



Volume 6, Issue 8, August 2016

Coding Theory in the Time of Big Data (Dagstuhl Seminar 16321) <i>Martin Bossert, Eimear Byrne, and Emina Soljanin</i>	1
Integrating Process-Oriented and Event-Based Systems (Dagstuhl Seminar 16341) <i>David Eyers, Avigdor Gal, Hans-Arno Jacobsen, and Matthias Weidlich</i>	21
Foundations of Secure Scaling (Dagstuhl Seminar 16342) <i>Lejla Batina, Swarup Bhunia, and Patrick Schaumont</i>	65
Next Generation Sequencing – Algorithms, and Software For Biomedical Applications (Dagstuhl Seminar 16351) <i>Gene Myers, Mihai Pop, Knut Reinert, and Tandy Warnow</i>	91

ISSN 2192-5283

Published online and open access by

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

Publication date

January, 2017

Bibliographic information published by the Deutsche Nationalbibliothek

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

License

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



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

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

The copyright is retained by the corresponding authors.

Aims and Scope

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

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

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

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

Editorial Board

- Gilles Barthe
- Bernd Becker
- Stephan Diehl
- Hans Hagen
- Hannes Hartenstein
- Oliver Kohlbacher
- Stephan Merz
- Bernhard Mitschang
- Bernhard Nebel
- Bernt Schiele
- Nicole Schweikardt
- Raimund Seidel (*Editor-in-Chief*)
- Arjen P. de Vries
- Klaus Wehrle
- Reinhard Wilhelm

Editorial Office

Marc Herbstritt (*Managing Editor*)
Jutka Gasiórowski (*Editorial Assistance*)
Dagmar Glaser (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)

Contact

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

Digital Object Identifier: 10.4230/DagRep.6.8.i

Coding Theory in the Time of Big Data

Edited by

Martin Bossert¹, Eimear Byrne², and Emina Soljanin³

¹ Universität Ulm, DE, martin.bossert@uni-ulm.de

² University College Dublin, IE, ebyrne@ucd.ie

³ Rutgers University – Piscataway, US, emina.soljanin@rutgers.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16321 “Coding Theory in the Time of Big Data”. The overarching technical theme was on how fundamentals of coding theory could be applied to data storage and transmission in the context of big data and conversely, on new problems in coding theory arising from such applications.

Seminar August 7–12, 2016 – <http://www.dagstuhl.de/16321>

1998 ACM Subject Classification E.4 Coding and Information Theory, E.3 Data Encryption, F.2 Analysis of Algorithms and Problem Complexity

Keywords and phrases Algebraic coding theory, Caching problems, Coding theory, Complexity theory, Cryptography, Distributed storage, Error-correction, Index coding, Information theory, Randomized algorithms, Streaming algorithms

Digital Object Identifier 10.4230/DagRep.6.8.1

Edited in cooperation with Allison Beemer and Carolyn Mayer

1 Executive Summary

Eimear Byrne

Martin Bossert

Emina Soljanin

License © Creative Commons BY 3.0 Unported license
© Eimear Byrne, Martin Bossert, and Emina Soljanin

The Dagstuhl Seminar 16321 *Coding Theory in the Time of Big Data*, held in August 7–12, 2016, was the third of a series of Dagstuhl seminars relating modern aspects of coding theory and its applications in computer science. The overarching technical theme was on how fundamentals of coding theory could be applied to data storage and transmission in the context of big data and conversely, on emerging topics in coding theory arising from such applications. In Dagstuhl Seminar 11461 the main topics discussed were list decoding, codes on graphs, network coding and the relations between them. The themes of distributed storage, network coding and polar codes were central to Dagstuhl Seminar 13351.

The conference was organised into six main working groups, as listed below:

1. Distributed Storage & Index Coding,
2. Private Information Retrieval for Storage Codes,
3. DNA-Based Storage,
4. Age & Delay of Information.
5. Code-Based Cryptography,
6. Rank-Metric Codes.



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

Coding Theory in the Time of Big Data, *Dagstuhl Reports*, Vol. 6, Issue 8, pp. 1–20

Editors: Martin Bossert, Eimear Byrne, and Emina Soljanin



Dagstuhl Reports

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

The amount of data that is being stored is scaling at a rapid pace making efficient data storage an important problem that inspires several lines of scientific research. During the seminar, several discussions were conducted on the theme of using classical and new techniques from coding theory to store/compute data efficiently in distributed storage systems. A number of open problems were identified, such as the design of codes with optimal repair bandwidth, fundamental trade-offs between storage & communication cost, applications to content distribution networks, connections between fundamental limits of storage/caching and the index coding problem and applications of coding theory for parallel computing. A theoretical framework and numerical simulation for the long term reliability of a distributed storage system were presented by Luby.

DNA-based storage was recently proposed to address new challenges to handle extremely high volume recording media to propose new compression methods for non-traditional data formats. Since DNA may be easily replicated and a massive amount of information stored reliably with minimal space requirements, it has enormous potential as a method of big data storage. This was the focus of the DNA working group. Problems such read and write cost, insertion and deletion errors arising in sequences, error reduction were discussed. Milenkovic gave an introductory talk describing several problems associated with whole genome, sequencing read, RNA-seq and ChIP-seq data compression, and outlined the first portable DNA-based rewritable and random access storage system.

Private information retrieval (PIR) enables a user to retrieve a data item from a database without disclosing the identity of the item retrieved, while the data itself may be public. The PIR working group considered this problem in the context of storage codes, in particular for dynamic coded storage and adversarial PIR, with some extensions to asynchronized systems, batch codes and private keyword search. Hollanti gave a tutorial overview of recent results in the area.

Age of information is a metric for status updating systems, where a monitor is interested in staying timely about the status of a source. The optimal updating strategy that minimizes the average age exists when the updating rate is constrained by limited network resources. Streaming source coding problems can be applied to the problem of age analysis. The main focus the Age & Delay working group was to introduce the age of information concept to participating coding theorists and explore potential age and delay problems in coding and storage. An adaptive arithmetic coding scheme was proposed as a potential solution to avoid huge decoding delay. Several possible delay problems in file downloading from multiple servers were discussed. Two PhD students, Zhong and Najm gave a tutorial overview of the topic.

Code-based crypto-systems are some of the very few that resist quantum-based attacks. In the case subfield subcodes such as the Goppa or Srivastava codes no successful attack is known yet. Moderate-density parity-check (MDPC) codes have been proposed for key size reduction in such cryptosystems. The group identified open problems such as investigating other subfield subcodes and attacks on MDPC structured codes. An overview was presented by Bossert.

Rank-metric codes have applications in random network coding, coded-caching and in code-based cryptography. The working group focussed on maximum rank distance (MRD) codes, specifically their classifications and on algebraic methods for constructing and decoding families of them. New nontrivial classifications were obtained. Further research directions on the classification problem were identified such as adapting semi-field theory techniques and searches for codes with high symmetry. Given the known limitation of list decoders for Gabidulin codes, the group worked on adapting decoders for Gabidulin codes to recent

families of MRD codes. Sheekey presented recent results on MRD codes and described links to semifields.

A total of 44 researchers participated in the seminar across these working groups. In addition, several participants took the opportunity to collaborate with others on specific related projects. There were 16 talks in total, several related to storage of big data and others on topics such as maximum rank distance codes, chip-to-chip communication, the MDS conjecture, the SAGE computer algebra system, age of information, the edge removal problem, convolutional codes and network coding. Among the talks given were some tutorial presentations, aimed at introducing researchers to fundamentals of a related working group. The working groups focussed on identifying and addressing new and/or important open problems in the area. Age & Delay, PIR for storage codes and DNA-based storage were new topics to many participants and generated considerable interest.

2 Table of Contents**Executive Summary**

<i>Eimear Byrne, Martin Bossert, and Emina Soljanin</i>	1
---	---

Overview of Talks

An Explicit, Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and All-Node Repair <i>Vijay Kumar</i>	6
A Coding Theory Inspired Approach to Computing Large Linear Transforms in Parallel/Distributed Systems <i>Viveck Cadambe</i>	6
Coding and Distributed Caching for Content Delivery <i>Alex Dimakis</i>	7
MDS Conjecture and the Projective Line <i>Iwan M. Duursma</i>	7
The Missing Link: An Introduction to the Edge Removal Problem <i>Michelle Effros</i>	7
Alphabet Size for Network Coding – Vectors Outperform Scalars <i>Tuvi Etzion</i>	8
An Algebraic Framework for Physical-Layer Network Coding <i>Elisa Gorla and Alberto Ravagnani</i>	8
Repairing Reed-Solomon Codes <i>Venkatesan Guruswami</i>	9
Private Information Retrieval from Coded Data in Distributed Storage Systems <i>Camilla Hollanti</i>	9
Linear Systems and Convolutional Codes <i>Julia Lieb</i>	10
A Mathematical Theory of Distributed Storage <i>Michael Luby</i>	10
DNA-Based Storage and Storing DNA <i>Olga Milenkovic</i>	11
SageMath for Research and Teaching in Coding Theory <i>Johan Rosenkilde</i>	11
Maximum Rank Distance Codes: Constructions, Classifications, and Applications <i>John Sheekey</i>	11
A Theory of Coding for Chip-to-Chip Communication <i>M. Amin Shokrollahi</i>	12
Age of Information <i>Jing Zhong and Elie Najm</i>	14

Working groups

Code-Based Cryptography
Martin Bossert 15

Distributed Storage & Index Coding
Viveck Cadambe 15

Private Information Retrieval for Storage Codes
Camilla Hollanti 17

DNA-Based Storage
Olgica Milenkovic 18

Rank-Metric Codes
John Sheekey 18

Age & Delay of Information
Jing Zhong 19

Participants 20

3 Overview of Talks

3.1 An Explicit, Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and All-Node Repair

P. Vijay Kumar (Indian Institute of Science -Bangalore, IN)

License © Creative Commons BY 3.0 Unported license

© Vijay Kumar

Joint work of Birenjith Sasidharan, Myna Vajha, P. Vijay Kumar

Main reference B. Sasidharan, M. Vajha, P. V. Kumar, “An Explicit, Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and All-Node Repair,” arXiv:1607.07335v3 [cs.IT], 2016.

URL <https://arxiv.org/abs/1607.07335v3>

This talk presents an explicit construction for an $((n, k, d = n - 1), (\alpha, \beta))$ regenerating code over a finite field of size Q operating at the Minimum Storage Regeneration (MSR) point. The MSR code can be constructed to have rate k/n as close to 1 as desired, sub-packetization given by $r^{n/r}$, for $r = (n - k)$, field size no larger than n and where all code symbols can be repaired with the same minimum data download. The construction modifies a prior construction by Sasidharan et al. which required far larger field-size. A building block appearing in the construction is a scalar MDS code of block length n . The code has a simple layered structure with coupling across layers, that allows both node repair and data recovery to be carried out by making multiple calls to a decoder for the scalar MDS code. The construction can be extended to handle the case of $d < (n - 1)$ under a mild restriction on the choice of helper nodes. While this work was carried out independently, there is considerable overlap with a prior construction by Ye and Barg. It is shown here that essentially the same architecture can be employed to construct MSR codes using vector binary MDS codes as building blocks in place of scalar MDS codes. The advantage here is that computations can now be carried out over a field of smaller size potentially even over the binary field as we demonstrate in an example.

3.2 A Coding Theory Inspired Approach to Computing Large Linear Transforms in Parallel/Distributed Systems

Viveck Cadambe (Pennsylvania State University – University Park, US)

License © Creative Commons BY 3.0 Unported license

© Viveck Cadambe

Joint work of Viveck Cadambe, Pulkit Grover, Sanghamitra Dutta

Main reference Sanghamitra Dutta, Viveck Cadambe, Pulkit Grover, “Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products,” to appear in Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS’16), 2016; pre-proceedings version available.


URL <https://papers.nips.cc/paper/6329-short-dot-computing-large-linear-transforms-distributedly-using-coded-short-dot-products>

In distributed and parallel computing, the performance of a computation is often limited by “stragglers”, a usually small set of processors that slow down the entire computation. Job replication has been recently proposed as a method for overcoming the straggler effect. In this talk, we will describe a simple, coding-theory inspired method to compute matrix-vector multiplications that is robust to stragglers in distributed systems. In particular, our method

performs a small set of carefully designed redundant computations such that the desired matrix-vector multiplication can be obtained from any sufficiently large subset of processors.

3.3 Coding and Distributed Caching for Content Delivery


Alex Dimakis (University of Texas – Austin, US)

License  Creative Commons BY 3.0 Unported license
© Alex Dimakis

Smartphone and tablet proliferation is generating an enormous increase in the demand for multimedia content. Modern wireless networks cannot support this demand and its large projected growth. We explain how caching of popular content can play a fundamental role in addressing this problem and how several novel mathematical and algorithmic problems arise. We focus on the Femtocaching problem and the Coded Caching problem introduced by Maddah-Ali and Niesen and discuss how caching is very promising for giving gains that scale surprisingly well in the size of the wireless system. Unfortunately, we show that for these gains to appear, the cached files must be separated in a number of blocks that scales exponentially in the number of users and files. We show how this problem can be resolved if we modify the Maddah-Ali and Niesen scheme to place and deliver coded packets in a less optimistic way.

3.4 MDS Conjecture and the Projective Line

Iwan M. Duursma (University of Illinois – Urbana Champaign, US)

License  Creative Commons BY 3.0 Unported license
© Iwan M. Duursma

Simeon Ball famously solved the MDS conjecture for prime fields and together with Jan de Beule extended the proof to codes in characteristic p with $k \leq 2p - 2$. Ameera Chowdhury formulated the proof in a combinatorial setting of inclusion matrices. The proofs rely on Segre's tangent lemma and polynomial interpolation. We replace these two main ingredients with a new simple duality relation for the projective line, which greatly simplifies the overall proof.

