Andreas Hildebrandt

The training of integrative bioinformatics experts will be a key challenge for educators throughout the next decade. This challenge is complicated by the fact that the field is evolving rapidly, and that several types of undergraduate and/or Master's degrees could feed into such a programme. It is therefore likely that there will not be a single such curriculum, but rather a set of courses, from which a choice is made based on the pre-existing knowledge of the trainee. At the same time, the level of education on which this training is to take place is flexible. Basic programming courses in often-used languages such as Python, for instance, should preferentially begin at the undergraduate level at the latest (it would be far better to start much earlier, e.g., in secondary education), while training in advanced mathematical modelling is more likely to take place at the Master's or at the post-graduate level. Overall, online training courses could be a very interesting means to educate people, but care should be taken that the courses stay up-to-date. This is challenging in a fast evolving field, and will require substantial time investment.

3.6 Analysis and interpretation of public proteomics data in orthogonal contexts

Juan Antonio Vizcaino (EBI – Hinxton, GB)

The availability of public proteomics datasets continues to increase, and a plateau has clearly not been reached yet. Many possibilities for data reuse exist and some of these forms are increasingly popular. Two particularly rewarding but difficult scenarios for reuse of proteomics data are 're-analysis' and 're-purpose'. The difference between the two is subtle: in the case of the former, the analysis settings change compared with the original study, but the goals of the study do not. In the case of the latter, both analysis settings as well as goals can be different from the original. In the case of 're-analysis', examples are found in widely used resources, e.g. Peptide Atlas, MassIVE and ProteomicsDB. 'Re-purpose' examples can be found in proteogenomics studies and in the detection of new PTMs/sequence variants. Existing challenges for the reuse of data were highlighted as well. There is a lack of suitable annotation for many data sets, which prevents re-use to extract biological meaning. Moreover, there is a need for robust computational infrastructure to provide the required calculation power. A special mention is made of the difficulty in matching different datasets coming from different 'omics approaches (where the data is also spread across different 'omics specific data repositories). A slowly emerging issue that should be taken into account already, is access controlled data in the case of clinical samples. Of course, any re-analysis will run into challenges related to false discovery rate estimations in the context of large search spaces (for instance, when searching for single amino acid substitutions) and when false positives are combined across different data sets. At the same time, there are opportunities available that have not been covered so far. The re-analysis of atypical data sets, such as metaproteomics experiments or data independent acquisition (DIA) analyses provides an obvious example, but hinges on the availability of dedicated algorithms, and specialized resources such as spectral libraries. A more future-oriented goal is the integration of proteomics and metabolomics data sets to elucidate metabolite fluxes and influences on the proteome.

3.7 Integrated multi-omics for enhanced metaproteomics

Paul Wilmes (University of Luxembourg, LU)

Metaproteomics involves analysing the protein complement of microbial consortia. Peptide and protein identification is challenged by the inherent complexity of the samples. The generation of concomitant metagenomic and metatranscriptomic data allows the construction of sample-specific protein databases which facilitates enhanced data usage including for protein identification. Furthermore, exploitation of the de Bruijn metagenomic and metatranscriptomic assembly graphs allows the resolution of variant paths which in turn enables strain-level resolution of peptides and proteins. These approaches greatly enhance peptide and protein identification rates. Consequently, integrated multi-omic analyses of microbial communities overall result in much improved metaproteomic coverage.

4 Working groups

4.1 False Discovery Rates in Spectral Library Searching and Data Independent Acquisition Identification

Eric Deutsch (Institute for Systems Biology – Seattle, US), Robert Chalkley (UC – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Nico Jehmlich (UFZ – Leipzig, DE), Bernhard Küster (TU München, DE), Hannes Röst (University of Toronto, CA), Timo Sachsenberg (Universität Tübingen, DE), Stephen Tate (SCIEX – Concord, CA), Mathias Wilhelm (TU München, DE), and Paul Wilmes (University of Luxembourg, LU)

The breakout group on data indepenten acquisition (DIA) spectral library false discovery rate (FDR) in the Dagstuhl Seminar on Computational Proteomics, containing 11 participants, discussed the current issues in estimating and maintaining the reliability during generation of spectral libraries and in subsequent analyses. As far as the generation of spectral libraries is concerned, determining the FDR at the creation of the library is based on data dependent acquisition (DDA) FDR estimates. However, when it comes to extending the spectral libraries for new entries, maintenance and estimation of the FDR within the library is not really solved other than re-searching all the data again. The within library FDR should be propagated and considered in the search results to compensate for the reliability of the library itself.

Estimating the FDR on spectral library search results is a challenge itself, as decoy generation is not as easy as for database searches. Current methods seem to over- or underestimate the true FDR in the data sets. As an action item this breakout group aims to generate one (or more) gold standard spectral library data sets to evaluate current and future approaches for FDR estimation. This could also allow for checking whether decoy generation is the way to estimate FDR for spectral library searching. It was recognized that approaches being applied for DDA may not be valid for DIA. FDR calculation for DIA data is even more complicated than for DDA because of the increased complexity. We may have to come up with new/better solutions to estimate FDR on DIA searches.

The group also discussed a few additional related topics, including how post-translational modification (PTM) site localisation could be handled and how FDR estimation approaches from other fields could be adopted.

4.2 Spectral Libraries in Proteomics

Eric Deutsch (Institute for Systems Biology – Seattle, US), Nuno Bandeira (University of California – San Diego, US), Sebastian Böcker (Universität Jena, DE), Robert Chalkley (UC – San Francisco, US), John Cottrell (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Nico Jehmlich (UFZ – Leipzig, DE), Lukas Käll (KTH – Royal Institute of Technology, SE), Hannes Röst (University of Toronto, CA), Timo Sachsenberg (Universität Tübingen, DE), Stephen Tate (SCIEX – Concord, CA), Hans Vissers (Waters Corporation – Wilmslow, GB), Pieter-Jan Volders (Ghent University, BE), Mathias Walzer (EBI – Hinxton, GB), Ana L. Wang (Scripps Research Institute – La Jolla, US), Mathias Wilhelm (TU München, DE), and Dennis Wolan (Scripps Research Institute – La Jolla, US)

The eighteen participants of the Spectral Libraries Breakout Group of the 2017 Computational Proteomics Dagstuhl Seminar discussed the current state and future directions for the generation and use of peptide tandem mass spectrometry spectral libraries. Their use in proteomics is growing slowly, but there are multiple challenges in the field that must be addressed to further increase the adoption of spectral libraries and related techniques. This Spectral Libraries Breakout Group aims to generate and publish a set of recommendations for addressing these challenges, building on prior work of the Proteomics Standards Initiative (PSI).

The primary bottlenecks are the paucity of high quality and comprehensive libraries, and the general difficulty of adopting spectral library searching into existing workflows. There are several existing spectral library formats, but none of them capture a satisfactory level of metadata, and therefore a logical next advance is to design a more advanced, PSI-approved spectral library format that can encode all of the desired metadata.

The group discussed a series of metadata requirements, organized into three levels of completeness or quality, tentatively dubbed bronze, silver, and gold. The metadata would be encoded at the collection (library) level (e.g., methods details, such as whether library spectra are consensus or representative spectra), at the individual entry (peptide ion) level (e.g., FDR of identification used for inclusion in the library), and at the peak (fragment ion) level (e.g., intensity variance).

The group discussed strategies for encoding mass modifications in a consistent manner (there was agreement that the mzTab specification seems adequate) and the requirement for encoding high quality and commonly-seen but as-yet unidentified spectra. The exact style of the new standard format (e.g., enhancement of the currently most popular MSP format, XML, heavily optimized binary formats, database-based storage, etc.) remains the subject of vigorous debate.

The group also discussed a few additional related topics, including strategies for comparing two spectra, techniques for generating representative spectra for a library, approaches for selection of optimal signature ions for targeted workflows, and issues surrounding the merging of two or more libraries into one.

4.3 Assessment of proteome coverage

Gerben Menschaert (Ghent University, BE), Marco Hennrich (EMBL – Heidelberg, DE), Bernhard Küster (TU München, DE), Kathryn Lilley (University of Cambridge, GB), Frédérique Lisacek (Swiss Institute of Bioinformatics, CH), Lennart Martens (Ghent University, BE), Elien Vandermarliere (Ghent University, BE), Juan Antonio Vizcaino (EBI – Hinxton, GB), and Henrik Zauber (Max-Delbrück-Centrum – Berlin, DE)

The impact of various biological processes on the coverage of the complete proteoform space varies. The following topics are encountered when considered in order of importance. First is the occurrence of splice isoforms, which can be cell or tissue type specific. In order to study these, it will therefore be important to rely on matching data (e.g., from genomics, transcriptomics, and proteomics). It is also interesting to see that emerging technologies can potentially be used to improve our understanding of isoforms and their annotation (for instance, the nanopore technology).

Another, somewhat related topic is that of ORF delineation. Current approaches do exist, but these are still at a reasonably early stage in development: adding extra unannotated (re-annotated) open reading frames (ORFs) to the search space is probably the most mature. Moreover, these can be supplemented with (potential) alternative start sites. The impact of these additions on identification rate is rather limited, because the database size increase remains modest. At the same time, there is an entire family of potential open reading frames that are currently too small to be picked up by gene prediction algorithms, and these (upstream (uORFs) and/or small ORFs (sORFs)) should also be investigated. An important question is whether these are actually active at the protein level, and if these can thus be picked up by mass spectrometry.

A further expansion of the potential sequence space is conferred by single amino-acid variations, which can even be increased further in the context of disease. While potentially detectable with existing methods (although it will be challenging due to the dependency of detection probability on the abundance of the parent protein, and the ionization potential and possible modification status of the peptide in question), it remains challenging to include the frequency information for these variations. It should also be noted in this context that, while variation is included in databases such as the UniProt KnowledgeBase, frequencies are not included for these variations. When combined with ribosome profiling, the sequence space can be further expanded to include frameshifts as well as stop-codon read-through. Although it should be noted that the occurrence rate of these events may be quite low, and the biological relevance could be quite limited.

Finally, beyond the sequence space, the chemical space can be extended as well, through post-translational modifications (PTMs). Many of these occur frequently, and thus have a large impact on the total proteoform space. Moreover, mass spectrometry remains the primary means of exploring these PTMs, but is in turn hindered by a lack of knowledge on the underlying biological processes and mechanisms that carry out and regulate these modifications, which makes it hard to predict what we could possibly expect. Throughout, a question that remains essentially unanswered, is how to derive biological meaning from results that are obtained from an extension of our coverage.

4.4 False Discovery Rate Estimation Issues in Large Database Searches and Proposal of Benchmarking Challenges

Thilo Muth (Robert Koch Institut – Berlin, DE), Magnus Arntzen (Norwegian University of Life Sciences – As, NO), Sebastian Böcker (Universität Jena, DE), John Cottrell (Matrix Science Ltd. – London, GB), Julien Gagneur (TU München, DE), Laurent Gatto (University of Cambridge, GB), Marco Hennrich (EMBL – Heidelberg, DE), Lukas Käll (KTH – Royal Institute of Technology, SE), Jeroen Krijgsveld (DKFZ – Heidelberg, DE), Phillip Pope (Norwegian University of Life Sciences – As, NO), Hans Vissers (Waters Corporation – Wilmslow, GB), Ana L. Wang (Scripps Research Institute – La Jolla, US), Dennis Wolan (Scripps Research Institute – La Jolla, US), and Henrik Zauber (Max-Delbrück-Centrum – Berlin, DE)

We first discussed issues of false discovery rate (FDR) estimation for large databases with an emphasis on metaproteomics. Different levels of FDR were recognized, such as peptide, protein, isoform and species FDR. There were concerns regarding the FDR control when using multiple search engine because of different scoring schemes. It was also discussed that the size and completeness of the search space (i.e. spectra, sequences and post-translational modifications (PTMs)) has an influence of FDR at all levels for target-decoy searches. One possibility is to reduce the search space, e.g. by limiting the database to the peptides which can be expected by using custom (metagenome/metatranscriptome) databases. While PSM and peptide FDR have been evaluated thoroughly so far, still no clear consensus can be found on how to assess the protein FDR. Secondly, we proposed to initiate two benchmarking challenges which are open for the community, one for metaproteomics and another for splicing isoforms. For assessing splicing isoforms, two different cell types will be grown and mixture series will be generated. One expects the quantification of isoforms to be proportional to the mixture ratio which allows for benchmarking their linear relationships, e.g. using R-squared of isoform quantity estimates vs. dilution ratio. Participants are requested to estimate isoform quantities for each sample individually. Moreover, transcriptome data are generated for each cell type and methods for MS-based isoform quantification are benchmarked using state-of-the-art RNA-sequencing isoform quantities as ground truth estimates. The metaproteomics challenge consists of three different options, (i) creating a metaproteome of a mock community of known isolates for evaluating peptide/protein/species FDR, (ii) providing a mock communities of different dilution series of variants (rare vs. abundant species) allowing also to assess fold change estimation, (iii) spiking the mock into a complex background community (e.g. with closely related species) to assess the recovery. Sample spectra and database consisting of the whole complex (real + mock) community will be provided to the participant. For the metaproteomics challenge, the connection with CAMI challenge for metagenomics (version 2) will be coordinated. The splicing isoform challenge may be linked to either ABRF (http://www.cosmosid.com/nist-challenge/) or DREAM http://dreamchallenges.org/) challenges.

4.5 Visualization of proteomics and multi-omics data

Magnus Palmblad (Leiden University Medical Center, NL), Magnus Arntzen (Norwegian University of Life Sciences – As, NO), Harald Barsnes (University of Bergen, NO), Ileana M. Cristea (Princeton University, US), Laurent Gatto (University of Cambridge, GB), Lydie Lane (Swiss Institute of Bioinformatics, CH), Bart Mesuere (Ghent University, BE), Thilo Muth (Robert Koch Institut – Berlin, DE), Phillip Pope (Norwegian University of Life Sciences – As, NO), Veit Schwämmle (University of Southern Denmark – Odense, DK), and Olga Vitek (Northeastern University – Boston, US)

The old saying "a picture is worth a thousand words" probably understates the necessity for appropriate visualization tools in data intensive sciences such as genomics or proteomics. In the breakout session, we contrasted interactive visualizations to explore data with reproducible generation of figures for reports or publications. We discussed the importance of mindful visualization – what is the question to be addressed, is the data available, what kind of transformations are required, and what software should be used? We covered these questions in the contexts of six use cases: (1) influence of PTMs on PPI networks, (2) alignment and visualization of unidentified features across datasets, (3) integrating spatially resolved quantitative omics data, (4) flux analysis integrating time-resolved omics data, (5) metaproteomics with taxonomies down to the strain level, and (6) Mapping PTM crosstalk and proteoforms to structures.

Network visualizations were found to address questions in all use cases. Careful attention should be paid to data representation, including using controlled vocabularies and ontologies for metadata used for the visualizations. Distinction was also made between visualizing many entities (proteins or metabolites) in one experiment versus showing the distribution of few entities across many datasets.

Though many powerful visualization software platforms exist, there is a need for refined tools for displaying PTMs or proteoform information in the context of PPI networks or pathways (use cases 1 and 4), systematic metadata annotation using controlled vocabularies (use cases 3, 4 and 5), and integrating alignment of unidentified LC-MS(/MS) features with study metadata. Network and pathway visualization tools must clearly distinguish between absolute and relative changes in abundance (all use cases) and between no change with no data. Potential pitfalls were also discussed, such as adding information lacking experimental evidence in visualizations and attempting to display too much information in one figure. Sometimes, visualization is more about what to hide than what to show.

4.6 Metadata Provision for Public Proteomics Data

Juan Antonio Vizcaino (EBI – Hinxton, GB), Lydie Lane (Swiss Institute of Bioinformatics, CH), Frédérique Lisacek (Swiss Institute of Bioinformatics, CH), Lennart Martens (Ghent University, BE), Gerben Menschaert (Ghent University, BE), Veit Schwämmle (University of Southern Denmark – Odense, DK), and Mathias Walzer (EBI – Hinxton, GB)

The value of public data increases with reuse, but such reuse requires proper metadata annotation. Unfortunately, metadata is currently only sparsely available, and mostly remains unstructured. The way to resolve this issue, and thus to add value to public data, revolves around two complementary strategies. The first strategy is to recover the already submitted meta data through post-hoc annotation; this can be achieved by structuring currently unstructured data, or by extracting metadata from in-depth analysis of the data proper. The second startegy is to increase the annotation of submitted proteomics data by the submitter. Importantly, formats already exist that allow metadata to be be structured, and that covers a variety of metadata pertaining to more or less the complete analytical workflow in proteomics. The working group therefore looked into existing solutions and readily available metadata, and reviewed these with respect to tangible applications to put metadata into a structured format.

At the same time, however, data repositories should endeavour to make the added value of metadata availability more visible, and to lower the threshold for entering metadata annotation. This will not only motivate submitters to provide these metadata, but will also enable the efficient annotation of these data.

The overall conclusion was a need for specific tools to annotate metadata, both at the site of the experimentalist, and preferrably in such a way that the user-specified metadata (as opposed to instrument-derived metadata, which tends to be captured more comprehensively and transparently already) is captured even before the project starts. In addition, efforts by ‘annotation super users’ (researchers who already actively reuse public proteomics data on a large scale) that connect existing data with metadata should be captured for subsequent general reuse.

4.7 Computational Proteomics Education

Pieter-Jan Volders (Ghent University, BE), Harald Barsnes (University of Bergen, NO), Lennart Martens (Ghent University, BE), Bart Mesuere (Ghent University, BE), Magnus Palmblad (Leiden University Medical Center, NL), and Elien Vandermarliere (Ghent University, BE)

Computational proteomics, and bioinformatics in general, attracts people with different backgrounds such as bio(medical) and computer sciences. We discussed the required skillset and different profiles of people working in the field distinguishing computational scientists, bioinformaticians and biologists. The key aspect is computational thinking. Moreover, bioinformaticians, and scientists in general, need to be taught enough of the neighbouring fields to communicate efficiently with everyone involved in a research project. This is partly reflected in the observation that the boundaries between being a biologist, a bioinformatician and a computer scientist are becoming increasingly vague. Next, we focused on training. Training someone in computational proteomics requires knowledge from (molecular) biology, statistics, computer science, mass spectrometry and general bioinformatics. We thus propose a curriculum with required skills and knowledge from those fields. We compiled a set of guidelines and a repository of online resources for both students and educators that can serve as a basis for designing educational programs in computational proteomics.