3.5 The Missing Link: An Introduction to the Edge Removal Problem

Michelle Effros (CalTech – Pasadena, US)


License  Creative Commons BY 3.0 Unported license
© Michelle Effros

In a world where massive networks carry massive quantities of data, it is surprising how little we know about even the most fundamental properties of the networks that we employ. This talk will introduce one such fundamental question: How much does a single link of some fixed capacity affect the capacity of the network in which it is employed? This simple question remains largely unsolved. Its solution turns out to be the key to many other fundamental questions whose solutions are also unknown. This talk will introduce the question and explore

both what is known about its solution and questions about reliability and data dependence to which it is linked.

3.6 Alphabet Size for Network Coding – Vectors Outperform Scalars

Tuvi Etzion (Technion – Haifa, IL)


License  Creative Commons BY 3.0 Unported license
© Tuvi Etzion

A survey on the known results on the alphabet size for solutions of multicast network is given. Vector network coding solutions based on rank-metric codes and subspace codes are considered. The main result of this paper is that vector solutions can significantly reduce the required field size compared to the optimal scalar linear solution for the same multicast network. The multicast networks considered in this paper have one source with h messages and the vector solution is over a field of size q with vectors of length t . The achieved gap of the field size between the optimal scalar linear solution and the vector solution is $q^{(h-2)t^2/h+o(t)}$ for any $q \geq 2$ and any even $h \geq 4$. If $h \geq 5$ is odd, then the achieved gap of the field size is $q^{(h-3)t^2/(h-1)+o(t)}$. Previously, only a gap of constant size had been shown for networks with a very large number of messages.

These results imply the same gap of the field size between the optimal scalar linear and any scalar *nonlinear* network coding solution for multicast networks. For three messages, we also show an advantage of vector network coding, while for two messages the problem remains open. Several networks are considered, all of them generalizations and modifications of the well-known combination network. The vector network codes that are used as a solution for those networks are based on subspace codes and in particular subspace codes obtained from rank-metric codes. Some of these codes form a new family of subspace codes which poses a new interesting research problem. Finally, the exposition given in this paper suggests a sequence of related problems for future research.

3.7 An Algebraic Framework for Physical-Layer Network Coding

Elisa Gorla (Université de Neuchâtel, CH) and Alberto Ravagnani (Université de Neuchâtel, CH)

License  Creative Commons BY 3.0 Unported license
© Elisa Gorla and Alberto Ravagnani

In this talk we will propose a new algebraic framework for physical-layer network coding. Our setup is based on nested-lattice-based physical-layer network coding as proposed by Nazer-Gastpar (2011), and on the algebraic approach proposed by Feng-Silva-Kschischang (2013) and Feng-Nobrega-Kschischang-Silva (2014). We will discuss the algebra which is used in our setup, and why our approach recovers and generalizes the previous ones.

3.8 Repairing Reed-Solomon Codes

Venkatesan Guruswami (Carnegie Mellon University – Pittsburgh, US)

License © Creative Commons BY 3.0 Unported license
© Venkatesan Guruswami

Joint work of Venkatesan Guruswami, Mary Wootters

Main reference V. Guruswami, M. Wootters, “Repairing Reed-Solomon Codes”, in Proc. of the 48th Annual ACM SIGACT Symp. on Theory of Computing (STOC’16), pp. 216–226, ACM, 2016; pre-print available as arXiv:1509.04764v2 [cs.IT].

URL <http://dx.doi.org/10.1145/2897518.2897525>

URL <https://arxiv.org/abs/1509.04764v2>

A fundamental fact about polynomial interpolation is that k evaluations of a degree- $(k - 1)$ polynomial f are sufficient to determine f . This is also necessary in a strong sense: given $k - 1$ evaluations, we learn nothing about the value of f on any k th point. Motivated by the exact repair problem for Reed-Solomon codes in distributed storage systems that are ubiquitous in the time of big data, we study a variant of the polynomial interpolation problem: instead of querying entire evaluations of f (which are elements of a large field F) to recover an unknown evaluation, we are allowed to query only a few bits of evaluations.

We show that in this model, one can do significantly better than in the traditional setting, in terms of the amount of information required to determine the missing evaluation. More precisely, only $O(k)$ bits are necessary to recover the missing evaluation, and this result is optimal for linear methods.

3.9 Private Information Retrieval from Coded Data in Distributed Storage Systems

Camilla Hollanti (Aalto University, FI)

License © Creative Commons BY 3.0 Unported license
© Camilla Hollanti

Private information retrieval (PIR) enables a user to retrieve a data item from a database without disclosing the identity of the item retrieved, while the data itself may be public. In this talk, I will give an introduction to PIR, concentrating in particular on recent advances in PIR from coded storage systems, where storage nodes may collude [1]. Some open problems will be shortly introduced, leaving further discussions to the PIR working group.

References

- 1 R. Tajeddine and S. El Rouayheb, “Private Information Retrieval from MDS Coded data in Distributed Storage Systems,” *2016 IEEE International Symposium on Information Theory (ISIT)*, Barcelona, 2016, pp. 1411–1415. Extended version available at <http://www.ece.iit.edu/~salim/PIRMDS.pdf>.

3.10 Linear Systems and Convolutional Codes

Julia Lieb (Universität Würzburg, DE)

License © Creative Commons BY 3.0 Unported license
© Julia Lieb

Main reference U. Helmke, J. Jordan, J. Lieb, “Probability estimates for reachability of linear systems defined over finite fields”, *Advances in Mathematics of Communications*, 10(1):63–78, 2016.

URL <http://dx.doi.org/10.3934/amc.2016.10.63>

Some properties such as reachability or observability are of considerable interest for dealing with linear systems. If one focuses on systems defined over finite fields, it becomes possible to count the number of systems with a certain property and therefore, to estimate probabilities [1]. Using a criterion for the reachability and observability of interconnected systems [2], we extend these probability estimations to networks of systems, e.g. parallel connection. Since there is a correspondence between linear systems and convolutional codes [3], these considerations could be transferred to interconnected convolutional codes. Hereby, the reachability and observability of the system correlate with the minimality and non-catastrophicity of the code.

References

- 1 U. Helmke, J. Jordan, J. Lieb. “Probability estimates for reachability of linear systems defined over finite fields”, *Advances in Mathematics of Communications*, Vol. 10, No. 1(2016), 63–78.
- 2 P. A. Fuhrmann, U. Helmke. “The Mathematics of Networks of Linear Systems,” Springer, 2015.
- 3 J. Rosenthal, E.V. York. “BCH convolutional codes,” *IEEE Trans. Inform. Theory*, Vol. 45, No. 6 (1999), 1833–1844.

3.11 A Mathematical Theory of Distributed Storage

Michael Luby (Qualcomm Inc. – San Diego, US)

License © Creative Commons BY 3.0 Unported license
© Michael Luby

We describe a natural and general model of distributed storage. A distributed storage system uses two fundamental mechanisms to ensure that stored objects remain recoverable as nodes fail and are replaced with new nodes: (1) storage overhead, i.e., the aggregate size of stored objects is less than the aggregate system storage capacity; (2) a repair mechanism, i.e., a repairer continually reads data stored at the nodes, performs computations on the read data, and writes the results of the computations back to the nodes.

We show lower bounds and upper bounds on trade-offs between storage overhead and repairer read rates (and write rates). The lower bounds are information-theoretic, i.e., we prove all repairers must operate above certain storage overhead/repairer read rate trade-offs. The upper bounds are algorithmic, i.e., we prove there is a repairer that operates below certain storage overhead/repairer read rate trade-offs. The lower and upper bound trade-offs are asymptotically equal as the storage overhead goes to zero.

3.12 DNA-Based Storage and Storing DNA

Olgica Milenkovic (University of Illinois – Urbana Champaign, US)

License © Creative Commons BY 3.0 Unported license
© Olgica Milenkovic

The surge of Big Data platforms has imposed new challenges to the storage community to identify extremely high volume recording media and to information theorists to propose new compression methods for nontraditional data formats. To address the first challenge, the new paradigm of DNA-based storage was recently proposed and implemented by a number of researchers. At the same time, to address the second challenge, several new initiatives for genomic read and RNA expression data compression were put forward by the National Institute of Health.

We describe several problems associated with whole genome, sequencing read, RNA-seq and ChIP-seq data compression, including parallel transform coding and large-alphabet source coding. We then proceed to outline the implementation of the first portable DNA-based rewritable and random access storage system. In this setting, we introduce new problems in prefix-synchronized, profile and Damerau-distance coding.

3.13 SageMath for Research and Teaching in Coding Theory

Johan Rosenkilde (Technical University of Denmark – Lyngby, DK)

License © Creative Commons BY 3.0 Unported license
© Johan Rosenkilde
Joint work of David Lucas, Daniel Augot, Clément Pernet, Johan Rosenkilde
URL <http://jsrn.dk/talks.html>

SageMath is a powerful open-source computer-algebra system. In excess of being open-source – a clear advantage and “ethical” feature for researchers – SageMath has user-friendly interfaces that support experimentation, such as the Jupyter Notebook and SageMathCloud. For Coding Theory, Sagemath has in recent years become vastly more powerful, mainly due to the ACTIS project, which employed David Lucas full-time for two years as software developer. We give a succinct demonstration of selected capabilities that SageMath now offers to the working and teaching coding theorist.

3.14 Maximum Rank Distance Codes: Constructions, Classifications, and Applications


John Sheekey (University College Dublin, IE)

License © Creative Commons BY 3.0 Unported license
© John Sheekey
Main reference J. Sheekey. “A new family of linear maximum rank distance codes”, *Advances in Mathematics of Communications*, 10(3):475–488, 2016.
URL <http://dx.doi.org/10.3934/amc.2016019>

We survey the known constructions and classifications of MRD codes. We illustrate their links with algebraic structures known as semifields and quasifields, and present some open problems.

3.15 A Theory of Coding for Chip-to-Chip Communication

M. Amin Shokrollahi (EPFL – Lausanne, CH)

License  Creative Commons BY 3.0 Unported license
© M. Amin Shokrollahi

Modern electronic devices consist of a multitude of IC components: the processor, the memory, the RF modem and the baseband chip (in wireless devices), and the graphics processor are only some examples of components scattered throughout a device. The increase of the volume of digital data that needs to be accessed and processed by such devices calls for ever faster communication between these IC's. Faster communication, however, often translates to higher susceptibility to various types of noise, and inevitably to a higher power consumption in order to combat the noise. This increase in power consumption is, for the most part, far from linear, and cannot be easily compensated for by Moore's Law. In this talk I will give a short overview of problems encountered in chip-to-chip communication, and will advocate the use of novel coding techniques to solve those problems. I will also briefly talk about Kandou Bus, and some of the approaches the company is taking to design, implement, and market such solutions.

References

- 1 A. Abbasfar, "Generalized differential vector signaling," Proc. of the ICC, pp. 1-5, 2009.
- 2 A. Abbasfar, "Simplified receiver for use in communication systems," U.S. patent no. 8,159,375.
- 3 A. Amirkhany, "Multi-carrier signaling for high speed electrical links," Ph.D. Thesis, Stanford University, 2008.
- 4 A. Amirkhany, A. Abbasfar, V. Stojanovic, and M. A. Horowitz, "Practical limits of multi-tone signaling over high-speed backplane electrical links," Proc. Of the ICC, pp. 2693–2698, 2007.
- 5 A. Amirkhany, K. Kaviani, A. Abbasfar, F. Shuaeb, W. Beyene, C. Hoshino, C. Madden, K. Chang, and C. Yuan, "A 4.1pJ/b 16Gb/s Coded Differential Bidirectional Parallel Electrical Link," ISSCC 2012, pp. 138–140.
- 6 A. Bechtolsheim, "Moore's law and networking," North American Network Operator's Group (NANOG) meeting, 2012. (<https://www.nanog.org/meetings/nanog55/presentations/Monday/Bechtolsheim.pdf>)
- 7 D. M. Chiarulli, J. D. Bakos, J. R. Martin, and S. P. Levitan, "Area, power, and pin efficient bus transceiver using multi-bit-differential signaling," IEEE International Symposium on Circuits and Systems, pp. 1662–1665, 2005.
- 8 H. Cronie and A. Shokrollahi, "Orthogonal differential vector signaling," U.S. Patent application no. 12/784414.
- 9 H. Cronie, A. Shokrollahi, and A. Tajalli, "Methods and systems for noise resilient, pin-efficient and low-power communications with sparse signaling codes," U.S. Patent no. 8,649,445.
- 10 K. Fukuda, H. Yamashita, G. Ono, R. Nemoto, N. Masuda, T. Takemoto, F. Yui, and T. Saito, "A 12.3-mW 12.5-Gb/s complete transceiver in 65-nm CMOS process," IEEE J. Solid State Circuits, 45, 2010.
- 11 K. Gharibdoust, A. Tajalli, and Y. Leblebici, "A 7.5 mW 7.5 Gb/s mixed NRZ/multi-tone serial-data transceiver for multi-drop memory interfaces in 40nm CMOS," ISSCC 2015, pp. 1–3, 2015.
- 12 M. Harwood et al., "A 12.5 Gb/s SerDes in 65nm CMOS using a Baud-Rate ADC with Digital Receiver Equalization and Clock Recovery," ISSCC 2007, pp. 436–439.

- 13 A. Healey and C. Morgan, "A comparison of 25 Gbps NRZ & PAM-4 Modulation used in legacy & premium backplane channels," DesignCon 2012.
- 14 A. Hormati, A. Shokrollahi, and R. Ulrich, "Method and apparatus for low power chip- to-chip communications with constrained ISI-ratio," U.S. Patent application no. 14/612241.
- 15 A. Hormati and A. Shokrollahi, "ISI tolerant signaling: a comparative study of PAM4 and ENRZ," DesignCon 2016.
- 16 A. Hormati, A. Tajalli, Ch. Walter, K. Gharibdoust, and A. Shokrollahi, "A Versatile Spectrum Shaping Scheme for Communicating Beyond Notches in Multi- Drop Interfaces," DesignCon 2016.
- 17 S. A. Ibrahim and B. Razavi, "Design requirements of 20-Gb/s serial links using multi-tone signaling," ISSCC 2009, pp. 1–4.
- 18 J. Lee, M. Chen, and H. Wang, "Design and Comparison of Three 20-Gb/s Backplane Transceivers for Duobinary, PAM4, and NRZ Data," JSSC, vol. 43, no. 9, 2008.
- 19 F. J. MacWilliams and N. J. A. Sloane. "The Theory of Error-Correcting Codes." North-Holland, 1988.
- 20 M. Mansuri, J. E. Jaussi, J. T. Kennedy, T. Hsueh, S. Shekhar, G. Balamurugan, F. O'Mahony, C. Roberts, R. Mooney, and B. Casper, "A scalable 0.128-to-1Tb/s 0.8- to-2.6pJ/b 64-lane parallel I/O in 32nm CMOS," ISSCC 2013, pp. 402–403.
- 21 A. Nazemi, Kangmin Hu, B. Catli, Delong Cui, U. Singh, T. He, Zhi Huang, Bo Zhang, A. Momtaz, and J. Cao, "A 36Gb/s PAM4 transmitter using an 8b 18GS/s DAC in 28nm CMOS," ISSCC 2015, pp. 58–60.
- 22 D. Oh, F. Ware, J.-H. Kim, A. Abbasfar, J. Wilson, L. Luo, R. Schmitt, and C. Yuan, "Pseudo- differential vector signaling for noise reduction in single-ended signaling systems," Designcon, 2009.
- 23 P. Orlik and H. Terao. Arrangements of Hyperplanes. Springer Verlag, 1992. Number 300 in Grundlehren der mathematischen Wissenschaften.
- 24 V. Parthasarathy, "PAM4 digital receiver performance and feasibility," IEEE 802.3bj meeting, 2012. (http://www.ieee802.org/3/bj/public/jan12/parthasarathy_01_0112.pdf)
- 25 D. V. Perino and J. B. Dillon, "Apparatus and method for multilevel signaling," U.S. Patent no. 6,005,895.
- 26 J. W. Poulton, S. Tell, and R. Palmer, "Multiwire differential signaling," University of North Carolina-Chapel Hill, 2003.
- 27 A. Shokrollahi, "Vector signaling codes with reduced receiver complexity," U.S. Patent application no. 14/313966.
- 28 A. Shokrollahi, "Vector signaling codes with high pin-efficiency and their applications to chip-to-chip communications and storage," U.S. Patent application no. 14/612252.
- 29 A. Shokrollahi and R. Ulrich, "Vector signaling codes with increased signal to noise characteristics," U.S. Patent application no. 62/015172.
- 30 A. Shokrollahi et al., "A Pin-Efficient 20.83 Gb/s/wire 0.94 pJ/bit Forwarded Clock CNRZ-5 coded Serial Link up to 12mm for MCM Packages in 28nmTechnology," ISSCC 2016.
- 31 A. Singh et al., "A pin- and power-efficient low-latency 8-to-12Gb/s/wire 8b8w- coded SerDes link for high-loss channels in 40nm technology," ISSCC 2014, pp. 442–443.
- 32 J. Poulton, W. J. Dally, , X. Chen, J. G. Eyles, Th. H. Greer, S. G. Tell, J. M. Wilson, and C. Th. Gray, "A 0.54 pJ/b 20 Gb/s ground-referenced single-ended short- reach serial link in 28 nm CMOS for advanced packaging applications," JSSC, vol. 48, pp. 3206–3218, 2013.
- 33 D. Slepian, "Permutation modulation," Proc. IEEE, vol. 53, pp. 228–236, 1965.
- 34 D. Stauffer, J. Trinko-Mechler, M. A. Sorna, K. Dramstad, C. R. Ogilvie, A. Mohammad, and J. D. Rockrohr. "High Speed SerDes Devices and Applications." Springer Verlag, 2009.
- 35 V. M. Stojanovic, A. Amirkhany, and J. Zerbe, "Multi-tone system with oversampled precoders," U.S. Patent no. 7,817,743.

- 36 A. Tajalli, H. Cronie, and A. Shokrollahi, “Methods and circuits for efficient processing and detection of balanced codes,” U.S. Patent no. 8,593,305.
- 37 Th. Toifl, C. Menolfi, M. Ruegg, R. Reutemann, P. Buchmann, M. Kossel, T. Morf, J. Weiss, and M. L. Schmatz, “A 22-Gb/s PAM-4 receiver in 90-nm CMOS SOI technology,” JSSC, vol. 41, pp. 954–965, 2006.
- 38 R. Ulrich, “Multilevel driver for high speed chip-to-chip communications,” U.S. patent application no. 14/315,306.
- 39 G. A. Wiley, “Three phase and polarity encoded serial interface,” U.S. Patent no. 8,472,551.
- 40 L. Yang and J. Armstrong, “Oversampling to reduce the effect of timing jitter on high speed OFDM systems,” IEEE Communication Letters, vol. 14, pp. 196–198, 2010.
- 41 S. Zogopoulos and W. Namgoong, “High-Speed Single-Ended Parallel Link Based on Three-Level Differential Encoding,” JSSC, Vol. 44, pp. 549–557, 2009.

3.16 Age of Information

Jing Zhong (Rutgers University – New Brunswick, US) and Elie Najm (EPFL – Lausanne, CH)

License © Creative Commons BY 3.0 Unported license
© Jing Zhong and Elie Najm

Age of information is a recently introduced timeliness metric for status updating systems, where a monitor is interested in staying timely about the status of the source which is connected to the monitor through some communication systems. In this talk, we will discuss the general methods to calculate the average age that can be applied to different service systems, and present several most recent results on age of information for a variety of queueing systems. The optimal updating strategy that minimizes the average age exists when the updating rate is constrained by limited network resources. We also connect age of information to coding theory by applying the age analysis to streaming source coding problems, where the receiver is required to decode the source message in a real-time fashion. Unlike the traditional source coding problem that focuses on the coding redundancy, the age-optimized source coding scheme balances the redundancy and higher moments of code lengths in order to minimize the queuing delay.

References

- 1 S. Kaul, R. D. Yates, and M. Gruteser, “Real-time status: How often should one update?” Proc. of IEEE INFOCOM, 2731–2735, 2012.
- 2 M. Costa, M. Codreanu and A. Ephremides, “Age of Information with Packet Management,” IEEE Int. Symp. on Info. Theory (ISIT), 1583–1587, 2014.
- 3 E. Najm, R. Nasser, “Age of Information: the Gamma Awakening,” IEEE Int. Symp. on Info. Theory (ISIT), 2574–2578, 2016.
- 4 J. Zhong, R. D. Yates, “Timeliness in Lossless Block Coding,” Data Compression Conf. (DCC), 2016.

4 Working groups

4.1 Code-Based Cryptography

Martin Bossert (Universität Ulm, DE)

License © Creative Commons BY 3.0 Unported license
© Martin Bossert

We discussed the status of coding-based cryptography. The McEliece and the Niederreiter cryptosystems are based on codes and remain the only methods. As of now these systems are some of the very few that resist quantum-based attacks. Lattice-based crypto is another such promising scheme. However, the usage of McEliece and Niederreiter with highly structured codes is broken. For example RS, RM, wild Goppa, LDPC, BCH, Gabidulin, and others cannot be used in these systems. The successful attacks are almost all based on the convolution theorem of the Fourier transform (Schur product, star product, etc.). However, the original Goppa code is not broken yet. It seems that for subfield subcodes like this Goppa code or for the Srivastava code no successful attack is known yet. Thus, an open problem is to check other subfield subcodes.

The concept of learning with errors is very similar to syndrome based methods and is viewed in the crypto community as provably secure. One recent approach is MDPC (moderate-density parity-check) codes with quasi-cyclic property for key size reduction (key size being one of the major criticisms of code-based cryptography). Possible attacks were discussed using the quasi-cyclicity and decoding by calculating low-weight codewords. Thus another open problem is to check possible attacks for MDPC codes. A further recent approach is to improve the permutation in the McEliece cryptosystem such that even structured codes can be used (not published yet).

A common argument against the use of code-based cryptography has been that signatures are not possible; however, this is not true since five methods are known and two of them are not broken yet ([1], [2]).

References

- 1 Nicolas T. Courtois, Matthieu Finiasz, and Nicolas Sendrier. “How to Achieve a McEliece-Based Digital Signature Scheme”, pages 157–174. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.
- 2 G. Kabatianskii, E. Krouk, and B. Smeets. “A Digital Signature Scheme Based on Random Error-Correcting Codes”, pages 161–167. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

4.2 Distributed Storage & Index Coding

Viveck Cadambe (Pennsylvania State University – University Park, US)

License © Creative Commons BY 3.0 Unported license
© Viveck Cadambe

In the big data era, the amount of data that is being stored is scaling at a rapid pace. This makes efficient data storage an important problem that inspires several lines of scientific research. During the seminar, several discussions were conducted on the theme of using classical and new techniques from coding theory to store/compute data efficiently in distributed storage systems. We summarize the discussions below.

1. Open Problems in Codes for Distributed Storage: Several open problems requiring design of new codes for storing data in distributed storage systems, and study of their information-theoretic fundamental limits were discussed. A list of specific open problems that are relevant to wide variety of applications was composed. Follow-up discussions included a summary of recent progress in some of the listed problems and related challenges. The topics discussed include:
 - a. design of codes with good/optimal repair bandwidth, and fundamental trade-offs between storage cost and communication cost,
 - b. fundamental limits of locality and availability of codes,
 - c. formulations that are applicable to caching and content distribution networks, and
 - d. connections between fundamental limits storage/caching and the index coding problem in network information theory.
2. A Mathematical Theory for Distributed Storage: Details related to a recently developed theoretical framework that studies the long term reliability of a distributed storage system were presented by its author (Michael Luby). The model, unlike others discussed in the working group, studies of the evolution of the storage system over time, estimating the likelihood of data loss. A numerical simulation of the mathematical model was also demonstrated. Follow up discussions included comparisons and connections between the exact and functional repair problem in coding theory and the presented framework.
3. Coding Theory Inspired Methods for Parallel/Distributed Computing: Mathematical details related to a coding theory inspired method for implementing linear transforms in parallel/distributed systems, were presented by its author (Viveck Cadambe). The utility of coding theory was to incorporate redundancy in the computations to make it robust to slow nodes (stragglers). The discussion also involved connections to classical topics in coding theory, such as the role of the field size of the operation and the decoding complexity.
4. Consistency Issues in Distributed Storage: In a brief discussion conducted jointly with the Private Information Retrieval for Storage Codes working group, consistency related issues in storage of multiple versions of data were discussed. A discussion of relevant applications and a related coding theoretic formulation ensued.

References

- 1 M. Ye and A. Barg, “Explicit constructions of high-rate MDS array codes with optimal repair bandwidth.” arXiv:1604.00454 (2016).
- 2 B. Sasidharan, M. Vajha, and P.V. Kumar, “An Explicit Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and All-Node Repair.” <http://arxiv.org/pdf/1607.07335.pdf>
- 3 N. Raviv, N. Silberstein and T. Etzion, “Constructions of high-rate minimum storage regenerating codes over small fields,” 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, 2016, pp. 61–65.
- 4 B. and P. V. Kumar, “High-rate regenerating codes through layering,” 2013 IEEE International Symposium on Information Theory Proc, pp. 1611–1615.
- 5 Z. Wang, A.G. Dimakis, and J. Bruck, “Rebuilding for array codes in distributed storage systems,” 2010 IEEE Globecom Workshops, 2010.
- 6 T. Westerback, R. Freij-Hollanti, and C. Hollanti, “Applications of polymatroid theory to distributed storage systems,” 53rd Annual Allerton Conference on Communication, Control, and Computing, 2015.
- 7 T. Westerback, R. Freij-Hollanti, T. Ernvall and C. Hollanti, “On the Combinatorics of Locally Repairable Codes via Matroid Theory,” in IEEE Transactions on Information Theory, vol. 62, no. 10, pp. 5296–5315, Oct. 2016.

- 8 M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory* 60.5 (2014): 2856–2867.
- 9 K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, “Finite length analysis of caching-aided coded multicasting,” 52nd Annual Allerton Conference on Communication, Control, and Computing, 2014.
- 10 N. Golrezaei, A. Molisch, A. G. Dimakis, and G. Caire, “Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution,” *IEEE Communications Magazine* 51.4 (2013): 142–149.
- 11 A. Golovnev, O. Regev, and O. Weinstein, “The Minrank of Random Graphs,” arXiv preprint arXiv:1607.04842 (2016).
- 12 Z. Wang and V. R. Cadambe, “Multi-Version Coding-An Information Theoretic Perspective of Consistent Distributed Storage,” arXiv preprint arXiv:1506.00684 (2015).
- 13 K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, K. Ramchandran, “Speeding up distributed machine learning using codes,” arXiv preprint arXiv:1512.02673 (2015).
- 14 S. Dutta, V. R. Cadambe, and P. Grover, “Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products,” to appear in Proceedings of The Thirtieth Annual Conference on Neural Information Processing Systems (NIPS), 2016.
- 15 I. Tamo and A. Barg, “A family of optimal locally recoverable codes,” *IEEE Transactions on Information Theory* 60.8 (2014): 4661–4676.
- 16 V. Guruswami and M. Wootters, “Repairing Reed-Solomon Codes,” Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, 2016.

4.3 Private Information Retrieval for Storage Codes

Camilla Hollanti (Aalto University, FI)

License © Creative Commons BY 3.0 Unported license
© Camilla Hollanti

The PIR working group discussed various topics related to (dynamic) coded storage and (adversarial) PIR, with some extensions to asynchronized systems and private keyword search. Recent results were reviewed related to matroids (Hollanti), batch codes (Skachek), and PIR, and this part of the discussion is likely to lead to a joint publication (Hollanti and Skachek, jointly with Thomas Westerber and Ragnar Freij-Hollanti).