Participants

- Magnus Arntzen
Norwegian University of Life Sciences – As, NO
- Nuno Bandeira
University of California – San Diego, US
- Harald Barsnes
University of Bergen, NO
- Sebastian Böcker
Universität Jena, DE
- Robert Chalkley
UC – San Francisco, US
- John Cottrell
Matrix Science Ltd. – London, GB
- Ileana M. Cristea
Princeton University, US
- Bernard Delanghe
Thermo Fisher GmbH – Bremen, DE
- Eric Deutsch
Institute for Systems Biology – Seattle, US
- Viktoria Dorfer
University of Applied Sciences Upper Austria, AT
- Julien Gagneur
TU München, DE
- Laurent Gatto
University of Cambridge, GB
- Marco Hennrich
EMBL – Heidelberg, DE
- Nico Jehmlich
UFZ – Leipzig, DE
- Lukas Käll
KTH – Royal Institute of Technology, SE
- Oliver Kohlbacher
Universität Tübingen, DE
- Jeroen Krijgsveld
DKFZ – Heidelberg, DE
- Bernhard Küster
TU München, DE
- Lydie Lane
Swiss Institute of Bioinformatics – Genève, CH
- Kathryn Lilley
University of Cambridge, GB
- Frédérique Lisacek
Swiss Institute of Bioinformatics – Genève, CH
- Lennart Martens
Ghent University, BE
- Gerben Menschaert
Ghent University, BE
- Bart Mesuere
Ghent University, BE
- Thilo Muth
Robert Koch Institut – Berlin, DE
- Magnus Palmblad
Leiden University Medical Center, NL
- Phillip Pope
Norwegian University of Life Sciences – As, NO
- Hannes Röst
University of Toronto, CA
- Timo Sachsenberg
Universität Tübingen, DE
- Veit Schwämmle
University of Southern Denmark – Odense, DK
- Stephen Tate
SCIEX – Concord, CA
- Elien Vandermarliere
Ghent University, BE
- Hans Vissers
Waters Corporation – Wilmslow, GB
- Olga Vitek
Northeastern University – Boston, US
- Juan Antonio Vizcaino
EBI – Hinxton, GB
- Pieter-Jan Volders
Ghent University, BE
- Mathias Walzer
EBI – Hinxton, GB
- Ana L. Wang
Scripps Research Institute – La Jolla, US
- Mathias Wilhelm
TU München, DE
- Paul Wilmes
University of Luxembourg, LU
- Dennis Wolan
Scripps Research Institute – La Jolla, US
- Henrik Zauber
Max-Delbrück-Centrum – Berlin, DE

Report from Dagstuhl Seminar 17431

Performance Portability in Extreme Scale Computing: Metrics, Challenges, Solutions

Edited by

Anshu Dubey[1], Paul H. J. Kelly[2], Bernd Mohr[3], and Jeffrey S. Vetter[4]

1 **Argonne National Laboratory, US,** adubey@anl.gov
2 **Imperial College London, GB,** p.kelly@imperial.ac.uk
3 **Jülich Supercomputing Centre, DE,** b.mohr@fz-juelich.de
4 **Oak Ridge National Laboratory, US,** vetter@ornl.gov

Abstract

This Dagstuhl Seminar represented a unique opportunity to bring together international experts from the three research communities essential to tackling the HPC performance portability challenge: developers of large-scale computational science software projects, researchers developing parallel programming technologies, and performance specialists. The major research questions for the seminar were to understand challenges, design metrics, and prioritize potential solutions for performance portability, management of data movement in complex applications, composability, and pathways to impact on the research community. The overall conclusion shared by all participants was that performance portability in extreme scale computing can be achieved, especially if parallel applications are designed with performance portability in mind from the beginning. Making legacy application performance portable still requires enormous efforts and expertise. In many instances it will likely require extensive refactoring.

Seminar October 22–27, 2017 – http://www.dagstuhl.de/17431
1998 ACM Subject Classification Parallel architectures, Parallel programming languages, Parallel algorithms, Performance,
Keywords and phrases Parallel programming, performance portability, productivity, scientific computing
Digital Object Identifier 10.4230/DagRep.7.10.84

1 Executive Summary

Anshu Dubey
Paul H. J. Kelly
Bernd Mohr
Jeffrey S. Vetter

This report documents the program and the outcomes of Dagstuhl Seminar 17431 "Performance Portability in Extreme Scale Computing: Metrics, Challenges, Solutions".

Performance Portability is a critical new challenge in extreme-scale computing. In essence, performance-portable applications can be efficiently executed on a wide variety of HPC architectures without significant manual modifications. For nearly two decades, HPC architectures and programming models remained relatively stable, which allowed growth of

Performance Portability in Extreme Scale Computing: Metrics, Challenges, Solutions, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 84–110
Editors: Anshu Dubey, Paul H. J. Kelly, Bernd Mohr, and Jeffrey S. Vetter
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

complex multidisciplinary applications whose lifecycles span multiple generations of HPC platforms.

Recently, however, platforms are growing much more complex, diverse, and heterogeneous - both within a single system and across systems and generations. Details already known from planned future systems indicate that this trend will continue (at least for the foreseeable future). Current and planned future large-scale HPC systems consist of complex configurations with a massive number of components. Each node has multiple multi-core sockets and often one or more additional accelerator units in the form of many-core nodes or GPGPUs, resulting in a heterogeneous system architecture. Memory hierarchies including caches, memory, and storage are also diversifying in order to meet multiple constraints: power, latency, bandwidth, persistence, reliability, and capacity. These factors are reducing portability, and forcing applications teams to either spend considerable effort porting and optimizing their applications for each specific platform, or risk owning applications that perform well on perhaps only one architecture. The latter option would still require porting and optimizing effort for each new generation of systems.

This Dagstuhl Seminar represented a unique opportunity to bring together international experts from the three research communities essential to tackling this performance portability challenge: developers of large-scale computational science software projects whose lifetime will span multiple generations of systems, researchers developing relevant parallel programming or system software technologies, and specialists in profiling, understanding, and modelling performance. The major research questions for the seminar were:

- To understand challenges, design metrics, and prioritize potential solutions for performance portability: Solutions will need to synthesize existing concepts across multiple fields, including performance and productivity modeling, programming models and compilation, architectures, system software.
- Management of data movement in complex applications: Diverse data movement patterns dictated by different devices form one of the largest impediments to portable performance. Addressing it will require cross-cutting solutions supporting more than one abstraction, and will allow scientists to balance tradeoffs in these factors prior to design, development, or procurement of an architecture, software stack, or application.
- Composability: Many applications require flexibility and composability because they address different physical regimes either within the same simulation, or in different instances of simulations.
- Pathways to impact on the research community: As the community becomes more reliant on both more complex architectures and software stacks, it is especially important that we develop the conceptual tools to facilitate research and practical solutions for performance portability. The impact of ignoring this topic could be potentially devastating to the quality and sustainability of computational science software, and consequently on the science and engineering research they support. Thus a key element of the seminar will be to tackle this challenge in major science community software projects.

The seminar started with a series of flash talks, where participants introduced themselves in a two-minute one-slide presentation summarizing their contribution or interest in the seminar by providing two to three bullet points on (i) Challenge/Opportunity (WHY?) (ii) Timeliness (WHY NOW?) (iii) Approaches (HOW?) and (iv) IMPACT (SO WHAT?). Each day started with a longer keynote presentation by a representative of one of the major stakeholders in the field, followed by short presentations by participants grouped in sessions with a common relevant theme. Each keynote or short talk session ended with an extensive question-and-answer session and open discussion slot in which all the speakers from the session took part.

The overall conclusion shared by all participants was that performance portability in extreme scale computing can be achieved, especially if parallel applications are designed with performance portability in mind from the beginning. Model complexity and performance portability both require that frameworks be designed with composable components incorporating layers of abstraction so that trade-offs can be reasoned about. Making legacy application performance portable still requires enormous efforts and expertise. In many instances it will likely require extensive refactoring. Similar design principles regarding formulation of a flexible and composable framework apply for legacy software refactoring, along with strong emphasis on rigorous verification built into the process. The seminar recognized the challenges faced by the applications in adopting abstractions; converting research prototypes to reliable production-grade product. The adverse structure of incentives for both applications and abstractions, and the complexity of formulating a process or collaboration between the two communities, may be bigger barriers than technical challenges in making performance portability feasible. It is critical that the involved communities and stakeholders are made aware of these challenges while seeking solutions for sustainable computational science projects.

2 Table of Contents

Executive Summary

Overview of Talks

Open problems

3 Overview of Talks

3.1 Automatic Detection of Large Extended Data-Race-Free Regions with Conflict Isolation

Alexandra Jimborean (Uppsala University, SE)

Joint work of Alexandra Jimborean, Jonatan Waern, Per Ekemark, Stefanos Kaxiras, Alberto Ros
Main reference Alexandra Jimborean, Per Ekemark, Jonatan Waern, Stefanos Kaxiras, Alberto Ros: "Automatic Detection of Large Extended Data-Race-Free Regions with Conflict Isolation", IEEE Trans. Parallel Distrib. Syst., Vol. 29(3), pp. 527–541, 2018.
URL http://dx.doi.org/10.1109/TPDS.2017.2771509

Data-race-free (DRF) parallel programming becomes a standard as newly adopted memory models of mainstream programming languages such as C++ or Java impose data-race-freedom as a requirement. We propose compiler techniques that automatically delineate extended data-race-free (xDRF) regions, namely regions of code that provide the same guarantees as the synchronization-free regions (in the context of DRF codes). xDRF regions stretch across synchronization boundaries, function calls and loop back-edges and preserve the data-race-free semantics, thus increasing the optimization opportunities exposed to the compiler and to the underlying architecture. We further enlarge xDRF regions with a conflict isolation (CI) technique, delineating what we call xDRF-CI regions while preserving the same properties as xDRF regions. Our compiler (1) precisely analyzes the threads' memory accessing behavior and data sharing in shared-memory, general-purpose parallel applications, (2) isolates data-sharing and (3) marks the limits of xDRF-CI code regions. The contribution of this work consists in a simple but effective method to alleviate the drawbacks of the compiler's conservative nature in order to be competitive with (and even surpass) an expert in delineating xDRF regions manually. We evaluate the potential of our technique by employing xDRF and xDRF-CI region classification in a state-of-the-art, dual-mode cache coherence protocol. We show that xDRF regions reduce the coherence bookkeeping and enable optimizations for performance (6.4 percent) and energy efficiency (12.2 percent) compared to a standard directory-based coherence protocol. Enhancing the xDRF analysis with the conflict isolation technique improves performance by 7.1 percent and energy efficiency by 15.9 percent.

3.2 Is it performance portability when I'm using DGEMM?

Michael Bader (TU München, DE)

The earthquake simulation software SeisSol uses a high-order discontinuous Galerkin scheme for discretisation of the seismic wave equations. The scheme is formulated via small element-local matrix chain products; respective matrices include the matrix of quantities, stiffness matrices, discrete Jacobians, etc. As matrices can be sparse or dense, the matrix chain products are analysed for sparsity patterns of matrices, including intermediate and result matrices. The LIBXSMM library is used to generated high-performance code on Intel architectures.

While the matrix notation provides a suitable abstraction layer for expressing the numerical scheme, only a limited level of performance portability is reached by this approach. The presentation's goal was to discuss the current status and possible routes for extension and improvement.

References

1 A. Heinecke, G. Henry, M. Hutchinson and H. Pabst: LIBXSMM: Accelerating Small Matrix Multiplications by Runtime Code Generation, SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. DOI: 10.1109/SC.2016.83

2 C. Uphoff and M. Bader: Generating high performance matrix kernels for earthquake simulations with viscoelastic attenuation. In W. W. Smari (ed.), Proceedings of the 2016 International Conference on High Performance Computing & Simulation (HPCS 2016), p. 908–916. IEEE, 2016.

3 C. Uphoff, S. Rettenberger, M. Bader, E. H. Madden, T. Ulrich, S. Wollherr and A.-A. Gabriel: Extreme Scale Multi-Physics Simulations of the Tsunamigenic 2004 Sumatra Megathrust Earthquake. SC '17: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2017

3.3 Performance Portability: Discussion Entry Points from Experience in OpenMP

Carlo Bertolli (IBM TJ Watson Research Center – Yorktown Heights, US)

In this talk I will provide a set of entry points for discussion related to performance portability challenges in the implementation of OpenMP on GPU-enabled systems. I will give an overview of how some of these challenges are being addressed by the OpenMP committee and what the main hard problems with proposed solutions are. I will also show how the performance portability issue spans beyond the traditional HPC community and is present in emerging fields such as cognitive computing and related hardware acceleration technology.

3.4 Portability of Performance in Generic Code

Mauro Bianco (CSCS – Lugano, CH)

Joint work of Mauro Bianco, Oliver Fuhrer, Carlos Osuna, Thomas Schulthess, Hannes Vogt

Generic Libraries can offer separation of concerns between program developers and performance specialists by providing abstractions for algorithmic motifs, so as to hide low-level details. As computing platforms diversify, the design of these libraries becomes more and more complex and more and more targeted to specific application fields. There are however some deep problems that make portability of performance extremely challenging in real-world applications. The effect is the tendency to limit the applicability of particular generic interfaces to sub-sets of problems in the targeted application fields, thus making them less and less general. By analyzing some of the characteristics of weather and climate applications, this talk will highlight some challenges in the development of a truly performance-portable layer for this class of applications. Some of these challenges raise more general questions about performance in HPC. The talk focuses on the engineering challenges, such as development complexity, maintainability and interoperability with other languages. The result of the development is a set of libraries for production applications for weather and climate simulations to be used by different institutions.

3.5 Automatic Empirical Performance Modeling

Alexandru Calotoiu (TU Darmstadt, DE)

Given the tremendous cost of an exascale system, its architecture must match the requirements of the applications it is supposed to run as precisely as possible. Conversely, applications must be designed such that building an appropriate system becomes feasible, motivating the idea of co-design. In this process, a fundamental aspect of the application requirements are the rates at which the demand for different resources grows as a code is scaled to a larger machine. However, if the anticipated scale exceeds the size of available platforms this demand can no longer be measured. This is clearly the case when designing an exascale system. Moreover, creating analytical models to predict these requirements is often too laborious - especially when the number and complexity of target applications is high. We discuss how automated performance modeling can be used to quickly predict application requirements for varying scales and problem sizes. Following this approach, we show how to determine the exascale requirements of a code and use them to illustrate system design tradeoffs.

3.6 If the HPC Community were to create a truly productive language...[how] would we ever know?

Bradford Chamberlain (Cray Inc. – Seattle, US)

In this talk, I strived to explore the relationship between productivity and performance portability, particularly in the context of programming languages for High Performance Computing. I began with a quick argument that we don't really have productive languages in HPC, and went on to characterize currently used HPC "languages" (broadly interpreted) and what I imagine from productive languages. From there, I gave a brief, personal historical perspective on the effort to define metrics for measuring productivity under the DARPA HPCS program (2002-2012). I wrapped up with an argument that I believe productivity is a highly personal, social choice and that, for that reason, our evaluations of productivity should be more personal and social, supporting subjective decisions rather than analytic. I introduced the (mainstream) Computer Language Benchmarks Game as an instance of measuring properties of languages through programs written in those languages, and one supporting interactions with the results and programs in various ways. This led to my first suggestion which is that the HPC community would benefit from a similar arena. From there, I turned to a brief description of Chapel, highlighting key policies that we have opted not to bake into the language in order to support productivity, a separation of concerns, and performance portability: user-defined parallel iterators, domain maps for implementing dense/sparse/associative domains & arrays, and locale models for representing the target architecture. I then posed the question as to how languages like Chapel can advance to the stage of being used "in HPC production," suggesting that opportunities (like a Dagstuhl seminar?) should be created for applications and languages people to pair program, cutting past superficial familiarity with each others' technologies.

3.7 Should I port my code to a DSL?

Aparna Chandramowlishwaran (University of California – Irvine, US)

Joint work of Aparna Chandramowlishwaran, Laleh Aghababaie Beni, Bahareh Mostafazadeh, Ferran Marti
Main reference Laleh Aghababaie Beni, Aparna Chandramowlishwaran: "PASCAL: A Parallel Algorithmic SCALable Framework for N-body Problems", in Proc. of the Euro-Par 2017: Parallel Processing - 23rd International Conference on Parallel and Distributed Computing, Santiago de Compostela, Spain, August 28 - September 1, 2017, Proceedings, Lecture Notes in Computer Science, Vol. 10417, pp. 482–496, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-64203-1_35

In this talk, I'll outline the key challenges in developing parallel algorithms and software for two classes of applications – N-body solvers and structured grid finite volume solvers on current and future platforms. Our goal is to reduce the apparent gap in performance between code generated from high-level forms and that of hand-tuned code, which we address using extensive characterization of the optimization space for these computations and automating the process through domain-specific languages (DSLs) and code generators. These application-specific compilers provide the domain scientists the ability to productively harness the power of these large machines and to enable large-scale scientific simulations and big data applications. However, the performance of DSLs has been a concern. Therefore, we specifically ask whether CFD and N-body applications expressed in these DSLs can deliver a sufficient combination of optimizations to compete with a hand-tuned code and what are its limitations.

3.8 Exascale Scientific Computing : The Road Ahead

Kemal A. Delic (Hewlett Packard – Grenoble, FR) and Martin A. Walker

Theory and experimentation have been the hallmark of scientific inquiries for centuries, while we observe the rising importance of simulation as the third, important pillar of scientific explorations. The key driving factor is slowly evolving field of hyperscale computing, recently augmented and accelerated through the confluence of artificial intelligence(AI), big data (BD) and the internet of things (IoT), making it again a leading force in scientific and industrial computing.

For the new scientific discoveries to happen, larger scale as well as different computing infrastructure is required to enable and support bigger data collections and more efficient/effective algorithms.

The rise of cloud computing,as a hyperscale computing infrastructure, has provided a new arena for the execution of scientific workloads. Such an infrastructure has also enabled great strides in the application of multilayered weighted networks often known as "neural networks" for scientific enquiry. The universal approximation theorem, implies that such networks can be constructed to solve partial differential equations, and therefore applied to computing the consequences of scientific theories.

Construction of such networks ("training", i.e. the determination of the number of nodes and topology of the network, together with the values of the weights on each edge, and the choice of trigger function at the nodes) requires extensive computing with large data volumes on very large infrastructure until a sufficiently accurate model has been achieved. The resulting model can then be executed ("inference"), either on general purpose or specialized hardware accelerators. We believe that critical applications such as high-frequency trading (HFT), self-driving cars and various drones, as wells as edge network devices will require both: special hybrid architectures combining general purpose CPUs and specialized GPU, TPU, FPGA for training, and special chips for on-board execution.

In conclusion, looking bravely into next 50 years, we see High Performance Computing being augmented and advanced with AI, BD and IoT, evolving into Neuromorphic Computing within the next 15 years and with Quantum Computing on a 50-year horizon.

This is an exciting opportunity for emerging generations of scientists advancing scientific knowledge using new types of scientific instruments based on advances in computing sciences and clever engineering, producing unprecedented exascale scientific machines. It is our high hope that they will augment and accelerate the pace of scientific inquiries. Emerging Exascale Computing Systems may enable advances similar to the scientific advances realized by the invention of both the microscope and telescope a few centuries ago.

3.9 Beyond algorithmic patterns: tackling the performance portability challenge with hardware paradigm optimisation patterns

Christophe Dubach (University of Edinburgh, GB)

Algorithmic patterns have emerged as a solution to exploit parallel hardware. Applications are expressed at a high-level, using a small set of well known patterns of code adapted to each application domain. This approach hides the hardware complexity away from programmers and shifts the responsibility for extracting performance to the library writer or compiler. However, producing efficient implementations remains a complicated task that needs to be repeated for each newly introduced high-level pattern or whenever hardware changes.

The first part of the presentation will show how typical optimisations performed by GPU programmers are expressible as optimisation patterns and how an optimisation space can be defined in terms of provably correct rewrite-rules. Our initial results show that this approach leads to performance portability on several classes of GPUs.

The second part of the talk will focus on showing how optimisation patterns could be generalised to a larger classes of hardware including CPUs, GPUs, FPGAs and multi-node clusters. The main idea is to develop an abstraction to reasons about hardware paradigms (e.g. memory hierarchy, parallelism hierarchy, synchronisation primitives, communication primitives) and a set of corresponding hardware-specific optimisation patterns. This would enable the automatic generation of an optimising code generator tailored for a particular instance of a heterogeneous parallel machine (e.g. cluster of GPUs or FPGAs) using a high-level description of the machine.

3.10 Kokkos: Performance Portability and Productivity for C++ Applications

H. Carter Edwards (Sandia National Labs – Albuquerque, US)

Joint work of H. Carter Edwards, Christian Trott, Daniel Sunderland, Daniel Ibanez, Nathan Ellingwood
Main reference H. Carter Edwards, Christian R. Trott, Daniel Sunderland: "Kokkos: Enabling manycore performance portability through polymorphic memory access patterns", J. Parallel Distrib. Comput., Vol. 74(12), pp. 3202–3216, 2014.
URL http://dx.doi.org/10.1016/j.jpdc.2014.07.003

Summary presentation of the Kokkos programming model and C++ library implementation for performance portability and productivity of intra-node parallel computations across diverse multicore and manycore architectures. This summary includes the "1+epsilon" versions of application code for performance portability, and the programming model abstractions enabling Kokkos to achieve this goal. An overview is given for Kokkos' data parallel and directed acyclic graph of tasks (task-dag) patterns to illustrate how Kokkos enables application development productivity for intra-node parallel algorithms.

3.11 How to define upper performance bounds using analytic performance models – Opportunities and Limitations

Jan Eitzinger (Universität Erlangen-Nürnberg, DE) and Georg Hager

Main reference Georg Hager, Jan Treibig, Johannes Habich, Gerhard Wellein: "Exploring performance and power properties of modern multi-core chips via simple machine models", Concurrency and Computation: Practice and Experience, Vol. 28(2), pp. 189–210, 2016.
URL http://dx.doi.org/10.1002/cpe.3180

Talking about application performance on computer systems ideally requires to define an analytic upper performance bound. Performance is defined by how a specific software interacts with the machine, which is the processor, its memory hierarchy and external IO devices as persistent storage and network. Many publications judge about 'good' or 'bad' performance in the light of comparisons to other implementations or hardware platforms. Still the insight created by such statements is usually very small. Automatic machine learning approaches fulfil certain purposes for predicting performance or are used to extrapolate measurements. But are those methods generating a deeper understanding about bottlenecks and optimisation opportunities? This talks tries to ignite a discussion about if and how (preferably simple) analytic performance modelling can help to make sense of an observed performance number and where its limitations are.

3.12 Deploying performance portable code

Todd Gamblin (LLNL – Livermore, US)

URL https://computation.llnl.gov/projects/spack-hpc-package-manager

Performance portability is a hot research topic, but are code developers really striving for it? HPC application teams are mainly tasked with producing good science, and ensuring performance portability takes a lot of time. Application teams fighting to make deadlines do some basic optimizations, but the risk/reward ratio for more sophisticated techniques is still far too high. Current performance portability techniques require significant effort on the part of the installer or the application developer. Most people installing HPC software are still building from source, from scratch, with compilers the code developer may or may not have tested with. The build is hard enough, even without considering a potential tuning phase. Package managers provide a natural harness around source builds of HPC code. However, most package managers in wide distribution sacrifice performance for potability and assume lowest-common-denominator flags.

In this talk, we discuss Spack, a package manager for HPC, and how the Spack community is looking to address performance portability at the software distribution level. We look at how performance portability can affect software source and binary distributions, and what type of additional infrastructure and tooling we would need to distribute tuned, optimized software for all HPC users. Our hope is to one day democratize performance portability.

3.13 Porting Atmosphere Kernels on Various System

Lin Gan (Tsinghua University – Beijing, CN)

A summary of previous work on porting atmosphere kernels on different platform, including the Sunway TaihuLight. Performance portability of atmosphere code is bad, so efforts have to be made and patience is required. For Sunway system, different software is being developed to make it easy for application to be ported.

3.14 Performance, Portability, and Dreams

William D. Gropp (University of Illinois – Urbana-Champaign, US)

URL http://wgropp.cs.illinois.edu/bib/talks/tdata/2017/dagstuhl-keynote.pdf

Why do we care about performance portability? A major reason is because a big part of the programming crisis is caused by the challenge of obtaining performance on even a single platform. And achieving performance is hard - systems are complex, behavior has random elements, and the behavior of interactions between parts is hard to predict. And after more than 20 years of relative architectural stability, processor architecture is diversifying, making the problem even worse.

But performance portability itself is not an absolute goal. Implicitly, performance portability is intended to reduce the amount of work needed to achieve adequate performance. How much programmer rework is acceptable to achieve performance portability? What other limitations, such as code complexity or sensitivity to input data, are acceptable? And there are dangers to making performance portability the goal. For example, one way to achieve performance portability is to make all performance mediocre. Then performance is similar on all platforms, but nowhere good. A good definition for performance portability is clearly necessary, but a workable definition is quite difficult. Most current definitions are either very difficult to apply (because they refer to a hard-to-determine theoretical achievable performance) or are susceptible to odd effects (when based on some specific code and that code's performance; leading, for example, to the case where any single code is performance portable by definition until a second code is created).

There are many different approaches to performance portability. These include enhancements to existing languages, new programming languages, libraries, tools, and even general techniques. The presentation provides a few examples that show that even for seemingly simple examples, performance is difficult to achieve without exploiting information known only at runtime. This suggests that approaches to performance portability need to include ways to adapt, perhaps at runtime, to different input data and different system behavior.

We discuss the Illinois Coding Environment (ICE), an example of an approach that uses annotations to an existing language to provide additional information that an guide performance optimizations, and uses a framework that can invoke third-party tools to apply performance enhancing transformations.

All approaches that rely on transformations to the user's code must address the issue of correctness - ensuring that any transformations do not introduce errors into the code. We point out that it is necessary to prove such transformations are correct, but that is not sufficient, because correctness requires that the entire system (including low-level software and all hardware) also correctly execute the transformed code.

The presentation ends by arguing that rather than try to define what performance portability is, the community should focus on the goals - making it easier for end users to run an application code effectively on different systems, and making it easier for developers to write, tune, and maintain an application for multiple systems.

3.15 Performance Portability Using Compiler-Directed Autotuning

Mary W. Hall (University of Utah – Salt Lake City, US)

As current and future architectures become increasingly diverse, the challenges of developing high-performance applications are becoming more onerous. The goal of compiler optimization in high-performance computing is to take as input a computation that is architecture independent and maintainable and produce as output efficient implementations of the computation that are specialized for the target architecture. A compiler that is specialized for an application domain can tailor its optimization strategy to increase effectiveness. This talk describes how domain-specific optimizations can be combined with standard polyhedral compiler transformation and code generation technology to achieve very high levels of performance, comparable to what is obtained manually by experts. Polyhedral frameworks permit composition of complex transformation sequences with robust code generation. Autotuning empirically evaluates a search space of possible implementations of a computation to identify the implementation that best meets its optimization criteria (e.g., performance, power, or both). Combining the three concepts, autotuning compilers generate this search space of highly-tuned implementations either automatically or with programmer guidance. We describe the application of this approach to three domains: geometric multigrid and the stencil computations within them, tensor contractions and sparse matrix computations.

3.16 User Interfaces to Performance: Kernel Transformation with Loopy

Andreas Klöckner (University of Illinois – Urbana-Champaign, US) and Matt Wala

Focusing on the optimization of computational kernels, Loopy is a transformation-based programming system embedded in Python that aims to assist with all stages of the performance engineering process from within a single, user-exposed intermediate representation (IR). This single IR is based on scalar assignemnt and polyhedrally-specified control flow. It is designed to easily capture code at many levels of abstraction, ranging from high-level, mathematical formulas to machine concerns such as vectorization, memory access patterns, parallelization, ILP, and many more. A large cross-section of tuning concerns are amenable to modification by transformations acting on the IR. The IR is capable of capturing data-dependent control flow, global barrier synchronization (compiled to multiple GPU kernels for compatibility), reductions, and prefix sums. In addition, the IR lends itself to counting and modeling computational expense, enabling manual and automated tuning. Multiple targets allow code generation for CPUs, Intel Knights machines as well as GPUs. A live demonstration of these capabilities will be part of this presentation.

Applications for which Loopy has been demonstrated to achieve good performance across architectures include high-order finite elements, chemical kinetics, and numerical linear algebra.

3.17 CnC for future-proof application development

Kathleen Knobe (Rice University – Houston, US)

The history of computing shows us that we have limited ability to predict important architectural features or important optimization concerns too far in the future. The applications that might have to be ported due to these changes are growing rapidly in number, size and complexity. The cost of re-implementing an app or a suite of apps for each new architectural advance or each new optimization goal is exorbitant.

Current approaches that address a range of foreseeable architectures and concerns are very helpful. CnC is, instead, an approach that addresses the fact that there will be unforeseen architectures with their unforeseen optimizations concerns.

The basic approach is one of separation of concerns. This is critical

- Not only for predicted architectures but for unpredicted architectures
- Not only for predicted optimization goals but for unpredicted system styles and unpredicted optimization goals.

The basic idea is to develop a program description that includes

- Everything about the meaning of the program
- Everything that might be useful for optimization

but it explicitly does not include anything that might have to be undone to improve performance for some new, unforeseen architecture or optimization goal. This application description is then paired with an appropriate tuning and/or runtime, and even a new approach to tuning and/or runtime.

The initial motivation for CnC was to support the separation of concerns between the activities of the domain-expert from those of the tuning-expert even within an individual but also among distinct professionals, each with their own expertise. This isolation now supports an unchanged application specification with a wide variety of specific tunings and even a variety of runtime styles. We believe it inherently supports new, unpredicted architectures and tuning goals. We also believe that a variety of current approaches to performance portability would pair well with this style of application specification.

The talk will describe the CnC program description language showing the rationale for its features. It will then describe briefly a wide variety of runtime systems and optimizations that have already been implemented as well as some that we're planning.

3.18 Performance Portability Challenges in FPGAs

Naoya Maruyama (LLNL – Livermore, US)

This talk will discuss performance portability challenges when using FPGAs as an accelerator. Developing applications that run portability across devices including FPGAs is now feasible by using common programming interfaces such as OpenCL. However, exploiting performance of FPGAs tends to require FPGA-specific parallelization and optimization, causing performance portability challenges. We will show several case studies using OpenCL-based benchmarks.

3.19 Crossing the Chasm: How to develop weather and climate models for next generation computers?

Chris Maynard (MetOffice – Exeter, GB)

Joint work of Bryan N. Lawrence, Michael Rezny, Reinhard Budich, Peter Bauer, Jörg Behrens, Mick Carter, Willem Deconinck, Rupert Ford, Christopher Maynard, Steven Mullerworth, Carlos Osuna, Andrew Porter, Kim Serradell, Sophie Valcke, Nils Wedi, Simon Wilson
Main reference Bryan N. Lawrence, Michael Rezny, Reinhard Budich, Peter Bauer, Jörg Behrens, Mick Carter, Willem Deconinck, Rupert Ford, Christopher Maynard, Steven Mullerworth, Carlos Osuna, Andrew Porter, Kim Serradell, Sophie Valcke, Nils Wedi, Simon Wilson: "Crossing the Chasm: How to develop weather and climate models for next generation computers?", Geosci. Model Dev. Discuss., in review, 2017.
URL https://doi.org/10.5194/gmd-2017-186

Weather and climate models are complex pieces of software which include many individual components, each of which is evolving under the pressure to exploit advances in computing to enhance some combination of a range of possible improvements (higher spatio/temporal resolution, increased fidelity in terms of resolved processes, more quantification of uncertainty etc). However, after many years of a relatively stable computing environment with little choice in processing architecture or programming paradigm (basically X86 processors using MPI for parallelism), the existing menu of processor choices includes significant diversity, and more is on the horizon. This computational diversity, coupled with ever increasing software complexity, leads to the very real possibility that weather and climate modelling will arrive at a chasm which will separate scientific aspiration from our ability to develop and/or rapidly adapt codes to the available hardware.

In this paper we review the hardware and software trends which are leading us towards this chasm, before describing current progress in addressing some of the tools which we may be able to use to bridge the chasm. This brief introduction to current tools and plans is followed by a discussion outlining the scientific requirements for quality model codes which have satisfactory performance and portability, while simultaneously supporting productive scientific evolution. We assert that the existing method of incremental model improvements employing small steps which adjust to the changing hardware environment is likely to be inadequate for crossing the chasm between aspiration and hardware at a satisfactory pace, in part because institutions cannot have all the relevant expertise in house. Instead, we outline a methodology based on large community efforts in engineering and standardisation, one which will depend on identifying a taxonomy of key activities – perhaps based on existing efforts to develop domain specific languages, identify common patterns in weather and climate codes, and develop community approaches to commonly needed tools, libraries etc – and then collaboratively building up those key components. Such a collaborative approach will depend on institutions, projects and individuals adopting new interdependencies and ways of working.

3.20 Performance portability: the good, the bad, and the ugly

Simon McIntosh-Smith (University of Bristol, GB)

URL http://uob-hpc.github.io

Achieving functional portability across a diverse range of computer architectures, such as CPUs and GPUs, is already a big challenge. Adding performance portability, where the same code performs well across those diverse architectures, is usually too ambitious a goal for scientific software developers. But what are the fundamental reasons behind this problem? Today, several parallel programming models can target a diverse range of hardware platforms: OpenMP 4.5, OpenCL, Kokkos and SYCL are just a small list of open source approaches for cross platform parallel programming. Why can't we write one program in, say OpenMP 4.5, and have this one code run well on GPUs from NVIDIA and AMD, and CPUs from Intel, IBM, Cavium et al. What are the fundamental technical problems that make this hard? And if we can enumerate and quantify these reasons, how can we then consciously design codes to avoid the main performance portability inhibitors?

In my HPC research group we have been studying performance portability since 2009. We have collected numerous case studies, where some codes are naturally performance portable, some can be adapted to become performance portable, while others appear to be naturally hostile to performance portability. Our initial research used OpenCL, but in the last few years our focus has been on OpenMP 4.x, and the emerging C++ parallel programming models that support cross-platform code generation (Kokkos, SYCL, Raja et al). Our target application areas have included life science codes (mostly molecular dynamics), and multi-physics codes (such as particle transport, heat diffusion and hydrodynamics).

In this talk I will describe our performance portability findings, including what we've found does work, what doesn't work, and where we think the most interesting open questions are.

3.21 AnyDSL: A Compiler-Framework for Domain-Specific Libraries (DSLs)

Richard Membarth (DFKI – Saarbrücken, DE)

AnyDSL is a framework for the rapid development of domain-specific libraries (DSLs). AnyDSL's main ingredient is AnyDSL's intermediate representation Thorin. In contrast to other intermediate representations, Thorin features certain abstractions which allow to maintain domain-specific types and control-flow. On these grounds, a DSL compiler gains two major advantages:

- The domain expert can focus on the semantics of the DSL. The DSL's code generator can leave low-level details like exact iteration order of looping constructs or detailed memory layout of data types open. Nevertheless, the code generator can emit Thorin code which acts as interchange format.
- The expert of a certain target machine just has to specify the required details once. These details are linked like a library to the abstract Thorin code. Thorin's analyses and transformations will then optimize the resulting Thorin code in a way such that the resulting Thorin code appears to be written by an expert of that target machine.

3.22 If you've scheduled loops, you've gone too far

Lawrence Mitchell (Imperial College London, GB)

Joint work of David A. Ham, Miklós Homolya, Paul H. J. Kelly, Fabio Luporini, Lawrence Mitchell
Main reference Miklós Homolya, Lawrence Mitchell, Fabio Luporini, David A. Ham: "TSFC: a structure-preserving form compiler", CoRR abs/1705.03667, 2017.
URL https://arxiv.org/abs/1705.003667

The optimal loop schedule for a given algorithm is typically hardware dependent, even with all other parameters of the algorithm fixed. When we manually port code to a new hardware platform, we must understand the loop structure and then perform, in tandem, data layout and loop reordering to achieve good performance. This is a difficult task. I argue that requiring a compiler system to perform the same task will never work: the scheduled loop nest does not offer enough information to the compiler for it to determine the algorithmic structure. Instead, we should strive for program transformation steps that operate on *unscheduled* DAGs. This is most easily achieved with DSLs, since no analysis is required. I will say some things about how we achieve this in the context of finite element codes, but will mostly be full of questions.

References

1 F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. J. Kelly, *Firedrake: automating the finite element method by composing abstractions*, ACM Transactions on Mathematical Software, 43 (2016), pp. 24:1–24:27, https://doi.org/10.1145/2998441, https://arxiv.org/abs/1501.01809.

2 M. Homolya, R. C. Kirby, and D. A. Ham, *Exposing and exploiting structure: optimal code generation for high-order finite element methods*, 2017, https://arxiv.org/abs/1711.02473.

3.23 Composing parallel codes while preserving portability of performance

Raymond Namyst (University of Bordeaux, FR)

Joint work of Raymond Namyst, Andra Hugo, Terry Cojean, Nathalie Furmento, Samuel Thibault, Pierre-André Wacrenier

Future exascale computers are expected to exhibit an unprecedented degree of parallelism together with a deeply hierarchical architecture, which only a few experts will be able to exploit efficiently if no significant progress is made in HPC software development. This situation is even more regrettable because there already exist many candidate applications with the potential to occupy the massive number of computing resources promised by exascale machines. Coupled applications for instance, which typically implement multi-physics and/or multi-scale simulations, exhibit high degrees of parallelism.

For such applications, designing a completely new "exascale programming model" can not be the answer, as rewriting coupled applications from scratch would require state-of-the-art skills in several domains (e.g. astrophysics and machine learning). Building these applications by reusing existing codes is thus the only reasonable way to go. The main challenge is thus about enabling a smooth transition to exascale that would leverage a significant base of existing codes and applications.

Unfortunately, reusing parallel codes in HPC applications is not commonplace at all. The main reason lies in current implementations of parallel libraries not being ready to run simultaneously over the same hardware resources, causing resource oversubscription, scheduling interferences, and other problems linked to unawareness of resource usage and compatibility of execution models. This problem – known as the parallel composability problem – significantly limits the way parallel codes can interact together.

This presentation covers the research topics raised by the challenge of "efficient parallel code reuse". In the light of our experience in the design of the StarPU task-based runtime system, we discuss how these challenges could be addressed by future runtime systems and programming environments.