Discussions have been continued via email (e.g., between Etzion and Hollanti), as well as during some visits that were prompted during this seminar (e.g., Eimear Byrne is visiting Hollanti and Greferath at Aalto University Sep.-Dec. 2016). During this seminar and working group discussions, open Ph.D. positions related to coded storage and PIR in the ANTA group at Aalto University were advertised, and a new student was found (from University of Neuchatel, M.Sc. advised by Elisa Gorla) and who started at Aalto University in October 2016.

4.4 DNA-Based Storage

Olga Milenkovic (University of Illinois – Urbana Champaign, US)

License  Creative Commons BY 3.0 Unported license
© Olga Milenkovic

The DNA working group focused on the problem of using DNA as information storage media. In this setting, information is encoded into strings over the alphabet of nucleobase symbols $\{A, T, G, C\}$. Because DNA may be easily replicated and a massive amount of information stored reliably with minimal space requirements, it has enormous potential as a method of storage in the time of big data. A current obstacle for practical implementations is that it remains costly to write information to and read from strands of DNA, though this cost is going down with time.

Recently, technologies for sequencing (reading) strands of DNA have improved dramatically, and will hopefully continue to become more reliable and accessible. In particular, nanopore technologies allow researchers to sequence longer strands of DNA than was previously possible. DNA is read as it passes through a nano-scale hole called a nanopore. Any given strand of DNA is read multiple times, until there is sufficient data for decoding. The errors introduced by this sequencing method include insertion and deletion errors: while being read, a strand may oscillate, doubling the sequence of read symbols back on itself. That is, if $\bar{s} \in \mathbb{F}_4^\ell$ is a sequence of ℓ symbols, and \bar{s}^R is the same sequence reversed, then we may have insertion error patterns of the form $\bar{s}t\bar{s}^R$, where $t \in \mathbb{F}_4$ is a transition symbol before the reversal of the strand. Thus, when designing codes for use in this application, we seek to avoid sequences which contain both a sequence of length ℓ , for some ℓ , and its reversal. Other constraints to consider include having an appropriate balance of bases, creating unique addresses which identify each strand of DNA, and avoiding particular subsequences which are biologically difficult to synthesize (write). After discussing the background of the problem, we briefly discussed several relevant open problems, which are given below.

- How many sequences exist in \mathbb{F}_4^n with the property that there are no “oscillating” subsequences of length ℓ ? That is, how many sequences avoid the pattern $\bar{s}t\bar{s}^R$ for $\bar{s} \in \mathbb{F}_4^\ell$?
- Can we mathematically justify methods for error reduction that seem to work in practice? For example: if we balance the nucleobase content in subblocks of a given length, we can avoid deletions when sequencing. However, this is an observed phenomenon, rather than a mathematically-developed strategy.

4.5 Rank-Metric Codes

John Sheekey (University College Dublin, IE)

License  Creative Commons BY 3.0 Unported license
© John Sheekey

Rank-metric codes are codes using matrices over a field as codewords, with the distance function determined by the rank of the difference of two matrices. The topic is experiencing a resurgence as of late, due in part to its applications in random network coding, and potential applications in code-based cryptography.

The Rank-Metric Codes working group began with a group discussion about the background literature and potential directions for research. The group then split into two main

subgroups; one focussing on computational classifications of MRD codes, and the other on algebraic methods for constructing and decoding families of codes.

The computational subgroup reviewed the known results, and discussed the feasibility of performing exhaustive searches for new parameters. Using a combination of pre-existing algorithms designed by some members, and incorporating some ideas from semifield theory, new nontrivial classifications were obtained. It was felt that there are many more parameters that are within reach with current computational power, particularly if semifield techniques can be adapted to the case of rectangular matrices. The group also considered searching for codes with prescribed automorphism groups.

The algebraic subgroup discussed the known results for decoding of Gabidulin codes, and discussed the possibility of extending these to the recently introduced family known as twisted Gabidulin codes. As the construction using linearized polynomials is similar for both families, initial investigations looked promising, and a coarse first estimate was obtained. However it was felt that there was much room for improvement by exploiting further properties of these codes, which will require a more in-depth analysis. The group also discussed potential constructions for new MRD codes similar to Gabidulin codes, as well as the need for codes with much less structure (for the purposes of avoiding weaknesses in the associated code-based cryptographic systems).

4.6 Age & Delay of Information

Jing Zhong (Rutgers University – New Brunswick, US)

License © Creative Commons BY 3.0 Unported license
© Jing Zhong

The main focus of our working group was to introduce the age of information concept to participating coding theorists and explore potential age and delay problems in coding and storage. A brief introduction of status updating age and its connection to streaming source coding was presented and discussed during the first meeting. The disadvantage of using arithmetic coding instead of fixed-to-variable block coding due to large decoding delay was discussed, and an adaptive arithmetic coding scheme for fixed length message was proposed as a potential solution to avoid huge decoding delay. In the group meetings of the last two days, we moved to a recent topic noticed by the industry, which is the application of rateless code to multipath packet transmissions in real-time video streaming. A tutorial about spatially coupled LDPC codes was given, and the comparison between spatially coupled LDPC codes and traditional punctured LDPC codes in terms of delay and decoding probability was reviewed. At the end of seminar, we discovered several possible delay problems in file downloading from multiple servers.

Participants

- Iryna Andriyanova
University of Cergy-Pontoise, FR
- Allison Beemer
Univ. of Nebraska – Lincoln, US
- Jessalyn Bolkema
Univ. of Nebraska – Lincoln, US
- Martin Bossert
Universität Ulm, DE
- Michael Braun
Hochschule Darmstadt, DE
- Eimear Byrne
University College Dublin, IE
- Viveck Cadambe
Pennsylvania State University –
University Park, US
- Gerard Cohen
Telecom Paris Tech, FR
- Alex Dimakis
University of Texas – Austin, US
- Iwan M. Duursma
University of Illinois –
Urbana Champaign, US
- Michelle Effros
CalTech – Pasadena, US
- Rafah El-Khatib
EPFL – Lausanne, CH
- Tuvi Etzion
Technion – Haifa, IL
- Christina Fragouli
University of California at Los
Angeles, US
- Heide Gluesing-Luerssen
University of Kentucky, US
- Elisa Gorla
Université de Neuchâtel, CH
- Marcus Greferath
Aalto University, FI
- Venkatesan Guruswami
Carnegie Mellon University –
Pittsburgh, US
- Daniel Heinlein
Universität Bayreuth, DE
- Camilla Hollanti
Aalto University, FI
- Anna-Lena
Horlemann-Trautmann
EPFL – Lausanne, CH
- Christine A. Kelley
Univ. of Nebraska – Lincoln, US
- P. Vijay Kumar
Indian Institute of Science –
Bangalore, IN
- Julia Lieb
Universität Würzburg, DE
- Hans-Andrea Loeliger
ETH – Zürich, CH
- Michael Luby
Qualcomm Inc. – San Diego, US
- Felice Manganiello
Clemson University, US
- Carolyn Mayer
Univ. of Nebraska – Lincoln, US
- Sihem Mesnager
University of Paris VIII, Telecom
Paris Tech, FR
- Olgica Milenkovic
University of Illinois –
Urbana Champaign, US
- Elie Najm
EPFL – Lausanne, CH
- Alessandro Neri
Universität Zürich, CH
- Mario Osvin Pavcevic
University of Zagreb, HR
- Sven Puchinger
Universität Ulm, DE
- Tovoheri Randrianarisoa
Universität Zürich, CH
- Alberto Ravagnani
Université de Neuchâtel, CH
- Johan Rosenkilde
Technical Univ. of Denmark –
Lyngby, DK
- Joachim Rosenthal
Universität Zürich, CH
- John Sheekey
University College Dublin, IE
- M. Amin Shokrollahi
EPFL – Lausanne, CH
- Vitaly Skachek
University of Tartu, EE
- Emina Soljanin
Rutgers Univ. – Piscataway, US
- Pascal Vontobel
The Chinese University of Hong
Kong, HK
- Jing Zhong
Rutgers University –
New Brunswick, US



Integrating Process-Oriented and Event-Based Systems

Edited by

David Eysers¹, Avigdor Gal², Hans-Arno Jacobsen³, and
Matthias Weidlich⁴

¹ University of Otago, NZ, dme@cs.otago.ac.nz

² Technion – Haifa, IL, avigal@ie.technion.ac.il

³ TU München, DE, arno.jacobsen@tum.de

⁴ HU Berlin, DE, matthias.weidlich@informatik.hu-berlin.de

Abstract

This report documents the programme and the outcomes of Dagstuhl Seminar 16341 on “Integrating Process-Oriented and Event-Based Systems”, which took place August 21–26, 2016, at Schloss Dagstuhl – Leibniz Center for Informatics. The seminar brought together researchers and practitioners from the communities that have been established for research on process-oriented information systems on the one hand, and event-based systems on the other hand. By exploring the use of processes in event handling (from the distribution of event processing to the assessment of event data quality), the use of events in processes (from rich event semantics in processes to support for flexible BPM), and the role of events in process choreographies, the seminar identified the diverse connections between the scientific fields. This report summarises the outcomes of the seminar by reviewing the state-of-the-art and outlining research challenges on the intersection of process-oriented and event-based systems.

Seminar August 21–26, 2016 – <http://www.dagstuhl.de/16341>

1998 ACM Subject Classification C.2.4 Distributed Systems, H. Information Systems

Keywords and phrases distributed systems, event-based systems, process-aware information systems

Digital Object Identifier 10.4230/DagRep.6.8.21

Edited in cooperation with Jan Sürmeli


1 Executive Summary

David Eysers

Avigdor Gal

Hans-Arno Jacobsen

Matthias Weidlich

License  Creative Commons BY 3.0 Unported license

© David Eysers, Avigdor Gal, Hans-Arno Jacobsen, and Matthias Weidlich

Background and Motivation

Process-oriented information systems are software systems that execute and manage a process, broadly defined as a coordinated execution of actions to achieve a certain goal. As such, they support Business Process Management (BPM) initiatives. Process-oriented systems have been traditionally used in domains such as business process automation, enterprise application integration, and collaborative work. Recently, there has also been a significant uptake of process-oriented information systems in transportation, logistics, and medical



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

Integrating Process-Oriented and Event-Based Systems, *Dagstuhl Reports*, Vol. 6, Issue 8, pp. 21–64

Editors: David Eysers, Avigdor Gal, Hans-Arno Jacobsen, and Matthias Weidlich



Dagstuhl Reports

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

infrastructures – domains that impose new challenges in terms of system reactivity and adaptability. Here, trends such as sensing of data (e.g., based on RFID technology) and advancing system integration (driven by technical standards such as EPCglobal) represent opportunities to strengthen the event-perspective in process-oriented systems in order to achieve more flexible and comprehensive process control.

Event-based systems, in turn, have been put forward to integrate heterogeneous systems in a flexible and scalable manner by separating communication from application logic. These systems provide interaction models, mechanisms for routing events between components, and techniques for the detection of composite events, i.e., for Complex Event Processing (CEP). Although event-based systems are typically positioned as general-purpose technology, they have found their way into many applications where event generation is comparatively deterministic and follows structured behaviour. In domains such as transportation, logistics, and the medical sector, events handled by event-based systems stem from the execution of processes, which are partially supported by process-oriented information systems. Exploiting the process-perspective, therefore, promises to lead to advancements in the design, analysis, and optimisation of event-based systems.

The increasing overlap of application scenarios that involve concepts and techniques of process-oriented as well as event-based systems, however, is only marginally supported by exchange and convergence of the related research fields. Strong communities have been established for research on either type of system. Yet, due to the missing link between these communities, manifold opportunities for ground-breaking research and broad impact in industry are missed out. Research efforts related to the underlying theory as well as specific platforms are duplicated and similar approaches are developed in both communities.

Breaking this disconnect had been the goal that the seminar aimed to achieve by identifying the links between conceptual models, formal analysis methods, and engineering techniques developed for either type of system.

Seminar Structure

Given that seminar attendees came from two rather disconnected communities, the first day of the seminar featured four tutorials to establish a joint understanding of essential concepts and terminology. First, Alessandro Margara presented an overview of the basic techniques to manage streams of events. Mathias Weske then gave a primer on BPM, elaborating on the main concepts, models, and the role of events for process management. An advanced view on techniques for event processing was given by Alejandro Buchmann. Stefanie Rinderle-Ma closed this part of the seminar with a tutorial on management, utilisation, and analysis of instance data in distributed process management.

The remainder of the seminar week was centred on break-out sessions, in which participants worked on particular groups on the intersection of process-oriented and event-based systems. In these working groups, participants discussed the relevant state-of-the-art and identified the research challenges under a near-, mid-, or long-term perspective. In addition, there were two sessions in which seminar participants gave a very short overview of their recent research work.

Topics and Key Challenges

The working groups focussed on a diverse set of topics, highlighting the key challenges that need to be addressed:

Event Models for BPM: Semantics of Events and Patterns. Starting from the observation that event models are well-established in both BPM and CEP and that their coupling has obvious benefits, the challenge relates to the question of how events can guide the evolution or adaptation of process instances.

Towards Automatic Event-Based Monitoring of Processes. Event-based monitoring of processes is influenced by the availability of patterns, the consequences of monitoring results, and the integration of contextual information. These dimensions render it particularly challenging to comprehensively discover and utilise patterns for process monitoring.

Patterns and Models for Communication. The communication models underlying an event-based middleware have diverse implications for the interplay of processes and event patterns – and a major challenge is the identification of requirements that are imposed by process scenarios on communication models.

Choreographies and Inter-Process Correlation. Common languages for the description of interacting processes lack capabilities for the specification of event-based processing. The challenge is to develop a better grounding of choreography languages and enable analysis of the information flow between processes.