References

1 C. Augonnet, S. Thibault, R. Namyst, P.-A. Wacrenier, StarPU: A Unified Platform for Task Scheduling on Heterogeneous Multicore Architectures, Concurr. Comput. : Pract. Exper. 23 (2011) 187–198.

2 H. Pan, B. Hindman, K. Asanović, Composing parallel software efficiently with lithe, SIGPLAN Not. 45 (2010) 376–387.

3.24 Performance Portability through Device Specialization

Simon Pennycook (Intel – Santa Clara, US)

Joint work of Simon Pennycook, Jason Sewall, Alejandro Duran, Victor Lee
Main reference S. J. Pennycook, J. D. Sewall, V. W. Lee, "Implications of a Metric for Performance Portability", In Future Generation Computer Systems, ISSN 0167-739X, 2017.
URL https://doi.org/10.1016/j.future.2017.08.007

We discuss the utility of shared definitions and metrics for "performance portability", and the implications of one such metric on the design of highly performance portable software. Specifically, we demonstrate that maintaining multiple implementations of some limited subset of application functionality can significantly improve performance portability, thereby motivating improved mechanisms for managing function variants in software.

3.25 Loop descriptors and cross-loop techniques: portability, locality and more

Istvan Reguly (Pazmany Peter Catholic University – Budapest, HU)

Joint work of Istvan Reguly, Mike Giles, Gihan Mudalige

This talk will outline what additional information can be added to "parallel for" loops about data accesses and how it enables high-level optimisations and portability. Adding in a user contract that data is not modified outside of these loops, we then show how the execution of these loops can be delayed, which enables cross-loop analysis and optimisations, showing a few examples involving cache-blocking tiling and checkpointing. This talk aims to prompt discussion on the further possible uses of per-loop and cross-loop techniques, and the relaxation of domain-specific semantics.

3.26 Characterizing Data Movement Costs: Tools Needed!

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

Data movement overheads dominate operation execution costs, both in terms of time (performance) as well as energy. But while the computational complexity of algorithms (in terms of number of elementary operations) is well understood, the data-movement complexity of algorithms is not. A significant complication for users to reason about the inherent data-movement complexity of computations is that unlike computational complexity, it is not additive under composition. Tools for characterizing data-movement complexity will be important in facilitating high performance, productivity, and performance-portability with high-level frameworks.

3.27 Performance Portability Through Code Generation – Experiences in the context of SaC

Sven-Bodo Scholz (Heriot-Watt University – Edinburgh, GB)

URL www.sac-home.org

This talk gives a bird's eye view on the performance portability experiences made in the context of compiling SaC into high-performance codes for a range of architectures. Particular emphasis is being put on the challenges met, the interplay between language design and those challenges, as well as a discussion on how these experiences compare to other approaches such as DSLs, staging or library based solutions.

3.28 Orthogonal Abstractions for Scheduling and Storage Mappings

Michelle Mills Strout (University of Arizona – Tucson, US)

With ever increasing amounts of HW parallelism, it is becoming more and more critical to reduce synchronization overhead and improve the data locality of computations. Reducing synchronization overhead can be done by creating many, decent-sized tasks that are only loosely synchronized. Improving data locality requires that the computation placed within a single task reuses data and that that data fits within the memory resources where the task is allocated. Reducing temporary storage is important as well. These activities require scheduling across loops and in some cases function calls and modifying how data values are mapped to storage locations.

Systems do this in a multitude of ways. For this discussion, some example approaches from projects such as OP2, Chapel, Kokkos, CnC, and Loop Chaining will be overviewed. What might an orthogonal abstraction stack for performance programming look like? How would such an abstraction stack interact with performance portability?

3.29 Mis-predicting performance

Nathan Tallent (Pacific Northwest National Lab. – Richland, US)

Main reference Ryan D. Friese, Nathan R. Tallent, Abhinav Vishnu, Darren J. Kerbyson, Adolfy Hoisie: "Generating Performance Models for Irregular Applications", in Proc. of the 2017 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2017, Orlando, FL, USA, May 29 - June 2, 2017, pp. 317–326, IEEE Computer Society, 2017.
URL http://dx.doi.org/10.1109/IPDPS.2017.61

Performance modeling gives insight into bottlenecks to performance portability. There are several approaches to performance modeling. Analytical models promise insight by representing program behavior using algebraic expressions that are a function of input parameters. However, they can be time-consuming to create and may miss important corner cases. Statistical and machine learning models can be more easily automated. However, they easily confuse cause and correlation; and are only as good as their training sets.

This talk focuses on some of the open questions in modeling performance. We use examples from recent work on modeling applications and workflows with irregular behavior, e.g, input-dependent solvers, irregular memory accesses, or unbiased branches. This irregular behavior often leads to long-latency events that are difficult to characterize when moving between different systems or programming models.

3.30 Performance Portability with Pragmas – Wishful Thinking or Real Opportunity?

Christian Terboven (RWTH Aachen, DE)

Since the evolution of OpenMP and OpenACC to de-facto standards in the HPC community, directive-based parallelization offers a higher level of abstraction than system- or device-level APIs. Both standards hide certain details from the programmer to focus on expressing parallelism at different levels. To answer the question if directives can offer an opportunity to enable performance portable programming, this contribution summarized and discussed recent developments within OpenMP.

OpenMP 4.5 was released in 2015 and introduced a feature called locks with hints. It enables the use of different lock types within a single program. Furthermore, it allows the specification of a lock hint that the runtime can exploit to determine the most efficient available implementation. For example the lock hint ‚speculative‘ may be realized by hardware-supported memory transactions, if supported by the target architecture.

OpenMP TR5 was released in 2016 and provided an early view of memory management support scheduled for inclusion in OpenMP 5.0. The addition of allocators as a concept in OpenMP enables the use of different kinds of memory. Instead of targeting a dedicated memory kind with a device-specific API, OpenMP TR5 allows to allocate from memory with a specific feature, for example memory with the highest available bandwidth can be selected. The concrete memory kind will be selected at runtime based on capabilities of the target architecture.

Both examples illustrate how the use of directives can hide hardware-level details and how the runtime system can apply optimizations under the hood. OpenMP and OpenACC differ in being prescriptive or descriptive. In the prescriptive approach of OpenMP, the programmer explicitly instructs how parallelism is extracted from the application code and mapped to the system. OpenACC tends to require fewer specific details from the programmer and leaves more freedom and room for optimization to the compiler and runtime. The use of directives over low-level APIs can contribute to the goal of performance portable programming, but has to be complemented by other measures.

References

1 H. Bae, J. Cownie, M. Klemm and C. Terboven: A User-Guided Locking API for the OpenMP Application Programming Interface. IWOMP 2014: 173–186, 2014.

2 J. Sewall, J. Pennycook, A. Duran, C. Terboven, X. Tian and R. Narayanaswamy: Developments in Memory Management in OpenMP. IJHPCN, to appear.

3.31 Thread and Data Placement on Multicore Architectures

Didem Unat (Koc University – Istanbul, TR)

Joint work of Didem Unat, Pirah Noor Soomro, Muhammad Aditya Sasongko, Mohammad Laghari
Main reference Mohammad Laghari, Didem Unat: "Object Placement for High Bandwidth Memory Augmented with High Capacity Memory", in Proc. of the 29th International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD 2017, Campinas, Brazil, October 17-20, 2017, pp. 129–136, IEEE Computer Society, 2017.
URL http://dx.doi.org/10.1109/SBAC-PAD.2017.24

Placement of parallel tasks and data in current multicore machines is one of the key aspects to achieve good performance. A thread placement policy is formulated by discovering machine's topology and analyzing application's behavior of sharing data among parallel task. This policy is used for binding application threads during its execution. Current binding options supported by the runtime systems do not analyze application's behavior and only provide basic mapping policies. To improve the performance programmer has to put an extra effort to understand core topology and memory hierarchy, then bind the tasks accordingly. This process is not portable; programmer has to repeat the analysis when the platform or application changes. We propose a thread mapping algorithm based on the communication pattern and machine topology. Our algorithm is fair in other words, it tries to pair threads that communicate most with each other and considers mutual preferences.

In addition to thread placement, in a memory subsystem where there are different types of memories, placement of data objects becomes a programming issue because the programmer has to decide which objects to place on which types of memory. We propose an object placement algorithm that places program objects to fast or slow memories by considering characteristics of each memory type. Our algorithm uses the reference counts and type of references (read or write) to make an initial placement of data. By placing objects according to our placement algorithm, we are able to achieve a speedup of up to 2x with 6 applications under various system configurations.

3.32 A Heterogeneous Talk: thoughts on portability, heterogeneous computing, and graphs

Ana Lucia Varbanescu (University of Amsterdam, NL)

Joint work of Stijn Heldens, Ana Lucia Varbanescu, Alexandru Iosup
URL https://github.com/stijnh/HyGraph

In this talk, we discuss performance portability from the perspective of programming models and heterogeneous computing. We further discuss the implications of irregular applications and data-dependent performance variability on such issues as performance portability and heterogeneous computing. Finally, we briefly introduce HyGraph, our runtime-based solution for heterogeneous graph processing. HyGraph is a novel graph-processing system for hybrid platforms which delivers performance by utilizing both the CPU and the GPU concurrently. Contrary to the state-of-the-art approach of statically partitioning the workload beforehand, HyGraph delivers performance by dynamic scheduling of jobs onto both the CPU and the GPU, thus providing automatic load balancing and superseding the need for the user to manually define a static workload distribution. Additionally, HyGraph minimizes the inter-process communication overhead by carefully overlapping computation and communication. Our results demonstrate that HyGraph can deliver system-level "efficiency portability" on different, large-scale systems.

3.33 Challenges with Different approaches for Performance Portability

Mohamed Wahib (AIST – Tokyo, JP)

The approaches for tackling the performance portability problem typically fall under one, or a mix, of the following: DSLs, compiler-based approaches, source-to-source translators, and libraries. Considering the diversity in pros and cons for each of those approaches, it is hard for the programmer to decide on which approach to follow. This is especially challenging when the programmer has to address performance portability for legacy code. This presentation compares experiences in addressing performance portability, using each of the approaches mentioned above. Main pitfalls and limitations, for some legacy real-world applications, are also discussed.

3.34 Deep memory hierarchies and performance portability

Michele Weiland (University of Edinburgh, GB)

Memory hierarchies are becoming deeper and more complex, and byte-addressable SCM is blurring the line between memory and storage. How users and applications can benefit from new memory technologies depends on how new layers in the memory hierarchy are exposed and used. The question is: whose responsibility is it to guarantee transparent use of deep memory hierarchies? Can the application developer expect this to be dealt with on a system level, or will it require direct intervention at the programming level? In this presentation I will look at the implications of complex memory hierarchies for performance and performance portability.

3.35 Musings on Performance Portability

Michael Wolfe (NVIDIA Corp., US)

The term "performance portability" has been used for at least twenty years. Some people argue that it is not achievable, and even that it's not desireable. This is folly. However, there is a tension between three P's: Performance, Productivity, and Portability. Each user or organization needs to optimize this function. Generally, any program can be made to run faster with greater programming effort, such as using machine language, but the productivity cost could be prohibitive. Similarly, programming in machine language reduces portability, even across different generations of the same instruction set architecture. But different people and organizations will have different pain thresholds for how much performance they are willing to give up to get some degree of productivity and of portability. Perhaps we should rename this discussion "Productive Performance Portability."

Let's review some successes in performance portability across the ages: 1957: The Fortran compiler from IBM, which had a goal to generate code that was as fast as hand-written machine language, allowed programs that could port across machines, and has been extremely successful for over 60 years now. 1977: Vectorizing compilers virtualized the details of vector code generation and allowed programmers to take advantage of vector instructions, and now SIMD instructions, without having to know details such as the vector or SIMD register length, and became the dominant method for programming vector computers. 1997: MPI was introduced earlier in the decade, but by 1977 it had replaced all previous message passing libraries, and remains the most common method for scalable programming, across thousands of nodes on a supercomputer network.

Now it's 2017, and we have highly scalable nodes, with many dozens or hundreds of compute units, all with shared memory or high-speed interconnected memories. Can we achieve productive performance portability across the range of such systems? I argue that we should be able to, because they all share many characteristics: Many processing elements, SIMD execution, multithreading, hardware caches. There have been a number of approaches, including languages, compilers, runtime systems, and numerical libraries. All of these may have a role to play. However, if there's a lesson to learn from history it's that we need to train programmers, many or most of whom are more interested in science than in programming, how to write programs that perform well and that will port to future computer systems as well. Vectorizing compilers played a role in training programmers for those machines, by giving immediate and precise feedback about how well the loops vectorized and why they did not. Such tools should be explored for future productive performance portability as well.

4 Open problems

4.1 Performance Portability via Automatic Region-based Auto-tuning

Philipp Gschwandtner (Universität Innsbruck, AT)

Joint work of Philipp Gschwandtner, Juan J. Durillo, Thomas Fahringer
Main reference Klaus Kofler, Juan José Durillo, Philipp Gschwandtner, Thomas Fahringer: "A Region-Aware Multi-Objective Auto-Tuner for Parallel Programs", in Proc. of the 46th International Conference on Parallel Processing Workshops, ICPP Workshops 2017, Bristol, United Kingdom, August 14-17, 2017, pp. 190–199, IEEE Computer Society, 2017.
URL http://dx.doi.org/10.1109/ICPPW.2017.37

Optimizing parallel programs for modern architectures and striving for performance portability is a notoriously difficult task. The increasing complexity of multi- and many-core platforms, hardware details such as topologies, caches, and links and multiple layers of nested parallelism pose a limiting factor when designing both faster hardware and software in contemporary HPC. Auto-tuning has become increasingly popular to mitigate this issue, but still lacks key features for pervasive use throughout the community. First, many auto-tuner systems require user directives that e.g. describe performance relationships or bottlenecks, or specify a list of tunable parameters and ranges of valid settings. Second, most parallel programs can be subdivided into several regions, whose optimal tunable parameter settings might differ, and whose optimization might have adverse effects on subsequent regions - harming overall performance. These issues are greatly aggravated by parameter setting overheads, non-contiguous, combinatorial parameter spaces, and multi-objective optimization environments. To address this problem, a fully automatic, region-based auto-tuner is needed, minimizing user effort in realizing performance portability. The auto-tuner should automatically identify program code regions of interest, find promising parameter spaces for a given set of user objectives, and explore them efficiently.

Participants

- Sadaf Alam
CSCS – Lugano, CH
- Michael Bader
TU München, DE
- Carlo Bertolli
IBM TJ Watson Research Center – Yorktown Heights, US
- Mauro Bianco
CSCS – Lugano, CH
- Alexandru Calotoiu
TU Darmstadt, DE
- Bradford Chamberlain
Cray Inc. – Seattle, US
- Aparna Chandramowlishwaran
University of California – Irvine, US
- Kemal A. Delic
Hewlett Packard – Grenoble, FR
- Christophe Dubach
University of Edinburgh, GB
- Anshu Dubey
Argonne National Laboratory, US
- H. Carter Edwards
Sandia National Labs – Albuquerque, US
- Jan Eitzinger
Universität Erlangen-Nürnberg, DE
- Todd Gamblin
LLNL – Livermore, US
- Lin Gan
Tsinghua University – Beijing, CN
- William D. Gropp
University of Illinois – Urbana-Champaign, US
- Philipp Gschwandtner
Universität Innsbruck, AT
- Mary W. Hall
University of Utah – Salt Lake City, US
- Robert J. Harrison
Brookhaven National Laboratory – Upton, US
- Alexandra Jimborean
Uppsala University, SE
- Paul H. J. Kelly
Imperial College London, GB
- Andreas Klöckner
University of Illinois – Urbana-Champaign, US
- Kathleen Knobe
Rice University – Houston, US
- Seyong Lee
Oak Ridge National Laboratory, US
- Naoya Maruyama
LLNL – Livermore, US
- Chris Maynard
MetOffice – Exeter, GB
- Simon McIntosh-Smith
University of Bristol, GB
- Richard Membarth
DFKI – Saarbrücken, DE
- Lawrence Mitchell
Imperial College London, GB
- Bernd Mohr
Jülich Supercomputing Centre, DE
- Raymond Namyst
University of Bordeaux, FR
- Simon Pennycook
Intel – Santa Clara, US
- Istvan Reguly
Pazmany Peter Catholic University – Budapest, HU
- P. (Saday) Sadayappan
Ohio State University – Columbus, US
- Sven-Bodo Scholz
Heriot-Watt University – Edinburgh, GB
- Michelle Mills Strout
University of Arizona – Tucson, US
- Nathan Tallent
Pacific Northwest National Lab. – Richland, US
- Christian Terboven
RWTH Aachen, DE
- Didem Unat
Koc University – Istanbul, TR
- Ana Lucia Varbanescu
University of Amsterdam, NL
- Jeffrey S. Vetter
Oak Ridge National Laboratory, US
- Mohamed Wahib
AIST – Tokyo, JP
- Michele Weiland
University of Edinburgh, GB
- Robert Wisniewski
Intel – Santa Clara, US
- Michael Wolfe
NVIDIA Corp., US

Report from Dagstuhl Seminar 17441

Big Stream Processing Systems

Edited by

Tilmann Rabl[1], Sherif Sakr[2], and Martin Hirzel[3]

1 **TU Berlin, DE,** rabl@tu-berlin.de
2 **KSAU-HS, Riyadh, SA,** sakrs@ksau-hs.edu.sa
3 **IBM TJ Watson Research Center – Yorktown Heights, US,** hirzel@us.ibm.com

Abstract

This report summarizes the Dagstuhl Seminar 17441 on "Big Stream Processing Systems" and documents its talks and discussions. The seminar brought together 29 researchers in various areas related to stream processing including systems, query languages, applications, semantic processing and benchmarking. The seminar program included four tutorials that have been delivered by experts in the various topics in addition to 29 lightening talks by the participants of the seminar. In this report, the abstracts of these talks are documented. Two working groups has been formed during the seminar. A report about the discussion outcomes of each group is presented in this report.

Seminar October 29–3, 2017 – http://www.dagstuhl.de/17441

1998 ACM Subject Classification C.2.4 Distributed Systems, C.2.4 Distributed Databases, H.2.4 Query Processing, D.3.2 Data-Flow Languages

Keywords and phrases Big Data, Big Streams, Stream Processing Systems, Benchmarking, Declarative Programming

Digital Object Identifier 10.4230/DagRep.7.10.111

1 Executive Summary

Martin Hirzel
Tilmann Rabl
Sherif Sakr

As the world gets more instrumented and connected, we are witnessing a flood of digital data that is getting generated, in a high velocity, from different hardware (e.g., sensors) or software in the format of streams of data. Examples of this phenomena are crucial for several applications and domains including financial markets, surveillance systems, manufacturing, smart cities and scalable monitoring infrastructure. In these applications and domains, there is a crucial requirement to collect, process, and analyze big streams of data in order to extract valuable information, discover new insights in real-time and to detect emerging patterns and outliers. Recently, several systems (e.g., Apache Apex, Apache Flink, Apache Storm, Heron, Spark Streaming,) have been introduced to tackle the real-time processing of big streaming data. However, there are several challenges and open problems that need to be addressed in order improve the state-of-the-art in this domain and push big stream processing systems to make them widely used by large number of users and enterprises. The aim of this seminar was to bring together active and prominent researchers, developers and practitioners actively working in the domain of big stream processing to discuss very relevant open challenges and

Big Stream Processing Systems, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 111–138
Editors: Tilmann Rabl, Sherif Sakr, and Martin Hirzel
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

research directions. The plan was to work on specific challenges including the trade-offs of the various design decisions of big stream processing systems, the declarative stream querying and processing languages, and the benchmarking challenges of big stream processing systems.

On Monday morning, the workshop officially kicked off with a round of introductions about the participants where adhoc clusters for the interests of the participants have been defined. The clusters have been revolving around the topics of systems, query languages, benchmarking, stream mining and semantic stream processing. The program of the seminar included 4 tutorials, one per day. On Monday, Martin Strohbach from AGT International presented different case studies and scenarios for large scale stream processing in different application domains. On Tuesday, we enjoyed the systems tutorial which has been presented by Paris Carbone from KTH Royal Institute of Technology, Thomas Weise from Data Torrent Inc. and Matthias J. Sax from Confluent Inc. Paris presented an interesting overview of the journey of stream processing systems, Thomas presented the recent updates about the Apache Apex system while Matthias presented an overview about the Apache Kafka and Kafka Streams projects. On Wednesday, Martin Hirzel from IBM TJ Watson Research Center presented a tutorial about the taxonomy and classifications of stream processing languages. On Thursday, Tilmann Rabl from TU Berlin presented a tutorial about the challenges of benchmarking big data systems in general in addition to the specific challenges for benchmarking big stream processing systems. All tutorials have been very informative, interactive and involved very deep technical discussions. On Thursday evening, we had a lively demo session where various participants demonstrated their systems to the audience on parallel round-table interactive discussions. On Wednesday, the participants split into two groups based on common interest in selected subset of the open challengers and problems. The selected 2 topics of the groups were systems and query languages. Thursday schedule was dedicated to working group efforts. Summary about the outcomes of these 2 groups is included in this report. It is expected that work from at least one of the groups to be submitted for publication, and we expect further research publications to result directly from the seminar.

We believe that the most interesting aspect of the seminar was providing the opportunity to freely engage in direct and interactive discussions with solid experts and researchers in various topics of the field with common focused passion and interest. We believe that this is a unique feature for Dagstuhl seminars. We received very positive feedback from the participants and we believe that most of the participants were excited with the scientific atmosphere at the seminar and reported that the program of the seminar was useful for them. In summary, we consider the organization of this seminar as a success. We are grateful for the Dagstuhl team for providing the opportunity and full support to organize it. The success of this seminar motivated us to plan for future follow-up seminars to continue the discussions on the rapid advancements on the domain and plan for narrower and more focused discussion with concrete outputs for the community.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Approximate data analytics systems

Pramod Bhatotia (University of Edinburgh, GB)

Joint work of Pramod Bhatotia, Le Quoc Do, Martin Beck, Dhanya Krishnan, Ruichuan Chen, Christof Fetzer, Thorsten Strufe, Volker Hilt, Istemi Akkus, Rodrigo Rodrigues, Spyros Blanas

We present approximate data analytics systems. Approximate computing aims for efficient execution of workflows where an approximate output is sufficient instead of the exact output. The idea behind approximate computing is to compute over a representative sample instead of the entire input dataset. Thus, approximate computing — based on the chosen sample size — can make a systematic trade-off between the output accuracy and computation efficiency.

In this talk, I presented data analytics system for approximate computing. Our work aims for an efficient mechanism based on approximation for large-scale data analytics. In particular, I presented four systems for approximate computing: (1) StreamApprox, a stream analytics systems for approximate computing, (2) PrivApprox, a privacy-preserving stream analytics system using approximate computing, (3) IncApprox, a data analytics system that combines incremental and approximate computing, and lastly, (4) ApproxJoin, a data analytics to support approximate distributed joins.

We have built our systems based on prominent distributed data analytics platforms, such as Apache Spark Streaming, and Apache Flink, which allow to transparently target a large class of existing applications. The source code of our approximate data analytics systems along with the full experimental evaluation setup is publicly available for the research community.

- Approx [1]: https://privapprox.github.io/
- StreamApprox [3]: https://streamapprox.github.io/
- IncApprox [2]: https://gitlab.com/tudinfse/incapprox

References

1 Do Le Quoc, Martin Beck, Pramod Bhatotia, Ruichuan Chen, Christof Fetzer, Thorsten Strufe. *PrivApprox: Privacy-Preserving Stream Analytics*. USENIX ATC, 2017.

2 Dhanya R. Krishnan, Do Le Quoc, Pramod Bhatotia, Christof Fetzer, Rodrigo Rodrigues. *IncApprox: A Data Analytics System for Incremental Approximate Computing*. WWW, 2016.

3 Do Le Quoc, Ruichuan Chen, Pramod Bhatotia, Christof Fetzer, Volker Hilt, Thorsten Strufe. *StreamApprox: approximate computing for stream analytics*. Middleware, 2017.

3.2 Data Stream Mining

Albert Bifet (Telecom ParisTech, FR)

Big Data and the Internet of Things (IoT) have the potential to fundamentally shift the way we interact with our surroundings. The challenge of deriving insights from the Internet of Things (IoT) has been recognized as one of the most exciting and key opportunities for both academia and industry. Advanced analysis of big data streams from sensors and devices is bound to become a key area of data mining research as the number of applications requiring such processing increases. Dealing with the evolution over time of such data streams, i.e., with concepts that drift or change completely, is one of the core issues in stream mining. I will present an overview of data stream mining, and I will introduce two popular open source tools for data stream mining.

3.3 Analysis of Queries on Big Graphs

Angela Bonifati (University Claude Bernard – Lyon, FR)

Joint work of Angela Bonifati, Nicky Advokaat, Guillaume Bagan, Radu Ciucanu, George Fletcher, Benoit Groz, Aurelien Lemay, Wim Martens, Thomas Timm
Main reference Angela Bonifati, Wim Martens, Thomas Timm: "An Analytical Study of Large SPARQL Query Logs", PVLDB 11(2): 149–161, 2017.
URL http://www.vldb.org/pvldb/vol11/p149-bonifati.pdf

My research revolves around graph data management by focusing in particular on graph query processing, graph benchmarking and graph query log analysis . Several modern graph query languages are capable of expressing sophisticated graph queries, which return nodes connected by arbitrarily complex labeled paths. Such paths can be synthesized by means of regular expressions and often involve recursion. Such graph queries are known as Regular Path Queries (RPQ) and correspond to Property Paths in Sparql 1.1 to make an example of a concrete graph query language. Graph queries arbitrarily combine RPQ with conjunctions and unions to constitute a comprehensive query language for graph databases. Recently, with my colleagues I have been investigating graph queries and their different fragments by studying the synthetic generation problem of graph instances and graph query workloads [1, 2], along with the complexity of graph query evaluation [3] and the analysis of a large corpus of real-world graph queries [4]. In the latter work, we have examined streaks of queries that are queries that originate from gradual modifications of a seed query. If we switch from batch processing to online processing, these streaks can be considered as streams of queries collected from SPARQL endpoints. In summary, graph benchmarking and graph log analysis are essential to shape the future capabilities of graph query engines. Hence, they have a strong potential to influence the work of our community on graph query processing and optimization.

References

1 Guillaume Bagan, Angela Bonifati, Radu Ciucanu, George Fletcher, Aurelien Lemay, and Nicky Advokaat. *Generating Flexible Workloads for Graph Databases. PVLDB*, 9(13):1447–1460, 2016.

2 Guillaume Bagan, Angela Bonifati, Radu Ciucanu, George Fletcher, Aurelien Lemay, and Nicky Advokaat. *gMark: Schema-Driven Generation of Graphs and Queries. IEEE Trans. on Knowl. Data Eng.*, 29(4): 856–869, 2017

3 Guillaume Bagan, Angela Bonifati, and Benoit Groz. *A trichotomy for regular simple path queries on graphs*. In *Proceedings of PODS*, pages 261–272, 2013.

4 Angela Bonifati, Wim Martens, Thomas Timm: An Analytical Study of Large SPARQL Query Logs. *PVLDB* 11(2): 149–161 (2017)

3.4 Privacy Preserving, Peta-scale Stream Analytics for Domain-Experts

Michael H. Böhlen (Universität Zürich, CH)

Joint work of Michael H. Böhlen, Abraham Bernstein, Daniele dell'Aglio, Muhammad Saad, Pengcheng Duan

Production of big data will soon outpace the availability of both storage and computer science experts who know how to handle such data. Moreover, society is increasingly concerned about data protection. Addressing these issues requires so-called stream-processing systems that continuously analyse incoming data (rather than store it) and allow non-computer scientists to specify its analysis in a privacy-preserving manner. We will develop a petabyte-scale stream analytics system that enables non-computer scientists to analyse high-performance data streams on commodity hardware. First, we provide a declarative language based on traditional querying but with extensions for statistical operations and capabilities for real-time operations. Second, the language permits users to specify the desired level of privacy. Third, the system translates the statistical functions and privacy specifications into executable computations. Finally, the runtime environment selects the best approach for optimising execution using existing systems (e.g. Apache Flink, Spark Streaming or Storm). To evaluate the robustness and functionality of our system, we will replicate the processing pipeline for the Australian Square Kilometre Array Pathfinder radio telescope. This will generate up to 2.5 gigabytes per second of raw data. To evaluate privacy preservation, we will analyse the TV viewing habits of around 3 million individuals.

3.5 Interconnecting the Web of Data Streams

Jean-Paul Calbimonte (HES-SO Valais – Sierre, CH)

Main reference Daniele Dell'Aglio, Danh Le Phuoc, Anh Lê Tuán, Muhammad Intizar Ali, Jean-Paul Calbimonte: "On a Web of Data Streams", in Proc. of the Workshop on Decentralizing the Semantic Web 2017 co-located with 16th Int'l Semantic Web Conf. (ISWC), 2017.
URL http://ceur-ws.org/Vol-1934/contribution-11.pdf
Main reference Jean-Paul Calbimonte: "Linked Data Notifications for RDF Streams", in Joint Proc. of the Web Stream Processing workshop (WSP 2017) and the 2nd Int'l Workshop on Ontology Modularity, Contextuality, and Evolution (WOMoCoE 2017) co-located with 16th Int'l Semantic Web Conference (ISWC 2017), pp. 66–73, Vienna, Austria, October 22nd, 2017.
URL http://ceur-ws.org/Vol-1936/paper-06.pdf

The Web evolves toward a vast network of data, making it possible to publish, discover, access, process, and consume information through standard protocols. This Web of Data increasingly includes data streams, whose velocity and variety challenge current methods and techniques for processing/consumption/publishing. Although semantic data models such as RDF have shown to be successful for addressing these issues for stored data, it remains an open problem for data streams on the Web. We propose using standard protocols such as Linked Data Notifications (LDN) as the backbone for sending, receiving and consuming stream elements on the Web. This will allow a wider reuse of stream, going beyond existing silos and technical and administrative barriers. To achieve this, we envision the use of semantically rich metadata that describes these streams, regardless of their format and structure, providing the means to enhance findability, accessibility, linkability of these streams.

3.6 Consistent Large-Scale Data Stream Processing

Paris Carbone (KTH Royal Institute of Technology, SE)

Joint work of Paris Carbone, Stephan Ewen, Gyula Fóra, Seif Haridi, Stefan Richter, Kostas Tzoumas
Main reference Paris Carbone, Stephan Ewen, Gyula Fóra, Seif Haridi, Stefan Richter, Kostas Tzoumas: "State Management in Apache Flink®: Consistent Stateful Distributed Stream Processing", PVLDB, Vol. 10(12), pp. 1718–1729, 2017.
URL http://www.vldb.org/pvldb/vol10/p1718-carbone.pdf

An early dogma on data stream processing technology labeled that tech as a fast, yet approximate, method for data analysis [5]. This argument has its roots on early design choices that considered limited in-memory data structures to maintain state, the lack of scale-out approaches seen in Map-Reduce and most importantly the conception that offering strong state consistency guarantees was non-trivial in the context of streams. Today, we are witnessing a 'big stream processing' revolution where stream processors such as Apache Flink, Kafka-Streams, Apex, Millwheel [4] (Beam/Dataflow Cloud [2] runner) etc. are being adopted as building blocks for scalable, continuous processing applications and pipelines. Modern data stream processors offer built-in state management with exactly-once end-to-end guarantees, eliminating the possibility of data loss or state inconsistencies as well as scaling-out in a data-parallel manner. The two dominant architectures to state management are 1) Externally persisted state in a transactional data store [4] and 2) Local state to compute nodes that is committed and replicated using consistent distributed snapshots flink [3, 1],ibmstreams,apex. Throughout this lightning talk and discussion we analyse the reasons behind locally maintained state, primarily in the context of Apache Flink [1] and its core snapshotting algorithm. We show that snapshots can be used for all operational needs of a long-running application such as live reconfiguration, fault tolerance, software patches and versioning. Finally, we address the costs of employing such an architecture both in terms of runtime overhead and reconfiguration time.

References

1 P. Carbone, S. Ewen, G. Fóra, S. Haridi, S. Richter, K. Tzoumas, "State management in Apache Flink®: consistent stateful distributed stream processing" *Proceedings of the VLDB Endowment*, 2017.

2 Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., et al.: The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. VLDB (2015)

3 P. Carbone, S. Ewen, S. Haridi, A. Katsifodimos, V. Markl, and K. Tzoumas, "Apache flink: Stream and batch processing in a single engine," *IEEE Data Engineering Bulletin*, p. 28, 2015.

4 Akidau T, Balikov A, Bekiroglu K, Chernyak S, Haberman J, Lax R, McVeety S, Mills D, Nordstrom P, Whittle S (2013) MillWheel: Fault-tolerant stream processing at internet scale. In: VLDB

5 The Lambda Architecture. http://lambda-architecture.net/

3.7 Tutorial: Data Stream Processing Systems

Paris Carbone (KTH Royal Institute of Technology, SE)

Conventional data management considers data, in its traditional definition, as facts and statistics organised and collected together for future reference or analysis. A wide family of database management systems designed around the principle of having data as a static component, yet allowing complex and flexible retrospective analysis on that data such as declarative adhoc sql queries. Large-scale data processing architectures aimed to scale out those technologies, examples are Map-Reduce and the Apache Spark stack. Yet, regardless of the undeniable scale-out we effectively got the same old perception of data processing in a "new outfit".

Data stream processing revolutionizes the way we define data in the first place, lifting its context from retrospective data set analysis to continuous unbounded processing coupled with persistent application state. Parts of that technology have been available in different primary forms and domains, such as network-centric processing on byte streams, functional programming (e.g., monads), actor programming and materialized views. However, we now see how all of these ideas have been put together to form a system architecture that is built and therefore keeps evolving around the notion of data as an unbounded collective stream of facts.

This new "Scalable Stream Processing" architecture has its own stack. Starting from the storage layer, dedicated partitioned logs such as Apache Kafka and Pravega replace distributed file systems (e.g., HDFS). Partitioned logs have unique characteristics that build on the notion of streams such as event stream producers, consumers, delivery guarantees as well as stream reconfiguration capabilities. At the middle, we have data compute systems such as Flink, Kafka-Streams, Apex, Timely-Dataflow and Spark-Streaming that offer continuous stateful stream processing with, most often, end-to-end exactly-once processing guarantees. Furthermore, the semantics of all these systems have been lifted from primitive per-event processing to declarative high-level abstractions such as stream windowing (triggers, event-time domain integration, sessions etc.). Finally, at the top-most layer we can see new forms of libraries and user-facing programming models covering relational event streams (stream sql), complex event processing (cep) as well as newly formed domain-specific languages and models for streams.

The ultimate aim of the tutorial is to offer a top-down view of all these concepts and pose potential insights of the upcoming needs of stream processing technology. On that end, we discuss the prospects and benefits of standardization, both in terms of runtime characteristics (e.g., snapshots), core programming models (e.g., Beam) and finally higher-level APIs (Stream SQL, Calcite, complex event processing APIs).

3.8 Cyber-Physical Social Systems for City-wide Infrastructures

Javier David Fernández-García (Wirtschaftsuniversität Wien, AT)

Main reference Aljbin Ahmeti, Saimir Bala, Fajar J. Ekaputra, Javier D. Fernández, Elmar Kiesling, Andreas Koller, Jan Mendling, Angelika Musil, Axel Polleres, Peb R. Aryan, Marta Sabou, Andreas Solti, Juergen Musil: "CitySPIN: Cyber-Physical Social Systems for City-wide Infrastructures", in Proc. of the 13th Int'l Conf. on Semantic Systems, Posters and Demos track, 2017.
URL http://ceur-ws.org/Vol-2044/paper21

The potential of Big Semantic Data is under-exploited when data management is based on traditional, human-readable RDF representations, which add unnecessary overheads when storing, exchanging and consuming RDF in the context of a large-scale and machine-understandable Semantic Web. In the first part of the talk, we briefly present our HDT [1] proposal, a compact data structure and binary serialization format that keeps big RDF datasets compressed while maintaining search and browse operations without prior decompression. As a practical use case, we show the recent project LOD-a-lot [4], which uses HDT to represent and query more than 28 billion triples of the current Linked Open Data network. Then, we inspect the challenges of using a write-once-read-multiple HDT format in a streaming scenario, providing details on two real-time solutions: i) SOLID [2], a lambda-based compact triplestore to manage evolving Big Semantic Data, and ii) ERI [3], a compact serialization format for RDF streams. Finally, we present CitySPIN [5], a project using such Big Semantic Data technologies to integrate and manage cyber-physical social systems in order to facilitate innovative Smart City infrastructure services.

References

1 J. D. Fernández, M. A. Martínez-Prieto, C. Gutiérrez, A. Polleres, and M. Arias. Binary RDF Representation for Publication and Exchange. *Journal of Web Semantics*, 19:22–41, 2013.

2 M. A. Martínez-Prieto, C. E. Cuesta, M. Arias, and J. D. Fernández. The solid architecture for real-time management of big semantic data. *Future Generation Computer Systems*, 47, 62–79, 2015.

3 J. D. Fernández, A. Llaves, and O. Corcho. Efficient RDF interchange (ERI) format for RDF data streams. In *Proceedings of the International Semantic Web Conference*, pp. 244–259, 2014.

4 J. D. Fernández, W. Beek, M. A. Martínez-Prieto, and M. Arias. LOD-a-lot: A Queryable Dump of the LOD cloud. In *Proceedings of the International Semantic Web Conference*, pp. 75–83, 2017.

5 A. Ahmeti, S. Bala, F. J. Ekaputra, J. D. Fernández, E. Kiesling, A. Koller, J. Mendling, A. Musil, A. Polleres, P. R. Aryan, M. Sabou, A. Solti, and J. Musil. CitySPIN: Cyber-Physical Social Systems for City-wide Infrastructures. In *13th International Conference on Semantic Systems*, Posters and Demos track, 2017.