Abstraction Levels: Processes versus Events. Observing that methods in BPM mainly proceed top-down, whereas event processing is often approached bottom-up, a key challenge is the identification of the right abstraction level on which concepts and methods shall be integrated.

Context in Events and Processes. The context of a process may influence event processing, and the context as materialised in complex events impacts the execution of a process. Yet, a suitable representation and dynamic evolution of context information is an open research challenge.

Integrated Platforms for BPM & CEP. The integration of traditional BPM or CEP engines promises accelerated application development and lower maintenance cost. To attain this end, the challenge of developing a unified model for events and processes, enabling well-grounded architectural decisions, needs to be addressed.

(Highly) Distributed Processes & The Role of Events. Events and processes can both be handled in a centralised or distributed infrastructure and open challenges relate to the tradeoffs regarding trustworthiness, reliability, and scalability.

Event Data Quality. Event data may be uncertain, which needs to be reflected in processes that are influenced by these events. The challenge is how to capture such uncertainty and make explicit how it influences decision making on the level of the process.

From Event Streams to Process Models and Back. Event patterns and processes are typically concerned with events on different levels of abstractions, which can be bridged only on the basis of a unifying formal model. Further challenges arise from the imprecision of event definitions in processes and the expressiveness of CEP languages when capturing procedural behaviour.

Compliance, Audit, Privacy and Security. Compliance checking of business processes may benefit from CEP systems and BPM tools may be useful to express service level agreements in event-based systems. Challenges, however, are methods for a structured integration of BPM and CEP technology and their alignment with informal compliance requirements.

Main Recommendations

From the discussions and the exchange of ideas during the workshop, a set of recommendations was able to be distilled in order to materialise the benefits of integrating process-oriented and event-based systems.

Build a community around BPM and CEP. The topics on the intersection of process-oriented and event-based systems provide a rich field for high-impact research. The number and diversity of open research questions call for a long-term research initiative, so that a respective community needs to be built up. To achieve this, it is recommended that joint workshops be initiated at the flagship conferences in either field, the BPM conference and the DEBS conference, and to evaluate potential co-location of the conferences in future.

Start research on integrated models. For many of the aforementioned challenges, the lack of integrated models, in which processes and events are first-class citizens, turns out to be a major issue. Research shall be devoted to creating such models, clarifying which basic notions of events exist, and considering the semantics of distributed event generation.

Facilitate joint research. Joint research is currently hindered not only by the disconnect of the research communities, but also by a lack of a common set of standard concepts in either community. There is a need for concise overviews of the most important concepts and methods in either field, e.g., by means of standard textbooks. Researchers from one field need to be able to quickly gather the level of understanding of the other field that is required for joint research initiatives.

Engage industry. The integration of process-oriented and event-based systems is driven by particular domains, such as logistics, health, and mobility. The prioritisation of challenges and the evaluation of developed solutions critically depends on the involvement of industrial partners from these domains. As such, it is recommended to reach out to industry to develop evaluation scenarios and benchmark datasets. One viable means for this are the research proposals on the EU and national levels that involve BPM and CEP experts from both academia and industry.

2 Table of Contents

Executive Summary

David Eysers, Avigdor Gal, Hans-Arno Jacobsen, and Matthias Weidlich 21

Working groups

Event Models for BPM: Semantics of Events and Patterns and Formal Methods for Reasoning on Events in the BPM Context
Anne Baumgraß, Alexander Artikis, Annika M. Hinze, Ken Moody, Wolfgang Reisig, Stefanie Rinderle-Ma, Stijn Vansummeren, and Matthias Weidlich 28

Compliance, Audit, Privacy and Security
David Eysers, Jean Bacon, Martin Jergler, Ken Moody, Stefanie Rinderle-Ma, and Stijn Vansummeren 30

Towards Automatic Event-Based Monitoring of Processes
Annika M. Hinze, Alexander Artikis, Anne Baumgraß, Alejandro P. Buchmann, Claudio Di Ciccio, Hans-Arno Jacobsen, Boris Koldehofe, Alessandro Margara, Pnina Soffer, and Holger Ziekow 31

Patterns and Models for Communication
Boris Koldehofe, Oliver Kopp, Wolfgang Reisig, Martin Ugarte, and Roman Vitenberg 34

Choreographies and Inter-Process Correlation
Oliver Kopp, Wolfgang Reisig, Jatinder Singh, Sergey Smirnov, Jan Sürmeli, Roman Vitenberg, Matthias Weidlich, and Kaiwen Zhang 35

Abstraction Levels: Processes versus Events
Sankalita Mandal, David Eysers, Agnes Koschmider, Ken Moody, Cesare Pautasso, Mohammad Sadoghi Hamedani, Wei Song, Lucinéia Heloisa Thom, and Lijie Wen 38

Context in Events and Processes
Alessandro Margara, Alejandro P. Buchmann, Sankalita Mandal, Cesare Pautasso, Arik Senderovich, Sergey Smirnov, Matthias Weidlich, and Mathias Weske 39

Integrated Platforms for BPM & CEP
Mohammad Sadoghi Hamedani, Alejandro P. Buchmann, Hans-Arno Jacobsen, Martin Jergler, Sankalita Mandal, Cesare Pautasso, Stefan Schulte, Jatinder Singh, Sergey Smirnov, and Mathias Weske 41

(Highly) Distributed Processes & The Role of Events
Stefan Schulte, Jean Bacon, Avigdor Gal, Martin Jergler, Stefanie Rinderle-Ma, Arik Senderovich, Vinay Setty, Martin Ugarte, and Stijn Vansummeren 42

Optimisation opportunities
Roman Vitenberg, Avigdor Gal, Alessandro Margara, Vinay Setty, Martin Ugarte, Matthias Weidlich, Lijie Wen, and Kaiwen Zhang 45

Event Data Quality
Kaiwen Zhang, Alexander Artikis, Anne Baumgraß, Avigdor Gal, Mohammad Sadoghi Hamedani, and Stefan Schulte 49

From Event Streams to Process Models and Back
Holger Ziekow, Jean Bacon, Claudio Di Ciccio, David Eysers, Boris Koldehofe, Oliver Kopp, Agnes Koschmider, Pnina Soffer, Wei Song, and Jan Sürmeli 50

Tutorials

Managing Streams of Events: An Overview <i>Alessandro Margara</i>	52
Business Process Management: Concepts, Models, Events <i>Mathias Weske</i>	52
Management, Utilisation, and Analysis of Instance Data in Distributed Process Settings <i>Stefanie Rinderle-Ma</i>	53
Event Processing <i>Alejandro P. Buchmann</i>	53

Overview of Talks

Online Learning for Complex Event Recognition <i>Alexander Artikis</i>	53
Smart Logistics in Practice – Using Event Processing for Comprehensive Transportation Monitoring <i>Anne Baumgraß</i>	54
Predictive Task Monitoring: Processing Flight Events to Foresee Diversions <i>Claudio Di Ciccio</i>	54
Smart Landscape: The Rugged Internet of Things <i>Annika M. Hinze</i>	55
BPM in Cloud Architectures: Enabling the Internet of Things Through Effective Business Processes Management with Events <i>Hans-Arno Jacobsen</i>	56
D2Worm – A Management Infrastructure for Distributed Data-centric Workflows <i>Martin Jergler</i>	56
Explicit Subscription for Enabling Event Buffering <i>Sankalita Mandal and Jan Sürmeli</i>	57
Associative Composition of Stream Processing Processes <i>Wolfgang Reisig</i>	57
RichNote: Adaptive Selection and Delivery of Rich Media Notifications to Mobile Users <i>Roman Vitenberg</i>	58
Real-time Explorative Event-based Systems <i>Mohammad Sadoghi Hamedani</i>	58
The ROAD from Sensor Data to Process Instances via Interaction Mining <i>Arik Senderovich</i>	59
Cost-Effective Resource Allocation for Deploying Pub/Sub on Cloud <i>Vinay Setty, Guido Urdaneta, Gunnar Kreitz, and Maarten van Steen</i>	59
Challenges of Data Integration in Cross-Organisational Processes <i>Sergey Smirnov</i>	60
Repairing Event Logs <i>Wei Song and Hans-Arno Jacobsen</i>	60

Efficient Handling of Out-of-Order Events
Jan Sürmeli 61

Research Issues on the Extraction of Process Models from Natural Language Text
Lucinéia Heloisa Thom and Renato César Borges Ferreira 62

Scalable, Expressive Publish/Subscribe Systems
Kaiwen Zhang, Hans-Arno Jacobsen, Mohammad Sadoghi Hamedani, and Roman Vitenberg 62

Participants 64

3 Working groups

3.1 Event Models for BPM: Semantics of Events and Patterns and Formal Methods for Reasoning on Events in the BPM Context

Anne Baumgraß (Synfioo – Potsdam, DE), Alexander Artikis (NCSR Demokritos – Athens, GR), Annika M. Hinze (University of Waikato, NZ), Ken Moody (University of Cambridge, GB), Wolfgang Reisig (HU Berlin, DE), Stefanie Rinderle-Ma (Universität Wien, AT), Stijn Vansummeren (Free University of Brussels, BE), and Matthias Weidlich (HU Berlin, DE)

License © Creative Commons BY 3.0 Unported license

© Anne Baumgraß, Alexander Artikis, Annika M. Hinze, Ken Moody, Wolfgang Reisig, Stefanie Rinderle-Ma, Stijn Vansummeren, and Matthias Weidlich

Event models are well known in the complex event processing (CEP) area, while process models are well known in the business process management (BPM) area. The combination of both is beneficial and promising but it is facing several challenges.

The benefits can be shown along three examples. First, in the field of logistics, phenomena such as missed connections, congestions, technical problems or strikes may impact on the timely arrival of trucks at a destination. Transportation itself is thus dealing with much uncertainty. Uncertainties are broadly investigated in the area of event processing, however not in process models or process instance definitions. A second example is that of discovering event patterns in order to predict hazardous situations (e.g., in forestry or mining) [1]. If no event logs exist that can be mined for event pattern discovery, patterns may have to be derived with incomplete information about adverse events (e.g., fatal accidents). Exploring the involved work processes may provide helpful information about events to monitor. Third, for care processes, uncertainties such as exceptional patient conditions might trigger therapy adaptations. Here, event models might be helpful in order to support nurses in deciding on which adaptation to apply.

From the CEP perspective the examples include two modelling topics: the modelling of time, and the modelling of uncertainty. Issues concerning the modelling of time include the meaning of time-points itself (do they represent application time, detection time, processing time?) [2] and the nature of time associated to events (e.g., is it a single time-point or an interval?) [3]. Uncertainty in events (e.g., due to noisy sensor readings or unreliable network transmission) can be managed and modelled by associating probabilities to events, and calculating derive probabilities for complex events [4]. In BPM, events are used to represent milestones, trigger instantiation, define deadlines, specify message exchange or communication and many more [5]. For this a simplistic definition of events is sufficient. It defines a discrete, atomic occurrence of a happening. However, we have to exclude events from the process model that can not be defined. The process model itself determines the exact execution, not allowing for executions that have not been defined. Furthermore, adaptive processes have been investigated in BPM [6, 7, 8]. User support for process adaptations has been addressed by few works [9, 10], but the treatment and utilisation of event models has not been considered yet.

We can benefit from the combination of both worlds in the following ways:

- Events trigger changes on the instance or process level
- Event classes in a process model are a means to abstract from the exact definition of an event and the production is moved to event pattern in CEP
- Probabilities may be introduced in process models to enrich them

- Temporal constraints allow for more flexible ways of handling event occurrences, e.g., through the use of intervals
- Consider rich event structures in process models
- Use process models as structured context for CEP and event pattern definitions

Derived from this we define the specific challenges:

- **Guided process instance adaptation:** How can events be used to guide process instance adaptation? Can we use event classes? How to identify them? How to identify the levels of abstraction?
- **Events change states of entities, while this state determines processing:** How can we understand the change of properties as an event which triggers/influences the process instances and determines the flow?
- **Event context through processes:** Can the process give more context that can be used to define pattern better? Can process knowledge give sense to situations that you discover with event processing? How does it help in the discovery of events?

References

- 1 J. Bowen, A. Hinze, and SJ Cunningham. Into the woods. In *Workshop on Human Work Interaction Design (HWID): Design for Challenging Work Environments, Interact conference, September 2015*.
- 2 Christian S. Jensen, James Clifford, Ramez Elmasri, Shashi K. Gadia, Patrick J. Hayes, and Sushil Jajodia. A consensus glossary of temporal database concepts. *SIGMOD Record*, 23(1):52–64, 1994.
- 3 Michael H. Böhlen, Renato Busatto, and Christian S. Jensen. Point-versus interval-based temporal data models. In *ICDE*, pages 192–200. IEEE Computer Society, 1998.
- 4 Gianpaolo Cugola, Alessandro Margara, Matteo Matteucci, and Giordano Tamburrelli. Introducing uncertainty in complex event processing: model, implementation, and validation. *Computing*, 97(2):103–144, 2015.
- 5 Mathias Weske. *Business Process Management – Concepts, Languages, Architectures, 2nd Edition*. Springer, 2012.
- 6 Stefanie Rinderle, Manfred Reichert, and Peter Dadam. Correctness criteria for dynamic changes in workflow systems – a survey. *Data Knowl. Eng.*, 50(1):9–34, 2004.
- 7 Barbara Weber, Manfred Reichert, and Stefanie Rinderle-Ma. Change patterns and change support features – enhancing flexibility in process-aware information systems. *Data Knowl. Eng.*, 66(3):438–466, 2008.
- 8 W. Song and H. A. Jacobsen. Static and dynamic process change. *IEEE Transactions on Services Computing*, PP(99):1–1, 2016.
- 9 Barbara Weber, Manfred Reichert, Stefanie Rinderle-Ma, and Werner Wild. Providing integrated life cycle support in process-aware information systems. *Int. J. Cooperative Inf. Syst.*, 18(1):115–165, 2009.
- 10 Georg Kaes and Stefanie Rinderle-Ma. Mining and querying process change information based on change trees. In *Service-Oriented Computing – 13th International Conference, ICSOC 2015, Goa, India, November 16-19, 2015*, pages 269–284, 2015.