3.9 Scaling SPADE to “Big Streams”

Ashish Gehani (SRI – Menlo Park, US)

Joint work of Ashish Gehani, Hasanat Kazmi, Hassaan Irshad
Main reference Ashish Gehani, Hasanat Kazmi, Hassaan Irshad: “Scaling SPADE to "Big Provenance"”, in Proc. of the 8th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2016, Washington, D.C., USA, June 8-9, 2016., USENIX Association, 2016.
URL https://www.usenix.org/conference/tapp16/workshop-program/presentation/gehani

Knowledge of the provenance of data has many uses, but also imposes novel challenges. Video provenance facilitates fine-grained attribution [1]. Operating system provenance enables principled forensic analysis [3], identification of the source of Grid infections [6], and authenticity claims that span trust domains [5]. Since provenance metadata can become voluminous, organizing it [4] and securing it [2] can be challenging. This has motivated policy-based [7] and flexible middleware-supported [8] approaches.

SPADE [13] is SRI’s open source system for managing provenance metadata. It has served as the research framework for exploring a range of ideas. These include automating the capture of application-level provenance through compiler instrumentation [15], accelerating distributed provenance queries via the use of sketches [12], optimizing the re-execution of workflows [11], diagnosing problems in mobile applications [10], vetting application for sensitive data flows [16], and studying tradeoffs in byte-, function-, and system-call level provenance tracking [14].

Most systems that track data provenance in distributed environments opt to collect it centrally giving rise to *“big streams”*. SPADE stores provenance at the host that it originated from, thereby decomposing a single big stream into multiple, concurrent ones. It avoids reconstruction of the entire global stream by introducing *coordination points* between independent provenance streams. In SPADE’s terminology, these are called *network artifacts* since they can be computed independently on each host using attributes of incoming and outgoing network flows. At query time, SPADE extracts relevant subsets from each host’s provenance stream. Using the network artifacts, these are stitched together into a single response that corresponds to the appropriate subset of the global provenance stream.

Since each host gives rise to a big stream of provenance, scaling continues to pose a challenge. Three strategies are adopted to ameliorate the issues that arise [9]. (1) Though SPADE allows a provenance stream to be abstracted online through the use of *filters*, the approach cannot be employed when details must be retained – in the case of forensic analysis, for example. Instead responses to big stream queries are rewritten with *transformers* to provide more comprehensible answers. (2) Merging auxiliary information into a big stream is problematic when large integration windows are needed. If the stream schema allow, content-based integration can be adopted to address this. (3) Deduplication of elements in a stream requires memory linear in the history’s size. For big streams, this cost becomes prohibitive. Leveraging persistent storage can address this at the cost of performance. A hybrid approach that combines caching and Bloom filters is developed to screen out duplicates with high probability.

References

1 Ashish Gehani and Ulf Lindqvist, **VEIL: A System for Certifying Video Provenance**, *9th IEEE International Symposium on Multimedia (ISM)*, 2007.

2 Ashish Gehani and Ulf Lindqvist, **Bonsai: Balanced Lineage Authentication**, *23rd Annual Computer Security Applications Conference (ACSAC), IEEE Computer Society*, 2007.

3 Ashish Gehani, Florent Kirchner, and Natarajan Shankar, **System Support for Forensic Inference**, *5th IFIP International Conference on Digital Forensics*, 2009.

4 Ashish Gehani, Minyoung Kim, and Jian Zhang, **Steps Toward Managing Lineage Metadata in Grid Clusters**, *1st Workshop on the Theory and Practice of Provenance (TaPP)* affiliated with the *7th USENIX Conference on File and Storage Technologies (FAST)*, 2009.

5 Ashish Gehani and Minyoung Kim, **Mendel: Efficiently Verifying the Lineage of Data Modified in Multiple Trust Domains**, *19th ACM International Symposium on High Performance Distributed Computing (HPDC)*, 2010.

6 Ashish Gehani, Basim Baig, Salman Mahmood, Dawood Tariq, and Fareed Zaffar, **Fine-Grained Tracking of Grid Infections**, *11th ACM/IEEE International Conference on Grid Computing (GRID)*, 2010.

7 Ashish Gehani, Dawood Tariq, Basim Baig, and Tanu Malik, **Policy-Based Integration of Provenance Metadata**, *12th IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY)*, 2011.

8 Ashish Gehani and Dawood Tariq, **SPADE: Support for Provenance Auditing in Distributed Environments**, *13th ACM/IFIP/USENIX International Conference on Middleware*, 2012.

9 Ashish Gehani, Hasanat Kazmi, and Hassaan Irshad, **Scaling SPADE to "Big Provenance"**, *8th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2016.

10 Nathaniel Husted, Sharjeel Qureshi, Dawood Tariq, and Ashish Gehani, **Android Provenance: Diagnosing Device Disorders**, *5th USENIX Workshop on the Theory and Practice of Provenance (TaPP)* affiliated with the *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2013.

11 Hasnain Lakhani, Rashid Tahir, Azeem Aqil, Fareed Zaffar, Dawood Tariq, and Ashish Gehani, **Optimized Rollback and Re-computation**, *46th IEEE Hawaii International Conference on Systems Science (HICSS)*, IEEE Computer Society, 2013.

12 Tanu Malik, Ashish Gehani, Dawood Tariq, and Fareed Zaffar, **Sketching Distributed Data Provenance**, *Data Provenance and Data Management for eScience, Studies in Computational Intelligence*, Vol. 426, Springer, 2013.

13 Support for Provenance Auditing in Distributed Environments, http://spade.csl.sri.com

14 Manolis Stamatogiannakis, Hasanat Kazmi, Hashim Sharif, Remco Vermeulen, Ashish Gehani, Herbert Bos, and Paul Groth, **Tradeoffs in Automatic Provenance Capture**, *Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science*, Vol. 9672, Springer, 2016.

15 Dawood Tariq, Maisem Ali, and Ashish Gehani, **Towards Automated Collection of Application-Level Data Provenance**, *4th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2012.

16 Chao Yang, Guangliang Yang, Ashish Gehani, Vinod Yegneswaran, Dawood Tariq, and Guofei Gu, **Using Provenance Patterns to Vet Sensitive Behaviors in Android Apps**, *11th International Conference on Security and Privacy in Communication Networks (SecureComm)*, 2015.

3.10 Benchmarking Enterprise Stream Processing Architectures

Günter Hesse (Hasso-Plattner-Institut – Potsdam, DE)

Main reference Guenter Hesse, Christoph Matthies, Benjamin Reissaus, Matthias Uflacker: "A New Application Benchmark for Data Stream Processing Architectures in an Enterprise Context: Doctoral Symposium", in Proc. of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, pp. 359–362, ACM, 2017.
URL http://dx.doi.org/10.1145/3093742.3093902

Data stream processing systems have become increasingly popular tools for analyzing large amounts of data that are generated in short periods of time. This is the case, for example in Industry 4.0 and Internet of Things scenarios, where masses of sensor and log data are continuously produced. This information can be leveraged in order to develop advanced applications, e.g., in the area of predictive maintenance.

Analysis of such data streams can become even more valuable when it takes advantage of traditional enterprise data, i.e., data from business systems such as an ERP system. Combining this transactional data with streaming data can unveil new insights with respect to processes and causal relations.

In recent years, many new data stream processing systems have been developed. Although a broad variety of systems allows for more choice, choosing the system that best suits a given use case becomes more difficult. Benchmarks are a common way of tackling this issue and allow a comprehensive comparison of different systems and setups. Currently, no suitable benchmark is available for comparing data stream processing architectures, especially when non-streaming enterprise data needs to be integrated.

With Senska, we will develop a new application benchmark for data stream processing architectures in an enterprise context that fills this gap.

3.11 Sliding-Window Aggregation in Worst-Case Constant Time

Martin Hirzel (IBM TJ Watson Research Center – Yorktown Heights, US)

Joint work of Kanat Tangwongsan, Martin Hirzel, Scott Schneider
Main reference Kanat Tangwongsan, Martin Hirzel, Scott Schneider: "Low-Latency Sliding-Window Aggregation in Worst-Case Constant Time", in Proc. of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, pp. 66–77, ACM, 2017.
URL http://dx.doi.org/10.1145/3093742.3093925

This talk briefly summarizes a paper with the same title that appeared at DEBS 2017 (International Conference on Distributed and Event-based Systems), where it won a best-paper award. Sliding-window aggregation is a widely-used approach for extracting insights from the most recent portion of a data stream. The aggregations of interest can usually be cast as binary operators that are associative, but they are not necessarily commutative nor invertible. Non-invertible operators, however, are difficult to support efficiently. The best published algorithms require O(log n) aggregation steps per window operation, where n is the sliding-window size at that point. This paper presents DABA, a novel algorithm for aggregating FIFO sliding windows using only O(1) aggregation steps per operation in the worst case (not just on average).

3.12 Tutorial: Stream Processing Languages

Martin Hirzel (IBM TJ Watson Research Center – Yorktown Heights, US)

This tutorial gives an overview of several styles of stream processing languages. The tutorial illustrates each style (relational, synchronous, explicit graph, etc.) with a representative example language. Of course, for each style, there is an entire family of languages, and this tutorial does not aim to be exhaustive. Overall, the field is diverse. Efforts to consolidate and standardize should be informed by an overview of the state of the art, which this tutorial provides.

3.13 Benchmarking Semantic Stream Processing Platforms for IoT Applications

Ali Intizar (National University of Ireland – Galway, IE)

With the growing popularity of Internet of Things (IoT) technologies and sensors deployment, more and more IoT-enabled applications are designed that can leverage the rich source of streaming data to gather knowledge, support data analytics and provide useful applications for end users such as smart city applications. Semantic Web and its underlying technologies are an ideal fit to support distributed heterogeneous applications designed over deployed sensors based infrastructure within smart cities. A merger of IoT and semantic Web has lead to the inception of RDF stream processing (RSP) and several RSP based streaming engines have been proposed. However, RSP technologies are still in their infancy and yet to be tested for their performance and scalability. Particularly for smart city applications IoT-enabled semantic solutions should be tested and benchmarked with the real deployments and infrastructure accessible within smart cities.

The Citybench Benchmark is a comprehensive benchmarking suite to evaluate RSP engines within smart city applications using real streaming data generated by smart cities. CityBench includes real-time IoT data streams generated from various sensors deployed within the city of Aarhus, Denmark. We provide a configurable testing infrastructure and a set of continuous queries covering a variety of data and application dependent characteristics and performance metrics, to be executed over RSP engines using CityBench datasets. This work can be used as a baseline to identify capabilities and limitations of existing RSP engines for smart city applications and provide support to smart city application developers to benchmark their applications.

3.14 Efficient evaluation of streaming queries comprising user-defined functions

Asterios Katsifodimos (TU Delft, NL)

Joint work of Asterios Katsifodimos, Paris Carbone, Jonas Traub, Volker Markl, Tilmann Rabl, Seif Haridi, Sebastian Bress
Main reference Paris Carbone, Jonas Traub, Asterios Katsifodimos, Seif Haridi, Volker Markl: "Cutty: Aggregate Sharing for User-Defined Windows", in Proc. of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, pp. 1201–1210, ACM, 2016.
URL http://dx.doi.org/10.1145/2983323.2983807

Aggregation queries on data streams are evaluated over evolving and often overlapping logical views called windows. While the aggregation of periodic windows were extensively studied in the past through the use of aggregate sharing techniques such as Panes and Pairs, little to no work has been put in optimizing the aggregation of very common, non-periodic windows. Typical examples of non-periodic windows are punctuations and sessions which can implement complex business logic and are often expressed as user-defined operators on platforms such as Google Dataflow or Apache Storm. The aggregation of such non-periodic or user-defined windows either falls back to expensive, best-effort aggregate sharing methods, or is not optimized at all. In my talk I presented a technique to perform efficient aggregate sharing for data stream windows, which are declared as user-defined functions (UDFs) and can contain arbitrary business logic. I introduced the concept of User-Defined Windows (UDWs), a simple, UDF-based programming abstraction that allows users to programmatically define custom windows. I then defined semantics for UDWs, based on which we designed Cutty [1], a low-cost aggregate sharing technique. I believe that user-defined windows is a noteworthy programming abstraction to be included in future stream programming languages.

On the other side of the spectrum, real-time sensor data enables diverse applications such as smart metering, traffic monitoring, and sport analysis. In the Internet of Things, billions of sensor nodes form a sensor cloud and offer data streams to analysis systems. However, it is impossible to transfer all available data with maximal frequencies to all applications. Therefore, we need to tailor data streams to the demand of applications. In recent work we contributed a technique that optimizes communication costs while maintaining the desired accuracy. Our technique [2] schedules reads across huge amounts of sensors based on the data-demands of a huge amount of concurrent queries. In the same spirit as Cutty, we introduce user-defined sampling functions that define the data-demand of queries and facilitate various adaptive sampling techniques, which decrease the amount of transferred data.

References

1 Paris Carbone, Jonas Traub, Asterios Katsifodimos, Seif Haridi, and Volker Markl. *Cutty: Aggregate sharing for user-defined windows*. In the Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1201–1210. ACM, 2016.

2 Jonas Traub, Sebastian Breß, Tilmann Rabl, Asterios Katsifodimos, and Volker Markl. *Optimized on-demand data streaming from sensor nodes*. In the Proceedings of the 2017 Symposium on Cloud Computing (SoCC '17). ACM, New York, NY, USA, 586–597.

3.15 Towards an effort for Adaptable Stream Processing Engines

Nikos Katsipoulakis (University of Pittsburgh, US)

Joint work of Nikos Katsipoulakis, Alexandros Labrinidis, Panos K. Chrysanthis

Online data flow processing requires that a Parallel Stream Processing Engine (pSPE) processes data by the time they become available and queries execute for a long period of time. Low operational cost and production of results in a timely fashion can be challenging goals for a pSPE, considering the volatile nature of streams. When the arrival rate of input data spikes, pSPEs need to be able to adjust their internal components, so that late delivery of results is avoided. This calls for adaptable pSPEs that can react to fluctuations in processing demands. This work aims to improve pSPEs' adaptability by revisiting internal components' design and load- balancing techniques. Adaptability is studied from the aspect of partitioning (Distribute), repartitioning (Divide), and load shedding (Drop). For each technique, the current state of the art is examined, both analytically and experimentally, and its shortcomings are exposed. In addition, new algorithms are proposed and this thesis' goal is (i) the investigation of decision-making algorithms, for selecting the best technique, (ii) the exploration of novel synergies among partitioning, repartitioning and load shedding.

References

1 N. R. Katsipoulakis, A. Labrinidis, and P. Chrysanthis, *A holistic view of stream partitioning costs.* PVLDB, pp. 1286–1297, 2017.

2 N. R. Katsipoulakis, C. Thoma, E. A. Gratta, et al., *Ce-storm: Confidential elastic processing of data streams.* ACM SIGMOD, pp. 859–864, 2015.

3 T. N. Pham, N. R. Katsipoulakis, P. K. Chrysanthis, and A. Labrinidis, *Uninterruptible migration of continuous queries without operator state migration.* SIGMOD Rec., vol. 46, no. 3, pp. 17–22, 2017.

3.16 Druid as an Example of a Federated Streaming Data System

Henning Kropp (Hortonworks – München, DE)

Streaming systems are often looked at as long running queries executed on an unbounded stream of data. In such a scenario a Federated Streaming Data System (FSDS) would execute such queries on heterogeneous and autonomous streaming systems.

In modern large enterprises heterogeneous streaming systems inevitable arise from different requirements, diverse system landscapes, evolving technology, and geographic distribution. FSDSs enable such companies to fully leverage the full potential of their streamed data. The success of unifying programming models like Apache Beam and Apache Calcite are a testimony for the potential value federated streaming data systems hold.

Further Druid (http://druid.io/) could be seen as an example of a federated streaming system. In it's architecture Duid distinguishes between a group of realtime and a group of historic nodes which use different kind of data access patterns to collect data from streams and answer queries. Combining the two types of node types Druid uses a catalog and query federation service, common to federated database systems, to transparently serve data from two different kind of systems. Although Druid is more like a streaming database then a streaming data system, it's architectures gives an idea of how a FSDS could look like.

3.17 Autonomous Semantic Stream Processing

Danh Le Phuoc (TU Berlin, DE)

Semantic Data Stream enables embedding computational semantics of contextual information and user/developer intentions into data stream generated from sensory data to structured (dynamic) knowledge base. By its intrinsic nature, semantic data stream sources are dynamically distributed in terms of spatial context, connectivity and distribution of the processing flow. To deal with this dynamicity, we propose the autonomous semantic stream processing model which sees semantic data streams as temporal RDF graphs (RDF streams). Via this model, distributed stream processing agents can autonomously coordinate their execution processes with their neighbour peers via exchanging control RDF streams among them. The control RDF streams enable a local RDF-based stream processing engine to continuously feed processing statuses from down stream processing agents such as data distribution, rate, available resources, processing capabilities and connectivity (network connections and link statuses, etc) to adaptively optimise its local processing pipelines. The adaptive optimisation of an processing agent can be done continuously in a collaboratively fashion with its neighbour peers which allow its to shift processing load and processing states to these peers. To realise the processing model, we built an autonomous processing kernel, called CQELS (Continuous Query Evaluation over Linked Stream) [1], which can run on small devices which collect sensor data as stream sources as well as on gateways with more processing power that can coordinate a small group of processing nodes. For dealing with big workload, the kernel can be also run as a cluster of powerful processing nodes on the Cloud. With our CQELS kernel, a distributed stream processing workflow is autonomously coordinated via its instances deployed in highly dynamic network topologies of heterogeneous processing nodes.

References

1 Phuoc, D.L., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A Native and Adaptive Approach for Unified Processing of Linked Streams and Linked Data. In: ISWC. pp. 370–388 (2011)

3.18 Collaborative Distributed Processing Pipelines (CDPPs)

Manfred Hauswirth

Joint work of Anja Feldmann, Manfred Hauswirth, Volker Markl
Main reference Anja Feldmann, Manfred Hauswirth, Volker Markl: "Enabling Wide Area Data Analytics with Collaborative Distributed Processing Pipelines (CDPPs)", in Proc. of the 37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017, Atlanta, GA, USA, June 5-8, 2017, pp. 1915–1918, IEEE Computer Society, 2017.
URL http://dx.doi.org/10.1109/ICDCS.2017.332

Novel concepts to organize the distribution and processing of information over the Internet in a secure and safe way which is scalable are needed. An important building block for achieving this goal are mobile edge clouds. A mobile edge cloud provides micro data centers which can move in the network according to load or other parameters and requirements. This is another form of virtualization which is transforming infrastructure everywhere, including network components, end-user devices, and eventually even sensors to transparently use any kind of resource. To work efficiently and scalably, mobile edge clouds need to be complemented by an appropriate world-wide network as well as sufficient backend computing power and storage. Combining all of this, we will see a seamless integration of network, storage, and computing in the future.

3.19 FlowDB: Integrating Stream Processing and Consistent State Management

Alessandro Margara (Polytechnic University of Milan, IT)

Joint work of Lorenzo Affetti, Alessandro Margara, Gianpaolo Cugola
Main reference Lorenzo Affetti, Alessandro Margara, Gianpaolo Cugola: "FlowDB: Integrating Stream Processing and Consistent State Management", in Proc. of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, pp. 134–145, ACM, 2017.
URL http://dx.doi.org/10.1145/3093742.3093929

Recent advances in stream processing technologies led to their adoption in many large companies, where they are becoming a core element in the data processing stack. In these settings, stream processors are often used in combination with various kinds of data management frameworks to build software architectures that combine data storage, processing, retrieval, and mining. However, the adoption of separate and heterogeneous subsystems makes these architectures overmuch complex, and this hinders the design, development, maintenance, and evolution of the overall system. In this talk, we propose a new model that integrates data management within a distributed stream processor. The model enables individual stream processing operators to persist data and make it visible and queryable from external components. It offers flexible mechanisms to control the consistency of data, including transactional updates plus ordering and integrity constraints. We implemented the model into the FlowDB prototype, and studied its overhead with respect to a pure stream processing system using real world case studies and synthetic workloads, proving the benefits of the proposed model by showing that FlowDB can outperform a state-of-the-art, in-memory distributed database in data management tasks.

3.20 Mining Big Data Streams – Better Algorithms or Faster Systems?

Gianmarco Morales (QCRI – Doha, QA)

Main reference Gianmarco De Francisci Morales, Albert Bifet: "SAMOA: scalable advanced massive online analysis", Journal of Machine Learning Research, Vol. 16, pp. 149–153, 2015.
URL http://dl.acm.org/citation.cfm?id=2789277

The rate at which the world produces data is growing steadily, thus creating ever larger streams of continuously evolving data. However, current (de-facto standard) solutions for big data analysis are not designed to mine evolving streams. So, should we find better algorithms to mine data streams, or should we focus on building faster systems?

In this talk, we debunk this false dichotomy between algorithms and systems, and we argue that the data mining and distributed systems community need to work together to bring about the next revolution in data analysis. In doing so, we introduce Apache SAMOA

(Scalable Advanced Massive Online Analysis), an open-source platform for mining big data streams (http://samoa.incubator.apache.org). Apache SAMOA provides a collection of distributed streaming algorithms for data mining tasks such as classification, regression, and clustering. It features a pluggable architecture that allows it to run on several distributed stream processing engines such as Storm, Flink, and Samza.

As a case study, we present one of SAMOA's main algorithms for classification, the Vertical Hoeffding Tree (VHT). Then, we analyze the algorithm from a distributed systems perspective, highlight the issue of load balancing, and describe a generalizable solution to it. Finally, we conclude by envisioning system-algorithm co-design as a promising direction for the future of big data analytics.

3.21 Data Streams in IoT Applications: Data Quality and Data Integration

Christoph Quix (Fraunhofer FIT – Sankt Augustin, DE)

Data integration is an open challenge that has been addressed in static data management for decades [1]. In data streams, the need of efficient data integration techniques is even more significant as the data has to be integrated while the stream is being processed. This means that the data can be processed only once and there should be not much overhead caused by the integration operations. Data quality issues also often arise in the context of data integration, as inconsistencies and incorrect values are revealed when several data sources are combined.

In data stream processing for Internet-of-Things (IoT) applications, the challenges of data integration and data quality are also very important as a network of interoperable devices and systems can only be established, if there are appropriate tools and techniques to integrate data and to verify the quality of the data. For example, in traffic applications, various data sources have to be combined for traffic state estimation or queue-end detection [3]. However, the techniques applied in this context are approximate techniques, such as data stream mining or map matching; thus, a result should always include a confidence value that indicates the quality. In industrial applications, the semantic heterogeneity of data, e.g., sensor data or data from ERP systems, requires a sophisticated semantic modeling of the data. In the context of Industry 4.0 applications, it is becoming less important in which factory a particular step of a manufacturing process is executed as the information of the production process needs to be efficiently exchanged along the value chain [2].

In this talk, we present our recent results on data integration and data quality management in data lakes [4, 5], and data streams [3].

References

1 Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, Johannes Gehrke, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, H. V. Jagadish, Donald Kossmann, Samuel Madden, Sharad Mehrotra, Tova Milo, Jeffrey F. Naughton, Raghu Ramakrishnan, Volker Markl, Christopher Olston, Beng Chin Ooi, Christopher Ré, Dan Suciu, Michael Stonebraker, Todd Walter, and Jennifer Widom. The beckman report on database research. *Commun. ACM*, 59(2):92–99, 2016.

2 Malte Brettel, Niklas Friederichsen, Michael Keller, and Marius Rosenberg. How Virtualization, Decentralization and Network Building Change the Manufacturing Landscape: An Industry 4.0 Perspective. *International Journal of Mechanical, Aerospace, Industrial and Mechatronics Engineering*, 8(1):37–44, 2014.

3 Sandra Geisler, Christoph Quix, Stefan Schiffer, and Matthias Jarke. An evaluation framework for traffic information systems based on data streams. *Transportation Research Part C*, 23:29–55, August 2012.

4 Rihan Hai, Sandra Geisler, and Christoph Quix. Constance: An intelligent data lake system. In Fatma Özcan, Georgia Koutrika, and Sam Madden, editors, *Proc. Intl. Conf. on Management of Data (SIGMOD)*, pages 2097–2100, San Francisco, CA, USA, 2016. ACM.

5 Matthias Jarke and Christoph Quix. On warehouses, lakes, and spaces: The changing role of conceptual modeling for data integration. In Jordi Cabot, Cristina Gómez, Oscar Pastor, Maria-Ribera Sancho, and Ernest Teniente, editors, *Conceptual Modeling Perspectives.*, pages 231–245. Springer, 2017.

3.22 Benchmarking Modern Streaming Systems

Tilmann Rabl (TU Berlin, DE)

Due to the recent trends of increasingly fast analysis of data, an increasing number of stream processing systems is built. Many of these include advanced features, such as a distributed architecture, different notions of time, and fault tolerance, with varying performance characteristics. These characteristics as well as basic stream processing operations pose specific challenges in benchmarking. In this talk, we identify several of these challenges, most notably the open world setup. In contrast to the closed world setup, where the streaming system under test has full control of the rate of incoming data, in a open world setup, the system has no influence on the data rate. While this is the typical setup of real deployments, there is no benchmark that properly tests this configuration. Our experiments demonstrate that current benchmarks fail to report correct latency measurements and overestimate throughput measures for real world setups.

3.23 Tutorial: Benchmarking

Tilmann Rabl (TU Berlin, DE)

In this tutorial, we cover why, when, and how to benchmark. Before introducing different types of benchmarks and standardized instantiations of those, we give a brief overview of performance estimation. After this, we give an overview of existing stream benchmarks and specific challenges when benchmarking streaming systems. Then, we exemplify the development process of modern industry standard benchmarks through TPCx-HS.

3.24 Collaborative Benchmarking of Computer Systems

Sherif Sakr (KSAU-HS – Riyadh, SA)

Joint work of Sherif Sakr, Amin Shafaat, Fuad Bajaber, Ahmed Barnawi, Omar Batarfi, Abdulrahman Altalhi
Main reference Sherif Sakr, Amin Shafaat, Fuad Bajaber, Ahmed Barnawi, Omar Batarfi, Abdulrahman H. Altalhi: "Liquid Benchmarking: A Platform for Democratizing the Performance Evaluation Process", in Proc. of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23-27, 2015., pp. 537–540, OpenProceedings.org, 2015.
URL http://dx.doi.org/10.5441/002/edbt.2015.52

Performances evaluation, reproducibility and benchmarking represent crucial aspects for assessing the practical impact of research results in the computer science

field. In spite of all the benefits (e.g., increasing impact, increasing visibility, improving the research quality) that can be gained from performing extensive experimental evaluation or providing reproducible software artifacts and detailed description of experimental setup, the required effort for achieving these goals remains prohibitive. In practice, conducting an independent, consistent and comprehensive performance evaluation and benchmarking is a very time and resource consuming process. As a result, the quality of published experimental results is usually limited and constrained by several factors such as: limited human power, limited time, or shortage of computing resources.

Liquid Benchmarking has been designed as an online and cloud-based platform for democratizing the performance evaluation and benchmarking processes. In particular, the platform facilitates the process of sharing the experimental artifacts (software implementations, datasets, computing resources, benchmarking tasks) as services where the end user can easily create, mashup, run the experiments and visualize the experimental results with zero installation or configuration efforts. In addition, the collaborative features of the platform enables the user to share and comment on the results of the conducted experiments so that it can guarantee a transparent scientific crediting process.

3.25 Low Latency Processing of Transactional Data Streams

Kai-Uwe Sattler (TU Ilmenau, DE)

Transactional database systems and data stream management systems have been thoroughly investigated over the past decades. While both system approaches follow completely different data processing models, i.e., push and pull based data forwarding, transactional stream processing combines both models. This means that stream queries writing to tables represent sequences of transactions and at the same time stream or batch queries on such tables get transaction isolation. In this talk, we present the PipeFabric framework – a lightweight publish-subscribe framework optimized for low-latency processing which comprises a library of operators for data stream processing including windows, aggregates, grouping, joins, CEP, matrix operations and a basic DSL for specifying dataflows. In addition to vectorized and data-parallel processing, PipeFabric provides support for tables as sinks and sources for streams as well as for maintaining persistent states of operators such as windows, aggregates, CEP or mining models. In this talk, we discuss challenges of persistent state management in low latency stream processing and sketch ideas for addressing these challenges.

3.26 Streams and Tables in Apache Kafka's Streams API

Matthias J. Sax (Confluent Inc – Palo Alto, US)

Apache Kafka introduces a novel approach for data stream processing compared to existing systems. Most state-of-the-art stream processing system use the abstraction of record- or fact-streams, that are append only sequences of immutable data items as first class citizen. Apache Kafka introduces a second type of data stream, called a changelog stream. A changelog stream is an append only sequence of updates and thus it models an evolving collection of data items. This evolving collection of data items can also be described as a table or materialized view and the individual records in the changelog stream are inserts/updates/deletes into this table. In contrast to a "static" database table, a Kafka table can be described as a continuously updating materialized view using the changelog topic as a kind of database redo log. This model opens a new processing paradigm for data stream processing and bridges the gap between static database tables and dynamic data streams. However, the current understanding of the duality is streams and tables is limited and we lack a semantically sound model that allows to define operator semantics that allows to reason about a computation or to do relational-style query optimization like operator reordering.

3.27 IoT Stream Processing Tutorial

Martin Strohbach (AGT International – Darmstadt, DE), Alexander Wiesmaier, and Arno Mittelbach

Main reference Vincenzo Gulisano, Zbigniew Jerzak, Roman Katerinenko, Martin Strohbach, Holger Ziekow: "The DEBS 2017 Grand Challenge", in Proc. of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, pp. 271–273, ACM, 2017.
URL http://dx.doi.org/10.1145/3093742.3096342

This tutorial covers the class of Internet of Things (IoT) streaming applications, i.e. applications that are concerned with interpreting and conceptualizing sensor data in real-time. It focuses on applications from two distinct domains. The first domain relates to Sports and Entertainment in which quantifiable insights about sports and entertainment events are created from sensors deployed at a venue. The second domain relates to Industry 4.0 in which sensor data from production machines is used to reduce energy costs and operations and maintenance costs.

The application examples are based on commercial deployments that run on top of AGT International's Internet of Things Analytics (IoTA) platform. IoTA is an IoT-based AI platform that provides cognitive and emotional computing skills to understand complex physical environments in real-time.

With the applications described for sports and entertainment we illustrate the specific characteristics of IoT streaming applications and the associated challenge of choosing an appropriate streaming infrastructure. This choice is influenced by the lack of standardized stream processing query languages, the multitude of available distributed streaming processing systems, required flexibility for a wide range of programming languages in which IoT analytics are being implmented, the focus on low latencies and a large number of shortlived processing pipelines, and the need to map processing results to a semantic data model.

We use the applications for Industry 4.0 to illustrate how stream processing applications can be benchmarked using the HOBBIT benchmarking platform. We describe applications in which we implemented solutions for reducing electricity costs for industrial customers by predicting energy peaks and applications in which we detect anomalies in machine states. We describe how we use the HOBBIT benchmarking platform in order to find anomalies from production machines in data streamed as RDF. The platform was used as part of this year's DEBS Grand Challenge.

3.28 Quantifying and Detecting Incidents in IoT Big Data Analytics

Hong-Linh Truong (TU Wien, AT)

Systems for IoT Big data analytics are extremely complex. Different software components at different software stacks from different infrastructures and providers are involved in handling different types of data. Various types of incidents may occur during execution of such a big data analytics due to problems occurring in software stacks, the data itself, and processing algorithms. Here incidents reflect unexpected situations that might happen within data themselves, machine learning algorithms, data pipelines, and underlying big data services and computing platforms. It is important to address any incident that prevents the pipeline running correctly or producing the expected quality of analytics. In this presentation, we show the motivation for quantifying, monitoring and analytics of incidents in IoT big data analytics systems and discuss our plan to tackle this important research.

3.29 Data Management of Big Spatio-temporal Data from Streaming and Archival Sources

Akrivi Vlachou (University of Thessaly – Lamia, GR)

An ever-increasing number of critical applications generate, collect, manage and process spatio-temporal data related to the mobility of entities in different domains. In this talk, an overview of the EU-funded project datAcron (http://www.datacron-project.eu/) is presented, which addresses time-critical mobility forecasting in maritime and aviation domains using Big Data analytics, focusing mainly on data management aspects. We describe a framework for semantic integration of big mobility data with other data sources, which is necessary for facilitating data analysis tasks, towards a unified representation of such data. First, data transformation from a wide variety of heterogeneous streaming and archival sources to a common representation (RDF) is performed. This is a typical situation in the analysis of mobility data, such as maritime and aviation, where streaming position data of moving objects need to be associated with static information (such as crossing sectors, protected geographical areas, weather forecasts, etc.) in order to provide semantically enriched trajectories. Next, spatio-temporal link discovery between spatio-temporal entities from diverse data sources is performed. Finally, our framework supports RDF queries that combine historical, static and streaming data. In this talk, we present an overview of the functionality of our framework as well as the technical challenges that are posed.

3.30 Stream Processing with Apache Apex

Thomas Weise (Mountain View, US)

Our environment is generating increasing volumes of data from a rapidly growing number of sources like mobile devices, sensors, industrial machines, web logs and more. Organizations are looking to convert these continuous streams of data into insights and competitive advantage, which requires systems that can process data at scale, with minimum delay and with accuracy.

The stream processing space has been evolving rapidly and adoption for real-world, business critical use cases is increasing. A new generation of systems supports large-scale, high-throughput, low-latency processing with correctness guarantees. Apache Apex, presented here, is one of these stream processing systems. The project started in 2012 with the vision to provide an alternative to MapReduce on Apache Hadoop YARN for low-latency processing. Originally a proprietary product, the project was open sourced as Apache Software Foundation project in 2015. Apex has been one of the innovators in the stream processing space with features such as distributed checkpointing, dynamic resource allocation/scaling and dynamic modification of the processing graph.

3.31 Lifetime-Based Memory Management in Distributed Data Processing Systems

Yongluan Zhou

Joint work of Lu Lu, Xuanhua Shi, Yongluan Zhou, Xiong Zhang, Hai Jin, Cheng Pei, Ligang He, Yuanzhen Geng
Main reference Lu Lu, Xuanhua Shi, Yongluan Zhou, Xiong Zhang, Hai Jin, Cheng Pei, Ligang He, Yuanzhen Geng: "Lifetime-Based Memory Management for Distributed Data Processing Systems", PVLDB 9(12): 936–947, 2016.
URL http://www.vldb.org/pvldb/vol9/p936-lu.pdf

In-memory caching of intermediate data and eager combining of data in shuffle buffers have been shown to be very effective in minimizing the re-computation and I/O cost in distributed data processing systems like Spark and Flink. However, it has also been widely reported that these techniques would create a large amount of long-living data objects in the heap, which may quickly saturate the garbage collector, especially when handling a large dataset, and hence would limit the scalability of the system. To eliminate this problem, we propose a lifetime-based memory management framework, which, by automatically analyzing the user-defined functions and data types, obtains the expected lifetime of the data objects, and then allocates and releases memory space accordingly to minimize the garbage collection overhead. In particular, we present Deca, a concrete implementation of our proposal on top of Spark, which transparently decomposes and groups objects with similar lifetimes into byte arrays and releases their space altogether when their lifetimes come to an end. An extensive experimental study using both synthetic and real datasets shows that, in comparing to Spark, Deca is able to 1) reduce the garbage collection time by up to 99.9%, 2) to achieve up to 22.7x speed up in terms of execution time in cases without data spilling and 41.6x speedup in cases with data spilling, and 3) to consume up to 46.6% less memory.

4 Working groups

4.1 Working Group: Languages and Abstractions

Martin Hirzel (IBM TJ Watson Research Center – Yorktown Heights, US)

Based on the definitions and survey from the corresponding tutorial, this working group identified and described three challenges faced by stream processing languages.

Variety of data models is a challenge for stream processing languages. A data model organizes elements of data with respect to their semantics, their logical composition into data structures, and their physical representation. Producers and consumers of streams to and from a streaming application dictate data models it must handle, and the application's own conversion and processing needs drive additional data-model variety. There is no consensus on what a stream data item is. At one extreme, in StreamIt, each data item is a simple number, while at the other extreme, C-SPARQL streams entire self-describing graphs. Streaming languages have so far failed to consolidate on a data model because data-model variety is a difficult challenge.

Data-model variety causes streaming-specific issues, since the data model affects the speed of serialization, transmission, compression, and dynamic checks for the presence or absence of certain fields, and because the online setting leaves no time for separate batch data integration. Some stream processing languages are designed around their data model, e.g., CQL on tuples or path expressions on XML trees. Furthermore, the data model enables streaming-language compilers to provide helpful error messages and optimizations.

The goal is for streaming languages to let the programmer use the logical data model they find most convenient while letting the compiler choose the best physical representation. Metrics of success are the expressive power of the language along with its throughput, latency, and resource consumption.

Veracity with simplicity is a challenge for stream processing languages. Veracity means producing accurate and factual results, and simplicity means avoiding unnecessary language complexity. There are several reasons why streaming veracity is hard. Sensors producing input data have limited precision, energy, and memory. In long-running and loosely-coupled streaming applications, sources come and go. And approximate stream algorithms and stream mining introduce additional uncertainty. This is compounded by the lack of ground truth in an online setting, and by the difficulty of anticipating and testing every eventuality.

Veracity causes streaming-specific issues, since it requires accurate real-time responses without having seen all the data, and because the online setting leaves no time for separate batch data cleansing. Also, streaming is often used in a distributed setting, where there can be no global clock. Some streaming languages are explorations in handling uncertainty on top of stream-relational algebra, but restricting stream operators to support retraction or uncertainty propagation limits expressiveness and raises complexity. A more general solution might use probabilistic programming to handle uncertainty in a principled way.