3.2 Compliance, Audit, Privacy and Security

David Eysers (University of Otago, NZ), Jean Bacon (University of Cambridge, GB), Martin Jergler (TU München, DE), Ken Moody (University of Cambridge, GB), Stefanie Rinderle-Ma (Universität Wien, AT), and Stijn Vansummeren (Free University of Brussels, BE)

License © Creative Commons BY 3.0 Unported license

© David Eysers, Jean Bacon, Martin Jergler, Ken Moody, Stefanie Rinderle-Ma, and Stijn Vansummeren

Applications must demonstrate compliance with policy, law and regulation. Audit is a means of achieving this. Regulations often concern privacy and security. In this space, there is a number of challenges and opportunities for the integration of process-oriented and event-based systems.

Reviewing the state-of-the-art, we observe that business process compliance has been concerned with checking/enforcing policies over business process models (design time, e.g., [4]) or process executions, mostly in the form of examination of event logs and traces (e.g., runtime monitoring, see survey in [3]). Security of business processes is often concerned with access control and anomaly detection, see survey in [2]. In event-based systems, there has been research on access control (AC) e.g., via parametrised RBAC [1] at the application level. AC is also well-established in the field of Business Process Management (BPM).

Near-term research challenges and opportunities in this space relate to the processing of audit logs, where engines for Complex Event Processing (CEP) may turn out to be a useful tool. BPM tools, in turn, may provide a useful means of expressing policies in event-based systems. Furthermore, Service Level Agreements (SLAs) are imposed on CEP engines, such as the requirement to provide composite event detection within a particular deadline, e.g., stock quote matching. There is the potential to impose similar SLAs on process-oriented systems.

Mid-term challenges relate to the observation that event-based systems tend to be developed *ad hoc*, and likewise are their application-level access control policies. BPM provides rich access control context that might be used to enrich access control within event-based systems. For example, principals' connections to process and task instances may carry across usefully to parameterised access control systems and enforcement of constraints such as Dynamic Separation of Duties (DSoD).

In the long run, a major challenge is that law is hard to translate into policy, both through being expressed in natural language, and interpreted within case law. There will be an increasing need for automation of enforcing and demonstrating compliance, given the emerging domains of Cloud Computing and the Internet of Things. Policies will need to be expressed and interpreted by machines at run-time, and be able to be validated against the law as it stands.

In terms of application logic the combination of different rules is a problem—policies may conflict at run-time. Conflict detection and resolution is difficult. Conflict may arise for many reasons, such as incorrect policy specification, an inability to meet obligations, or due to unexpected context.

Applications and audit logs are increasingly cloud hosted. While cloud tenants hold the legal responsibility, they do not have technical control over what cloud service providers are doing, and there is no transparency. For example, access to audit logs must be controlled. Also, they may be copied across international boundaries and jurisdictions.

References

- 1 Jean Bacon, David M. Eysers, Jatinder Singh, and Peter R. Pietzuch. Access control in publish/subscribe systems. In *DEBS'08: Proceedings of the second international conference on Distributed event-based systems*, pages 23–34, New York, NY, USA, 2008. ACM. doi:10.1145/1385989.1385993.
- 2 Maria Leitner and Stefanie Rinderle-Ma. A systematic review on security in process-aware information systems – constitution, challenges, and future directions. *Information and Software Technology*, 56(3):273–293, 2014. doi:10.1016/j.infsof.2013.12.004.
- 3 Linh Thao Ly, Fabrizio Maria Maggi, Marco Montali, Stefanie Rinderle-Ma, and Wil M.P. van der Aalst. Compliance monitoring in business processes: Functionalities, application, and tool-support. *Information Systems*, 54:209–234, 2015. doi:10.1016/j.is.2015.02.007.
- 4 Shazia Sadiq, Guido Governatori, and Kioumars Namiri. Modeling control objectives for business process compliance. In *Proceedings of the 5th International Conference on Business Process Management*, BPM'07, pages 149–164, Berlin, Heidelberg, 2007. Springer-Verlag. URL: <http://dl.acm.org/citation.cfm?id=1793114.1793130>

3.3 Towards Automatic Event-Based Monitoring of Processes

Annika M. Hinze (University of Waikato, NZ), Alexander Artikis (NCSR Demokritos – Athens, GR), Anne Baumgraß (Synfioo – Potsdam, DE), Alejandro P. Buchmann (TU Darmstadt, DE), Claudio Di Ciccio (Wirtschaftsuniversität Wien, AT), Hans-Arno Jacobsen (TU München, DE), Boris Koldehofe (TU Darmstadt, DE), Alessandro Margara (University of Lugano, CH), Pnina Soffer (Haifa University, IL), and Holger Ziekow (FH Furtwangen, DE)

License © Creative Commons BY 3.0 Unported license

© Annika M. Hinze, Alexander Artikis, Anne Baumgraß, Alejandro P. Buchmann, Claudio Di Ciccio, Hans-Arno Jacobsen, Boris Koldehofe, Alessandro Margara, Pnina Soffer, and Holger Ziekow

The dynamic interplay between event processing and process modelling was discussed along three dimensions, using a logistics example for illustration.

D1. Pattern discovery vs. Monitoring of known patterns

Pattern discovery is typically done once sufficient data is available (e.g., on event logs [18]) and rarely online while monitoring is done online. Online “event” discovery may be relevant in the case of many parallel event sources or in emergency situations that cannot rely on available event logs. Pattern discovery is typically done either via a process norm, i.e., detecting if a process deviates from expectation or via a process goal, i.e., prediction of success or failure of a process. For instance, the typical patterns of events about a delivery truck driving from A to B can be identified in an unsupervised manner, while supervised pattern learning requires labelling the event sequences to indicate whether the activity execution had positive or negative outcomes [10] (e.g., shipment activities that respectively delivered on time or delayed [9]). Once desired patterns are known, event streams are then evaluated online to identify the occurrence of these patterns. However, on-the-fly adaptation of the monitoring pattern may be necessary.

D2. Monitoring for adaptation vs. adaptive monitoring

Monitoring for adaptation aims to identify events that then trigger changes in the business process. While the announcement is on the process level, the adaptation is carried out on the instance level. For example, if a parcel delivery misses a shipping deadline, it may have to be redirected to go via train. For event processing, this kind of monitoring and action triggering is business as usual (e.g., change of flow depending on event sources and sinks). In contrast, monitoring adaptively is an event processing issue that can be guided by business process information. For example, the closer to the ETA of a delivery by truck, the more frequently the traffic situation needs to be monitored. Relevant information from processes could be: patterns to be monitored, constraints (deadlines, tolerance to false positives or negatives), utility functions, acceptable levels of monitoring, monitoring intervals, monitoring points, QoS as foundation to manage automatic adaptation. Changes in the monitoring may then be made, e.g., to the frequency or granularity of monitoring [5], the observed source of events and streams [2], the monitoring points within the process [1], and to variables and rules [6].

D3. Use of context for integration vs. context awareness for decision support

Context knowledge can drive the integration of heterogeneous sources. For example, instead of monitoring each parcel inside a container, the container ID is used to identify events of interest, and later the truck transporting the container. Without context knowledge, such integrations would not be possible. Context and situation awareness may enhance decision support, e.g., in the selecting between two alternative process paths [8]. When it is known that a thunderstorm may impair airplane landings, the shipment could be transported via truck instead.

Open Challenges

We thus envision the following open challenges, classified by their assessed time scope:

1. Using automatically discovered patterns in online monitoring of BPMSs (short-term)
2. Leveraging process background knowledge for pattern discovery (short-term)
3. Leveraging context to drive the integration of heterogeneous sources (short-term)
4. Monitoring events to guide process adaptation (mid-term)
5. Process information to guide the monitoring adaptation (mid-term)
6. Discovery of patterns at runtime (long-term)
7. Leveraging context and situation awareness to enhance decision support (long-term)

References

- 1 Anne Baumgrass, Cristina Cabanillas, and Claudio Di Ciccio. A conceptual architecture for an event-based information aggregation engine in smart logistics. In *EMISA*, pages 109–123. GI, September 2015.
- 2 Cristina Cabanillas, Claudio Di Ciccio, Jan Mendling, and Anne Baumgrass. Predictive task monitoring for business processes. In *BPM*, pages 424–432. Springer, September 2014. doi:10.1007/978-3-319-10172-9_31.
- 3 Tony Chau, Vinod Muthusamy, Hans-Arno Jacobsen, Elena Litani, Allen Chan, and Phil Coulthard. Automating SLA modelling. In Marsha Chechik, Mark R. Vigder, and Darlene A. Stewart, editors, *Proceedings of the 2008 conference of the Centre for Advanced Studies on Collaborative Research, October 27-30, 2008, Richmond Hill, Ontario, Canada*, page 10. IBM, 2008. doi:10.1145/1463788.1463802.

- 4 Alex King Yeung Cheung and Hans-Arno Jacobsen. Load balancing content-based publish/subscribe systems. *ACM Trans. Comput. Syst.*, 28(4):9, 2010. doi:10.1145/1880018.1880020.
- 5 Claudio Di Ciccio, Han van der Aa, Cristina Cabanillas, Jan Mendling, and Johannes Prescher. Detecting flight trajectory anomalies and predicting diversions in freight transportation. *Decision Support Systems*, 88:1–17, August 2016. doi:10.1016/j.dss.2016.05.004.
- 6 Sebastian Frischbier. *Runtime Support for Quality of Information Requirements in Event-based Systems*. PhD thesis, Technische Universität Darmstadt, 2016.
- 7 Guoli Li, Vinod Muthusamy, and Hans-Arno Jacobsen. A distributed service-oriented architecture for business process execution. *TWEB*, 4(1), 2010. doi:10.1145/1658373.1658375.
- 8 David C Luckham. *Event processing for business: organizing the real-time enterprise*. John Wiley & Sons, 2011.
- 9 Alessandro Margara, Gianpaolo Cugola, and Giordano Tamburrelli. Learning from the past: automated rule generation for complex event processing. In *DEBS*, pages 47–58. ACM, 2014.
- 10 Thomas M. Mitchell. *Machine Learning*. McGraw Hill series in computer science. McGraw-Hill, Inc., New York, NY, USA, first edition, 1997.
- 11 Vinod Muthusamy and Hans-Arno Jacobsen. BPM in cloud architectures: Business process management with SLAs and events. In Richard Hull, Jan Mendling, and Stefan Tai, editors, *Business Process Management – 8th International Conference, BPM 2010, Hoboken, NJ, USA, September 13-16, 2010. Proceedings*, volume 6336 of *Lecture Notes in Computer Science*, pages 5–10. Springer, 2010. doi:10.1007/978-3-642-15618-2_2.
- 12 Vinod Muthusamy, Hans-Arno Jacobsen, Tony Chau, Allen Chan, and Phil Coulthard. SLA-driven business process management in SOA. In Patrick Martin, Anatol W. Kark, and Darlene A. Stewart, editors, *Proceedings of the 2009 conference of the Centre for Advanced Studies on Collaborative Research, November 2-5, 2009, Toronto, Ontario, Canada*, pages 86–100. ACM, 2009. doi:10.1145/1723028.1723040.
- 13 Vinod Muthusamy, Haifeng Liu, and Hans-Arno Jacobsen. Predictive publish/subscribe matching. In Jean Bacon, Peter R. Pietzuch, Joe Sventek, and Ugur Çetintemel, editors, *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems, DEBS 2010, Cambridge, United Kingdom, July 12-15, 2010*, pages 14–25. ACM, 2010. doi:10.1145/1827418.1827423.
- 14 Vinod Muthusamy, Young Yoon, Mohammad Sadoghi, and Hans-Arno Jacobsen. eqosystem: supporting fluid distributed service-oriented workflows. In David M. Eyers, Opher Etzion, Avigdor Gal, Stanley B. Zdonik, and Paul Vincent, editors, *Proceedings of the Fifth ACM International Conference on Distributed Event-Based Systems, DEBS 2011, New York, NY, USA, July 11-15, 2011*, pages 381–382. ACM, 2011. doi:10.1145/2002259.2002320.
- 15 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, and Roman Vitenberg. Distributed event aggregation for content-based publish/subscribe systems. In Umesh Bellur and Ravi Kothari, editors, *The 8th ACM International Conference on Distributed Event-Based Systems, DEBS’14, Mumbai, India, May 26-29, 2014*, pages 95–106. ACM, 2014. doi:10.1145/2611286.2611302.
- 16 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, and Roman Vitenberg. Minimizing the communication cost of aggregation in publish/subscribe systems. In *35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015, Columbus, OH, USA, June 29 – July 2, 2015*, pages 462–473. IEEE Computer Society, 2015. doi:10.1109/ICDCS.2015.54.

- 17 Mohammad Sadoghi, Martin Jergler, Hans-Arno Jacobsen, Richard Hull, and Roman Vaculín. Safe distribution and parallel execution of data-centric workflows over the publish/subscribe abstraction. *IEEE Trans. Knowl. Data Eng.*, 27(10):2824–2838, 2015. doi:10.1109/TKDE.2015.2421331.
- 18 Wil M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011. doi:10.1007/978-3-642-19345-3.