The goal is for streaming languages to help minimize compounding uncertainty by being quality-aware and adaptive while remaining simple, expressive, and fast. This inherently leads to multiple metrics (e.g., precision, recall, throughput, latency) and harder-to-quantify objectives (simplicity, expressiveness). One can maximize one set of metrics while satisficing a threshold on the others, or one can seek Pareto-optimal solutions.

Adoption is a challenge for stream processing languages: while there are many languages, none have reached broad acceptance and use. The community should care about adoption of streaming languages because it would drive adoption of streaming technologies in general. A widely-adopted language is more attractive for students to learn, leading to a bigger pool of skilled people to hire for companies. Furthermore, a widely-adopted language would drive more mature libraries, tools, benchmarks, and optimizations. The lack of a dominant language indicates that adoption is a difficult goal.

Streaming as a domain is young, fast-moving, and prone to vendor lock-in. At the same time, not only is there no consensus on a streaming language, there is not even a consensus on which language features are most important and which can be omitted to reduce complexity. Furthermore, several recent streaming systems have a DSEL (domain-specific embedded language), which tends to have less well-isolated semantics and more host-language dependencies than a stand-alone DSL (domain-specific language).

The goal is for the community to agree upon one or a few languages that get widely adopted. Metrics for language adoption include lines of code, as well as mentions in resumes, job posting, courses, and support forums. Adoption can also be measured by the number of systems that support a language, open-source and open-governance implementations, and ultimately, an industry standard.

We hope this summary of our working group discussion helps guide future streaming-language research in novel and impactful directions.

4.2 Working Group: Systems and Applications

Tilmann Rabl (TU Berlin, DE)

In this working group, participants discussed characteristics and open challenges of stream processing systems. The discussions mainly focused on the topics state management, transactions, and pushing computation to the edge.

State management – Modern streaming systems are stateful, which means they can remember the state of the stream to some extent. A simple example is a counting operator that counts the number of elements seen so far. While even simple state like this poses several challenges in streaming setups (such as fault tolerance and consistency), many use cases call for more advanced state management capabilities. An example is the combination of streaming and batch data. This is for example required when combining the history of a user with her current activity or when finding matching advertisement campaigns with current activity a popular example of such a setup is modeled in the Yahoo! Streaming Benchmark [2]. Today, most setups deal with such challenges by combining different systems (e.g., a key value store for state and a streaming system for processing), however, it is desirable to have both in a single system for consistency and manageability reasons.

State can be considered the equivalent of a table in a database system. As a result, besides the combination of stream and state, several high level operations can be identified: conversion of streams to tables (e.g., storing a stream), conversion of tables to streams (e.g., scanning a table), as well operations only on tables or streams (joins, filters, etc.). The management of state opens the design space in between existing stream processing systems and database systems, which has only been partially explored by current systems. In contrast to database systems, stream systems typically operate in a reactive manner, i.e., they have no control over the incoming data stream, specifically, they do not control and define the consistency and order semantics in the stream. This requires advanced notions of time and order as for example specified for streams in the dataflow model [1], for state and stream this remains an open field of research.

Transactions

A further discussion topic where transactions in stream processing systems. The main difference between traditional database transactions and stream processing transactions is that in databases the computation moves and data stays (in the system), in stream processing systems the computation stays and the data moves to the computation (and out again).

Considering state management, the form of transactions as applied in databases can also be used in a stream processing system, if the state is managed in a transactional way. However, the operations on streams themselves can be transactional and then we can differentiate between single tuple transactions and multiple tuple transactions. For multiple tuple transactions, the transaction can only commit when all tuples are consumed. The tuples then have to pass the whole operator graph or at least the transactional subgraph.

The semantics of transactions on streams is currently still an open field of research.

By *pushing computation to the edge* of a network, stream processing can be highly distributed and decentralized. This is very useful when preprocessing or filtering can be done without a centralized view of the data. Especially, in setups with high communication cost or slow connections (e.g., mobile connections), it makes sense to not send all data to a central server, but distribute the computation. A logical first step is filtering, but aggregations and even more complex operations can be pushed to the edge, if possible. Many modern scenarios prohibit centralized data storage, which further encourages distributed setups with early aggregations. Primary points of research are the declarativity for specifying highly distributed data processing programs and the architecture of systems to support these use cases.

Other topics discussed were ad hoc queries and graph stream processing. Most current systems only discuss long running queries, but in many use cases (e.g., sports, automotive) streams can be short lived as can be stream queries.

References

1 T. Akidau, R. Bradshaw, C. Chambers, S. Chernyak, R. J. Fernández-Moctezuma, R. Lax, S. McVeety, D. Mills, F. Perry, E. Schmidt, and S. Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8:1792–1803, 2015.

2 S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum, K. Patil, B. Peng, and P. Poulosky. Benchmarking streaming computation engines: Storm, flink and spark streaming. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1789–1792, 2016.

Participants

- Pramod Bhatotia
University of Edinburgh, GB
- Albert Bifet
Telecom ParisTech, FR
- Michael H. Böhlen
Universität Zürich, CH
- Angela Bonifati
University Claude Bernard – Lyon, FR
- Jean-Paul Calbimonte
HES-SO Valais – Sierre, CH
- Paris Carbone
KTH Royal Institute of Technology, SE
- Emanuele Della Valle
Polytechnic University of Milan, IT
- Javier D. Fernández-García
Wirtschaftsuniversität Wien, AT
- Ashish Gehani
SRI – Menlo Park, US
- Manfred Hauswirth
TU Berlin, DE
- Günter Hesse
Hasso-Plattner-Institut – Potsdam, DE
- Martin Hirzel
IBM TJ Watson Research Center – Yorktown Heights, US
- Ali Intizar
National University of Ireland – Galway, IE
- Asterios Katsifodimos
TU Delft, NL
- Nikos Katsipoulakis
University of Pittsburgh, US
- Henning Kropp
Hortonworks – München, DE
- Danh Le Phuoc
TU Berlin, DE
- Alessandro Margara
Polytechnic University of Milan, IT
- Gianmarco Morales
QCRI – Doha, QA
- Christoph Quix
Fraunhofer FIT – Sankt Augustin, DE
- Tilmann Rabl
TU Berlin, DE
- Sherif Sakr
KSAU-HS – Riyadh, SA
- Kai-Uwe Sattler
TU Ilmenau, DE
- Matthias J. Sax
Confluent Inc – Palo Alto, US
- Martin Strohbach
AGT International – Darmstadt, DE
- Hong-Linh Truong
TU Wien, AT
- Akrivi Vlachou
University of Thessaly – Lamia, GR
- Thomas Weise
Mountain View, US
- Yongluan Zhou
University of Copenhagen, DK

Report from Dagstuhl Perspectives Workshop 17442

Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP

Edited by

Nicola Ferro[1], Norbert Fuhr[2], Gregory Grefenstette[3], and Joseph A. Konstan[4]

1 **University of Padova, Italy** ferro@dei.unipd.it
2 **University of Duisburg-Essen, Germany** norbert.fuhr@uni-due.de
3 **Institute for Human Machine Cognition, USA** ggrefenstette@ihmc.us
4 **University of Minnesota, Minneapolis, USA** konstan@umn.edu

Abstract

This reports briefly describes the organization and the plenary talks given during the Dagstuhl Perspectives Workshop 17442. The goal of this workshop was to investigate the state-of-the-art and to delineate a roadmap and research challenges for performance modeling and prediction in three neighbour domains, namely information retrieval (IR), recommender systems (RecSys), and natural language processing (NLP).

Seminar 30 October–03 November, 2017 – www.dagstuhl.de/17442
1998 ACM Subject Classification H.3 Information Storage and Retrieval, I.2.7 Artificial Intelligence – Natural Language Processing
Keywords and phrases Information Systems, Formal models, Evaluation, Simulation, User Interaction
Digital Object Identifier 10.4230/DagRep.7.10.139

1 Executive Summary

Nicola Ferro
Norbert Fuhr
Gregory Grefenstette
Joseph A. Konstan

Information systems, which manage, access, extract and process non-structured information, typically deal with vague and implicit information needs, natural language and complex user tasks. Examples of such systems are information retrieval (IR) systems, recommender systems (RecSys), and applications of natural language processing (NLP) such as e.g. machine translation, document classification, sentiment analysis or search engines. The discipline behind these systems differs from other areas of computer science, and other fields of science and engineering in general, due to the lack of models that allow us to predict system performances in a specific operational context and to design systems ahead to achieve a desired level of effectiveness. In the type of information systems we want to look at, we deal with domains characterized by complex algorithms, dependent on many parameters and confronted with uncertainty both in the information to be processed and the needs to be

Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 139–146
Editors: Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

addressed, where the lack of predictive models is somehow bypassed by massive trials of as many combinations as possible.

These approaches relying on massive experimentation, construction of testbeds, and heuristics are neither indefinitely scaled as the complexity of systems and tasks increases nor applicable outside the context of big Internet companies, which still have the resources to cope with them.

The workshop was organized as follows. The first day was devoted to plenary talks focused on providing a general introduction to IR, RecSys, and NLP and on digging into some specific issues in performance modeling and prediction in these three domains. The second day, participants split into three groups – IR, RecSys, and NLP – and explored performance modeling and prediction issues and challenges within each domain; the working groups then reconvened to present the output of their discussion in a plenary session in order to cross-fertilize across disciplines and to identify cross-discipline themes to be further investigated. The third day, participant split into groups which explored these themes – namely measures, performance analysis, documenting and understanding assumptions, application features, and modeling performance – and reported back in plenary sessions to keep all the participants aligned with the ongoing discussions. The fourth and fifth days have been devoted to the drafting of this report and the manifesto originated from the workshop.

This documents reports the overview of the talks given by the participants on the first day. The outcomes of the working groups – both within-discipline themes and cross-discipline themes – as well as the identified research challenges and directions are presented in the Dagstuhl Manifesto corresponding to this Perspectives Workshop [1].

Acknowledgements. We thank Schloss Dagstuhl for hosting us.

References

1 N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. Manifesto from Dagstuhl Perspectives Workshop 17442 – Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP. *Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany*, 7(1), 2018.

2 Table of Contents

3 Overview of the Talks

3.1 The Validity Problems of IR

Norbert Fuhr (University of Duisburg-Essen, DE)

Current IR experiments often suffer from flaws that affect the internal validity, such as e.g. invalid or inappropriate metrics, poor test design, multiple testing without correction, or lack of reproducibility. External validity deals with the extent to which the findings of a study can be generalized. For addressing this issue, we must deepen out understanding of the models used, especially their underlying assumptions, and devise methods for checking these assumptions in a new setting. Furthermore, we need to investigate the relationship between application properties and performance, i.e. characteristics of the controlled variables (documents, topics and relevance assessments) of an IR experiment and the evaluation result.

3.2 Recommender Systems: The Evaluation Challenge

Joseph A. Konstan (University of Minnesota – Minneapolis, US)

Recommender systems have become ubiquitous, helping businesses market and users find desired information and products. They employ a variety of techniques including non-personalized summary statistics, content-based information filtering, and personalized collaborative filtering, often using latent-factor models based on or approximating matrix factorization. Evaluating recommender system performance is challenging because the most accessible measures such as predictive accuracy, rank performance, etc., all fail to capture the actual utility of the system–recommending items the user would not have selected anyway without the aid of the recommender. We review a variety of algorithms, offline and online evaluation metrics, and the challenge of effectively evaluating performance of recommender systems in the context of actual use.

3.3 Evaluation in Natural Language Processing

Gregory Grefenstette (Institute for Human Machine Cognition, US)

In this talk, I present the two main ways that Natural Language Processing (NLP) systems are evaluated. One way is calculating the improvement in some applications that use NLP processes to produce their results. Examples of these applications are Summarisation, Question Answering, Plagiarism Detection, Speech Recognition, Entity Extraction, Classification, Machine Translation, Author Identification, Image Labeling, Information Retrieval and Recommendation, among others. The second way is intrinsic evaluation of individual NLP modules, such as Language Identification, Tokenisation, Morphological Analysis, Part-of-Speech Tagging, Chunking, Shallow Parsing and Semantic Role Labelling, Deeper Parsing, Co-reference resolution, Topic Detection and Taxonomy/Thesaurus Extraction. We will explain how automated evaluation systems are set up, run and results reported, based upon gold standards and common metrics. For prediction, we will also describe some ways to characterize collections (used for training or testing). Finally, we will give an example of how much data is needed to produce expected results for analogy tests in word embeddings systems.

3.4 Bad for IR, Worse for Recommenders: Missing Data and the External Validity of Offline Evaluations

Michael D. Ekstrand (Boise State University, US)

Missing data impedes the realistic offline evaluation of information retrieval and recommender systems. Data sets do not have complete data on the relevance of items or documents to users or queries. The information retrieval community has developed several techniques that attempt to address these problems, but these techniques are not applicable to evaluating recommender systems due to the personalized and entirely subjective nature of relevance in recommender applications. Further, the nature of recommendation tasks and the subjectivity of relevance mean that this missing data is particularly detrimental to the validity of recommender evaluations. In this talk, I review the problem of missing data in information retrieval and recommendation tasks, the methods IR has developed, and explain why those methods are not suitable for evaluating recommenders. I also describe some additional concerns in recommender system evaluation that arise from missing data, and demonstrate that proposed solutions depend on missing theoretical knowledge or unrealistic assumptions.

3.5 Advanced Performance Modelling (and Prediction?) Techniques in IR

Nicola Ferro (University of Padova, IT)

Trying to explain the performance of a set of Information Retrieval (IR) systems across a set of topics is a preliminary step indispensable to start envisioning how to predict the performance of such systems. In this talk we discuss the different types of performance models which have been developed so far, which are all based on General Linear Mixed Models (GLMM) and ANalysis Of VAriance (ANOVA).

We start from the Topic and System effects models [1, 6]. We then consider the breakdown of the System effect into those of its components, namely stop lists, stemmers, and IR models [3, 4]. We discuss the use of simulation for showing the importance of the Topic*System interaction effect [5] as well as very recent work on using random partitions of the document corpus to estimate this effect [7]. Finally, we report on preliminary results about the Sub-Corpus effect and System*Sub-Corpus interaction effect [2].

We conclude by discussing how these explanatory models might be turned into predictive ones by using features describing these different factors and regression-like techniques.

References

1 D. Banks, P. Over, and N.-F. Zhang. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34, May 1999.

2 N. Ferro and M. Sanderson. Sub-corpora Impact on System Effectiveness. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 901–904. ACM Press, New York, USA, 2017.

3 N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel, editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 25–34. ACM Press, New York, USA, 2016.

4 N. Ferro and G. Silvello. Towards an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)*, 2017.

5 S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012.

6 J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.

7 E. M. Voorhees, D. Samarov, and I. Soboroff. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)*, 36(2):12:1–12:21, September 2017.

3.6 Objective or Subjective measures?

Martijn C. Willemsen (Eindhoven University of Technology, NL)

Recommenders are traditionally evaluated using offline evaluation on historical data. More recently, focus has shifted to online evaluation of objective behavioral data using AB testing. However, such behavior is hard to interpret without using subjective measures that help interpreting the meaning of the behavior. For example lower click-rates might not be reflecting reduced interest, but increased engagement of a user consuming the recommended content from beginning to end without additional interactions. In this talk I first introduce our user-centric evaluation framework [3] and subsequently show in three cases how objective and subjective measures go hand in hand in predicting and understanding user behavior and system effectiveness. The first case demonstrates how we can build a better prediction model for user segments based on subjective survey data of only 3000 users than on the behavioral data of all 100k users [2]. In the second case I show how objective measures of similarity, obscurity and accuracy can be linked to subjective perceptions of diversity, novelty and satisfaction. These subjective measures can explain the different relative preferences of users for three classical recommender algorithms (item-item, user-user and SVD) [1]. In the final case I show how choice difficulty of recommendation lists can be reduced by using latent-feature diversification, which reduces similarity between items while maintaining sufficient levels of attractiveness. The study shows that a diverse 5-item set is experienced as more satisfactory than a top-5 item set, despite the lower predicted accuracy of the list and the lower average rank of the items chosen by the user [4].

References

1 Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. 10.1145/2645710.2645737.

2 Mark P. Graus, Martijn C. Willemsen, and Kevin Swelsen. Understanding Real-Life Website Adaptations by Investigating the Relations Between User Behavior and User Experience. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless, editors, *User Modeling, Adaptation and Personalization*, number 9146 in Lecture Notes in Computer Science, pages 350–356. Springer International Publishing, June 2015. ISBN 978-3-319-20266-2 978-3-319-20267-9. URL http://link.springer.com/chapter/10.1007/978-3-319-20267-9_30.

3 Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, March 2012. ISSN 0924-1868, 1573-1391. 10.1007/s11257-011-9118-4.

4 Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, October 2016. ISSN 0924-1868, 1573-1391. 10.1007/s11257-016-9178-6.

3.7 User Utterance Understanding in Conversational Systems

Bernardo Magnini (Fondazione Bruno Kessler, IT)

In the context of the recent resurge of Artificial Intelligence, Conversational Agents have been attracting the attention of the NLP community. Conversational systems offer an interesting scenario for cross-domain predictability in NLP, for two reasons: (i) task oriented conversational agents are being developed in a huge numbers of application scenarios (e.g. virtual coaching, personal assistant, e-commerce, etc.) in different domains (e.g. food, sport) and for different languages; (ii) there are very few conversational datasets available for training models. In this context predictability is crucial for successfully develop high quality conversational systems. However, it opens several fundamental research questions. Which are the characteristics of the language (e.g. specific terminology, typical conversational patterns) of a certain domain that mainly affect the system performance? Which are the relevant characteristics of the application domain (e.g. complexity of entities and properties)? Which are the characteristics of the task (i.e. the problem to be solved by conversation, like booking a restaurant, or recommending a book)? How these three levels are related one with the other to determine predictability?

Participants

- Pablo Castells
Autonomous University of Madrid, ES
- Elizabeth M. Daly
IBM Research – Dublin, IE
- Thierry Declerck
DFKI – Saarbrücken, DE
- Michael D. Ekstrand
Boise State University, US
- Nicola Ferro
University of Padova, IT
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Werner Geyer
IBM TJ Watson Research Center – Cambridge, US
- Julio Gonzalo
UNED – Madrid, ES
- Gregory Grefenstette
IHMC – Paris, FR
- Joseph Konstan
University of Minnesota – Minneapolis, US
- Tsvi Kuflik
Haifa University, IL
- Krister Lindén
University of Helsinki, FI
- Bernardo Magnini
Bruno Kessler Foundation – Trento, IT
- Jian-Yun Nie
University of Montréal, CA
- Raffaele Perego
CNR – Pisa, IT
- Bracha Shapira
Ben Gurion University – Beer Sheva, IL
- Ian Soboroff
NIST – Gaithersburg, US
- Nava Tintarev
TU Delft, NL
- Karin Verspoor
The University of Melbourne, AU
- Martijn Willemsen
TU Eindhoven, NL
- Justin Zobel
The University of Melbourne, AU