3.4 Patterns and Models for Communication

Boris Koldehofe (TU Darmstadt, DE), Oliver Kopp (Universität Stuttgart, DE), Wolfgang Reisig (HU Berlin, DE), Martin Ugarte (Free University of Brussels, BE), and Roman Vitenberg (University of Oslo, NO)

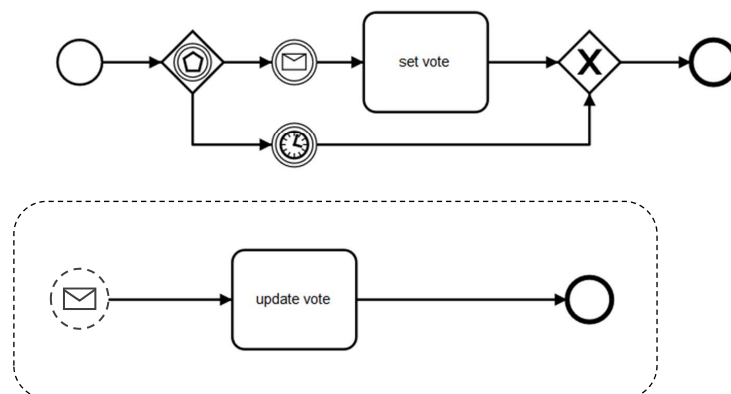
License © Creative Commons BY 3.0 Unported license

© Boris Koldehofe, Oliver Kopp, Wolfgang Reisig, Martin Ugarte, and Roman Vitenberg

Communication in processes, in particular by means of events, has many implicit effects on the modelled process, which are typically disregarded. Often many assumptions are implicit, e.g. it is assumed that the services are always on or all events are eventually received. Blocking communication may prevent processes from making progress if the expected event fails to occur. Counter measures like timeouts introduce assumptions on the timeliness of a communication system which may not be supported by a communication middleware. In general the behaviour of processes is affected by the behaviour of the middleware. The concrete properties of a middleware are often not exposed and are thus unclear for the process modeller. This poses an important source for badly specified processes and erroneous deviations of the process execution from the intended process behaviour.

For example, consider the process to collect votes for deciding on the public money to spend on a big project such as the Stuttgart 21 underground train station. Each admitted voter can send a vote and in addition update the vote within a given voting period. The process modelled in Figure 1 is underspecified in several ways: best effort delivery of events, reordered messages as well as timeouts may impose all changes to the number of votes collected.

Therefore it will be important to reflect these properties during process modelling. This will enable process modelling implementation/design to opt for a particular middleware, based



■ **Figure 1** Patterns and Models for Communication.

on the provided properties. The requirement specification for the process model may need to be mapped to the lower level specification of the communication middleware. Although there are works addressing failure models and formal specifications of middleware properties, e.g. [3, 1], there seems to be a significant modelling gap in connecting the middleware and the process layer. In closing this gap, several mid- and long-term challenges are arising.

A first challenge detected is to collect the right terms to assert the requirements or conditions that are imposed over communication. This is identified as a short-term challenge, given that it should be a starting point for middleware requirement specification. Consequently, a mid-term challenge is to provide a way for exposing middleware properties to the process modeller. Tasks like reliability and latency are today left out of the process modelling, given that modellers do not take into account the characteristics of the middleware. Another mid-term challenge is to understand that these specifications may affect (positively or negatively) the verification of processes. For example, requiring that events arrive in order provides certain guarantees on consistency; whereas an out of order arrival of a process could impact the process (e.g. loss of information). Once the previous problems are understood, a long term challenge would be to automatically determine from process descriptions suitable middleware components that are compliant, provide a good (or optimal) set of QoS guarantees, and to minimise the cost for deployment and execution. Such components may be available at a marketplace that offers appropriate components, building tools, and methods for dynamic adaptations, e.g. transitions between middleware components [SHK+15].

References

- 1 Algirdas Avizienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable Secur. Comput.*, 1(1):11–33, January 2004. doi:10.1109/TDSC.2004.2.
- 2 Oliver Kopp, Frank Leymann, and Daniel Wutke. Fault handling in the web service stack. In *Service-Oriented Computing*, pages 303–317. Springer Science + Business Media, 2010. doi:10.1007/978-3-642-17358-5_21.
- 3 J. C. Laprie, editor. *Dependability: Basic Concepts and Terminology : In English, French, German, Italian and Japanese (Dependable Computing and Fault-Tolerant Systems)*. Springer, 1991.
- 4 Ralf Steinmetz, Melanie Holloway, Boris Koldehofe, Björn Richerzhagen, and Nils Richerzhagen. Towards future internet communications – role of scalable adaptive mechanisms. In *27th Annual Conference on Symbiosis – Synergy of Humans & Technology*, pages 59–61. Eurographics Association, 2016.

3.5 Choreographies and Inter-Process Correlation

Oliver Kopp (Universität Stuttgart, DE), Wolfgang Reisig (HU Berlin, DE), Jatinder Singh (University of Cambridge, GB), Sergey Smirnov (SAP SE – Walldorf, DE), Jan Sürmeli (HU Berlin, DE), Roman Vitenberg (University of Oslo, NO), Matthias Weidlich (HU Berlin, DE), and Kaiwen Zhang (TU München, DE)

License © Creative Commons BY 3.0 Unported license

© Oliver Kopp, Wolfgang Reisig, Jatinder Singh, Sergey Smirnov, Jan Sürmeli, Roman Vitenberg, Matthias Weidlich, and Kaiwen Zhang

Choreographies enable modelling inter-organisational processes [9]. The idea is to abstract from the processes of each organisation and provide a global view. There are multiple choreography languages available, which offer two different modelling styles: interaction

models and interconnection models [8]. In interaction models (‘choreography diagrams’ in BPMN 2.0), each message exchange is made atomic, whereas in interconnection models (‘collaboration diagrams’ in BPMN 2.0), the abstract process of each participant is shown. There are various studies on the expressiveness of the languages (e.g., [5, 2]). For implementing choreographies, each participant can be implemented using processes, web services, or other approaches [3]. We see advantages in also considering event-based systems [1] as they offer an established technology, especially with respect to performance, scalability, community size, and commercial support.

There exist approaches mapping (a subset of) orchestration languages [7] and declarative process modelling languages [6] to event-based systems. Choreography modelling languages lack advanced event-based concepts, such as subscription and correlation between matched events and process instances. The CASU framework [4] presents a discussion regarding processes and instantiation. This has to be extended to the whole life-cycle of choreographies. Moreover, the formal understanding about the relation between choreography models and event-based implementations is missing. Finally, a holistic, systematic top-down approach from (conceptual) choreography models to event-based (implementation) models is required. We thus propose the following research goal for the near and far future.

3.5.1 Short-Term Goals

Extension of existing choreography modelling methods

The goal is to facilitate a systematic approach to the transformation of conceptual choreography models to models for their event-based implementation. This requires the identification and addition of missing notions in choreography modelling languages, such as subscription, correlation, and non-functional properties. Likewise, it is necessary to extend the methodological concepts accordingly, such as roles, steps, best practices and deliverables.

Introduction of intermediate implementation models

We aim at bridging the gap between conceptual models and their implementation. To this end, we propose to introduce an intermediate model, establishing a clear relation between a conceptual model and its implementation.

Identification of evaluation goals

The developed concepts have to be evaluated and validated. We aim to bring up concrete evaluation scenarios, concepts, and performance indicators to ensure a scientifically grounded evaluation.

3.5.2 Mid-Term Goals

Analysis

Analysis aspects regarding information flow have to be developed. For example, it should be ensured by the analysis that partners that do not directly interact do not gain additional information. The goal is to identify relevant analysis questions and to answer them. This includes concrete classical correctness and completeness properties.

Crisp simplistic theoretical foundation

The semantics of both the extended choreography language and event-based systems have to be formally defined. This enables formal analysis techniques to answer the analysis questions raised above.

3.5.3 Long-Term Goals

Increased level of automation

The mapping between the choreography model, the intermediate model and the event-based implementation should be as automatic as possible. At least, a skeleton of the intermediate model should be derived from the choreography model.

Implementation of logic as event-based systems

An approach to move choreography logic into event-based systems should be developed, preserving BPM benefits such as monitoring and analysis. Both the internal logic and the logic of the called services itself should be moved into event-based systems.

References

- 1 Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv.*, 44(3):15:1–15:62, June 2012. doi:10.1145/2187671.2187677.
- 2 Gero Decker, Alistair P. Barros, Frank Michael Kraft, and Niels Lohmann. Non-desynchronizable Service Choreographies. In Athman Bouguettaya, Ingolf Krüger, and Tiziana Margaria, editors, *ISCOC 2008: 6th International Conference on Service-Oriented Computing*, volume 5364 of *LNCS*, pages 331–346, 2008.
- 3 Gero Decker, Oliver Kopp, Frank Leymann, and Mathias Weske. Interacting services: From specification to execution. *Data & Knowledge Engineering*, 68(10):946–972, April 2009.
- 4 Gero Decker and Jan Mendling. Process Instantiation. *Data & Knowledge Engineering*, 68:777–792, 2009.
- 5 Gero Decker and Mathias Weske. Local enforceability in interaction Petri nets. In *Business Process Management*. Springer, 2007.
- 6 Martin Jergler, Mohammad Sadoghi, and Hans-Arno Jacobsen. D2WORM: A management infrastructure for distributed data-centric workflows. In *SIGMOD Conference*, pages 1427–1432. ACM, 2015.
- 7 Guoli Li, Vinod Muthusamy, and Hans-Arno Jacobsen. A distributed service-oriented architecture for business process execution. *ACM Trans. Web*, 4(1):1–33, Jan. 2010. doi:10.1145/1658373.1658375.
- 8 Andreas Schönberger. Do we need a refined choreography notion? In *3rd Central-European Workshop on Services and their Composition (ZEUS 2011)*. CEUR-WS.org, 2011.
- 9 Wil M. P. van der Aalst and Mathias Weske. The P2P approach to interorganizational workflows. In *Seminal Contributions to Information Systems Engineering*, pages 289–305. Springer, 2013. doi:10.1007/978-3-642-36926-1_23.

3.6 Abstraction Levels: Processes versus Events

Sankalita Mandal (Hasso-Plattner-Institut – Potsdam, DE), David Eyers (University of Otago, NZ), Agnes Koschmider (KIT – Karlsruher Institut für Technologie, DE), Ken Moody (University of Cambridge, GB), Cesare Pautasso (University of Lugano, CH), Mohammad Sadoghi Hamedani (Purdue University, US), Wei Song (Nanjing University of Science & Technology, CN), Lucinéia Heloisa Thom (Federal University of Rio Grande do Sul, BR), and Lijie Wen (Tsinghua University Beijing, CN)

License © Creative Commons BY 3.0 Unported license

© Sankalita Mandal, David Eyers, Agnes Koschmider, Ken Moody, Cesare Pautasso, Mohammad Sadoghi Hamedani, Wei Song, Lucinéia Heloisa Thom, and Lijie Wen

URL https://docs.google.com/document/d/10SSdxSAmjT2WbLHJmO_RgqrP5_1Wewbi52N-JMR3oqw

3.6.1 How to characterise and compare processes versus events?

Modelling

In general, process management follows the top-down approach where we start from a business goal and then model and execute the activities required to achieve it. So the outcome of a process would be the actions that are taken and the goal that is fulfilled collectively by those actions. Event processing typically follows a bottom-up approach where we start with the raw events and aggregate them based on some specific patterns or rules. Thus, the outcome of event processing are higher level or complex events. Though this aggregation may take time due to event buffering, simple events occur instantaneously. Processes (and their activities) instead have a duration. Processes model complete state machines, while events focus on transitions. From a graph representation perspective, processes correspond to nodes while events are on the edges. Unlike the arbitrary (though directed) topology structure for processes, the graph topology for event aggregation can be very well depicted by a tree structure. So far, process models are targeted towards the business analysts and the CEP developers take care of specifying the event hierarchies.

Runtime execution

Both process and event processing performance is measured in terms of latency and throughput. Scalability may depend on: number of process instances, size of process models, number of users versus number of rules and number of event types and stream sources. Both facilitate building flexible systems as processes can be dynamically adapted and evolved. Likewise, the set of CEP rules can be changed on the fly. In terms of monitoring, processes are concerned with the progress of their execution and give an explicit and persistent representation of their state and execution history, which could also ease debugging and testing. While (some) process engines offer transactional guarantees to deal with outages during the execution of processes, it remains to be seen whether process or event engines can be safely introduced within a safety critical system.

3.6.2 How are processes and events connected?

Processes are triggered by events. Processes catch events, which will produce a transition of their execution state and trigger the execution of further activities and the emission of events following the control flow structure of the process model. Events are consumed or produced by processes. Process start, process completion, and many other intermediate events can be explicitly represented in a process model. Low-level events produced during the execution of

a process and its activities can be logged and used for monitoring, auditing, etc. In general, the execution semantics of a process model corresponds to a partial order of events.

3.6.3 What are some challenges to integrate processes and events on a conceptual level?

How to find the suitable abstraction level for different modelling goals (time, space, cost)? How to deal with conflicting sources in large-scale systems integration? How to handle unexpected events? The “serious” Smart City scenario may provide a suitable challenging context to evaluate integrated event/process platforms.

3.7 Context in Events and Processes

Alessandro Margara (University of Lugano, CH), Alejandro P. Buchmann (TU Darmstadt, DE), Sankalita Mandal (Hasso-Plattner-Institut – Potsdam, DE), Cesare Pautasso (University of Lugano, CH), Arik Senderovich (Technion – Haifa, IL), Sergey Smirnov (SAP SE – Walldorf, DE), Matthias Weidlich (HU Berlin, DE), and Mathias Weske (Hasso-Plattner-Institut – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license

© Alessandro Margara, Alejandro P. Buchmann, Sankalita Mandal, Cesare Pautasso, Arik Senderovich, Sergey Smirnov, Matthias Weidlich, and Mathias Weske

Context-aware computing was discussed by Schilit and Theimer in 1994 to be software that “adapts according to its location of use, the collection of nearby people and objects, as well as changes to those objects over time”. In other words, context is the model of the environment that provides shared knowledge useful for the processes or event patterns at hand. For instance, the currency of a financial transaction depends on the country in which a process or an event pattern is executed: abstracting the concept of currency and modelling it as part of the context can help to simplify the pattern or process definition, avoid duplicate definitions and terminology ambiguity and biases. Thus, modelling context separately with respect to the core event patterns and processes strengthens separation of concerns, eases maintainability, and facilitates the reuse of the shared context between different processes. Several models have been proposed in the literature that could be used as a starting point to represent the context of processes and event patterns.

When considering the integration of business processes and event based systems, events can provide and contribute to detect the context for the process analysis, execution and external interaction. Similarly, processes provide the context for event pattern evaluation, optimisation and filtering. More in detail:

- The process provides the context in which event patterns operate.
- The process determines which events are relevant (filtering).
- The process determines if events will no longer arrive (optimisation).
- The process determines validity windows for pattern detection (session windows).
- The events and the situations that can be inferred through event pattern detection define the context in which the processes operate.
- Context events influence how processes make decisions (control flow).
- Context events determine which process variant is selected for deployment (variability).
- Context events constrain the binding of resources with activities (dynamic binding).
- Context compatibility is required for event subscriptions (interop, external data flow).

- Context can provide a global state shared between multiple instances (internal data flow).
- Context may impact the granularity of the monitoring events that are collected from one or more process instances.

Challenges

Representation of context: Which formalisms and models can be adopted to represent the context for processes and for event patterns? Are existing models adequate or we need new formalisms? Is a single formalism sufficient to capture both the context of processes and the context of event patterns?

Scope of context: How to distinguish what is included in the process and what is context? Context alignment at design time – how to determine the points for integrating context into processes? Should process and context remain orthogonal/independently modelled? If the process and event do not share the context, how to convert from one to another?

Context alignment at runtime: How to relate the running process with correct (smart) context? Which events are relevant for the current instance? How to discover the hidden context which explains variations in process outcomes?

References

- 1 Cristiana Bolchini, Carlo Curino, Elisa Quintarelli, Fabio A. Schreiber, and Letizia Tanca. A data-oriented survey of context models. *SIGMOD Record*, 36(4):19–26, 2007. doi:10.1145/1361348.1361353.
- 2 Ludger Fiege, Mariano Cilia, Gero Mühl, and Alejandro P. Buchmann. Publish-subscribe grows up: Support for management, visibility control, and heterogeneity. *IEEE Internet Computing*, 10(1):48–55, 2006. doi:10.1109/MIC.2006.17.
- 3 Tobias Freudenreich. *Simplifying the use of event-based systems with context mediation and declarative descriptions*. PhD thesis, Technische Universität, Darmstadt, 2015. URL: <http://tuprints.ulb.tu-darmstadt.de/5131/>.
- 4 Cesare Pautasso and Gustavo Alonso. Flexible binding for reusable composition of web services. In Thomas Gschwind, Uwe Aßmann, and Oscar Nierstrasz, editors, *Software Composition, 4th International Workshop, SC 2005, Edinburgh, UK, April 9, 2005, Revised Selected Papers*, volume 3628 of *Lecture Notes in Computer Science*, pages 151–166. Springer, 2005. doi:10.1007/11550679_12.
- 5 Michael Rosemann and Jan Recker. Context-aware process design exploring the extrinsic drivers for process flexibility. In Gil Regev, Pnina Soffer, and Rainer Schmidt, editors, *Proceedings of the CAISE*06 Workshop on Business Process Modelling, Development, and Support BPMDS'06, Luxembourg, June 5-9, 2006*, volume 236 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006. URL: <http://ceur-ws.org/Vol-236/paper9.pdf>.
- 6 Oumaima Saidani and Selmin Nurcan. Towards context aware business process modelling. In *8th Workshop on Business Process Modeling, Development, and Support (BPMDS'07), CAiSE*, volume 7, page 1, 2007.

3.8 Integrated Platforms for BPM & CEP

Mohammad Sadoghi Hamedani (Purdue University, US), Alejandro P. Buchmann (TU Darmstadt, DE), Hans-Arno Jacobsen (TU München, DE), Martin Jergler (TU München, DE), Sankalita Mandal (Hasso-Plattner-Institut – Potsdam, DE), Cesare Pautasso (University of Lugano, CH), Stefan Schulte (TU Wien, AT), Jatinder Singh (University of Cambridge, GB), Sergey Smirnov (SAP SE – Walldorf, DE), and Mathias Weske (Hasso-Plattner-Institut – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license

© Mohammad Sadoghi Hamedani, Alejandro P. Buchmann, Hans-Arno Jacobsen, Martin Jergler, Sankalita Mandal, Cesare Pautasso, Stefan Schulte, Jatinder Singh, Sergey Smirnov, and Mathias Weske

In general, BPM platforms are intended for expressive and flexible modelling of (business) processes while CEP platforms are designed for formulation and efficient detection of patterns or composite events. One may further view processes as partial ordering of events (i.e., a pattern or sequence of events) thereby giving rise to exploiting CEP as an efficient backend for executing BPM. On the one hand, unlike specialised BPM platforms, CEP has a loosely coupled architecture using a decentralised messaging substrate that is governed by publish/subscribe paradigm. As a result, the global state of CEP is distributed in the decentralised fabric of publish/subscribe (i.e., indirect state), which may have the potential to provide better compliant support for private data. On the other hand, BPM platforms rely on central messaging architecture with a tighter coupling among processes that rely on a more complex interaction and interfaces in comparison to basic publish and subscribe primitives. As a result, the global state of BPM is directly captured, which further simplifies the governance of the execution and the quality of service such as message ordering guarantees and transactional support. These benefits are gained at the cost of central execution of processes using a specialised (and possibly an *ad hoc*) runtime.

Therefore, we envision significant opportunities in integrating BPM and CEP platforms to accelerate the development and reduce the maintenance cost of building the BPM (or even CEP) engine. At one extreme, the CEP can be used exclusively as an enabler of BPM by mapping processes into a chain of rule firings that is formulated as a set of subscriptions/publications (e.g., [1-4]). Thus, the process is now represented as a lightweight, distributed agent that simply issues a set of subscriptions and publications on behalf of one or more processes. A weaker form of integration is to exploit the CEP engine only as notification or messaging substrate while the process execution runtime remains within the BPM platform. Alternatively, the CEP engine could be used only for the dynamic monitoring health and progress of processes using the generated events in BPM or the offline mining of the event log to support functionalities such as compliance and auditing. At the other extreme, one may explore the role of BPM as an enabler of CEP in order to benefit from the visual and flexible modelling power of BPM.

The key research problem in this space is to develop a unified event-process model (such as representing a process as a partial order of events) in order to enable building a unified engine that can model and execute both events and processes in a scalable, decentralised, distributed, and secured architecture (possibly by exploiting the inherent decoupling of the publish/subscribe paradigm).

References

- 1 Martin Jergler, Hans-Arno Jacobsen, Mohammad Sadoghi, Richard Hull, and Roman Vaculín. Safe distribution and parallel execution of data-centric workflows over the publish/sub-

- scribe abstraction. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 1498–1499. IEEE Computer Society, 2016. doi:10.1109/ICDE.2016.7498393.
- 2 Martin Jergler, Mohammad Sadoghi, and Hans-Arno Jacobsen. D2WORM: A management infrastructure for distributed data-centric workflows. In Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives, editors, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 – June 4, 2015*, pages 1427–1432. ACM, 2015. doi:10.1145/2723372.2735362.
 - 3 Haifeng Liu and Hans-Arno Jacobsen. A-ToPSS: A publish/subscribe system supporting imperfect information processing. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossman, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 – September 3 2004*, pages 1281–1284. Morgan Kaufmann, 2004. URL: <http://www.vldb.org/conf/2004/DEMP8.PDF>.
 - 4 H.-A. Jacobsen, M. Jergler, M. Sadoghi. Geo-distribution of flexible business processes over publish/subscribe paradigm. In *16th ACM/IFIP/USENIX International Middleware Conference (Middleware 2016)*, Trento, Italy, Dec. 2016.
 - 5 Mohammad Sadoghi, Martin Jergler, Hans-Arno Jacobsen, Richard Hull, and Roman Vaculín. Safe distribution and parallel execution of data-centric workflows over the publish/subscribe abstraction. *IEEE Trans. Knowl. Data Eng.*, 27(10):2824–2838, 2015. doi:10.1109/TKDE.2015.2421331.

3.9 (Highly) Distributed Processes & The Role of Events

Stefan Schulte (TU Wien, AT), Jean Bacon (University of Cambridge, GB), Avigdor Gal (Technion – Haifa, IL), Martin Jergler (TU München, DE), Stefanie Rinderle-Ma (Universität Wien, AT), Arik Senderovich (Technion – Haifa, IL), Vinay Setty (MPI für Informatik – Saarbrücken, DE), Martin Ugarte (Free University of Brussels, BE), and Stijn Vansummeren (Free University of Brussels, BE)

License © Creative Commons BY 3.0 Unported license

© Stefan Schulte, Jean Bacon, Avigdor Gal, Martin Jergler, Stefanie Rinderle-Ma, Arik Senderovich, Vinay Setty, Martin Ugarte, and Stijn Vansummeren

The discussion about the nature of process and event distribution yielded a two-dimensional table: Events can come from either a centralised source (e.g., a single patient monitor device) or distributed sources (e.g., surveillance cameras). Process management may be centralised (e.g., performed on a single machine or performed by a single organisational unit) or distributed, which may be done top-down, where processes are distributed for the purpose of efficiency, security, etc., or bottom-up. This classification leads to four combinations, namely centralised events-centralised processes (CECP), centralised events-distributed processes (CEDP), distributed events-centralised processes (DECP) and distributed events-distributed processes (DEDP).

An example that illustrates the four combinations involves a hospital that discharges patients to home monitoring. There is a process for each patient that continuously performs monitoring of vital signs: temperature, pulse rate, etc. If all is well, the data streams are stored locally. If a composite event detection indicates a possible emergency condition, an external event is sent to the hospital. The hospital process is responsible for responding to emergency events from all patients. This may involve invoking other emergency services such as sending an ambulance to the patient’s home.

3.9.1 CECF

The classical scenario in BPM is when processes are executed in a centralised manner. Specifically, the activities, decisions, and information flow of the business process happen in a single location. Further, the current scenario considers a situation when events are gathered and processed in a single location. Following our example, patient data is collected from wearable devices that the patient carries around. Further, the monitoring process (which we consider to be a business process) is also executed on the same device. Related work includes the analysis of small data, i.e. individual event gathering, works by Deborah Estrin, e.g. [1], and application paper from machine learning on detecting machine breakdowns [2]. From the BPM perspective, most processes that are considered in the literature except previous work on distributed processes, are centralised.

Challenges

1. Event data collection (privacy (of patients), trust (how much patients trust us, how much do we trust machine/patient sensors), new sensors to collect new kinds of data (perhaps some things are not measurable yet), sample rates (under-sampling vs. over-sampling of the device)).
2. Event processing (local processing on tiny devices, transmission of data).
3. BPM (process mining in small data devices (e.g. time series analysis) – real-time and off-line, granularity in analysis of such data: low-level recordings vs. high-level activity (abstraction gap), workflow recognition (much harder than recognising a single activity), combining historical data with recently collected data for analytics (short-term vs. long-term)).

3.9.2 DECF

Events from distributed sources are collected and processed by a centralised process, which is typically a centralised CEP engine. In the hospital example this occurs in a patient monitoring process. Every patient is equipped with a sensor measuring certain body parameters. If a combination of parameter values indicates a possible emergency (e.g. internal bleeding: low temperature, low blood pressure, high heart rate) an alarm is raised. A centralised monitoring process collects alarm events from individual patients and considers whether to send notifications to call for more doctors. Related work includes centralised event processing engines, e.g. Esper, event ordering Pub/Sub [3], and event aggregation in Pub/Sub.

Challenges

1. Out-of-order arrival of events from different event sources causes misdetection of complex events and thereby corrupts process execution: How can the order of events be guaranteed or how can misordering be detected?
2. Event loss (identification of lost events, compensation of event loss, high volume of event data that needs to be processed, high burst rates of events).

3.9.3 DPCE

This scenario refers to a centralised event source but the processes are distributed for load-balancing purposes. This scenario is critical for applications where fast processing is necessary. With a high volume stream of events centralised processing becomes a bottleneck. For fast

processing of events, distributed computing using platforms such as Apache Storm¹ and Spark streaming² is often necessary. For example, in a hospital, sensor data from patients' wearable devices is collected in a centralised location but for faster processing, such as scheduling logistics, inventory and global decision making, distributed processing is done. Other tasks like aggregation and top-k may require distribution as well [4, 5].

Challenges

Stock quote matching [6] is an example of a DPCE scenario. Such applications produce massive-scale streams of events from a centralised location. Analysing them in real-time in a centralised setting becomes a bottleneck. Faster response times are critical for many business processes in such scenarios. While distributed platforms such as Storm and Spark streaming can handle massive-scale streams, their response time needs to be improved. The event streams in such a scenario apart from being massive are also unpredictable and consequently require elastic provisioning of resources. Finally since the events are processed in a distributed fashion, correctness of results must be ensured as well.

3.9.4 DEDP

In this setting, both processes and events are distributed. In our example, the hospital may be monitoring multiple patients (generating distributed event streams) and has to coordinate (among its own departments) emergency services, where each department has its own process model and external services are independent (police, ambulance). This is a setting where observed event data is incomplete and uncertain (e.g. because of network delays or noisy sensors). Several approaches exist on how to establish choreographies taking care of criteria such as consistency between different partner processes, e.g. [7]. Less attention has been spent on aspects such as change in distributed processes, e.g. [8]. Distribution in CEP generally refers to parallelised techniques for detecting complex patterns. In this respect, some efforts have been devoted to generate, given a complex pattern, a query plan that allows for federation of sub-patterns (from RETE networks [9] and Information Flow Processing Systems [10] to Intrusion Detection Systems [11]).

Challenges

1. Methods to deal with distributed incompleteness and uncertainty in order to ensure reliability of process enactment.
2. Methods for resilient data integration are required to improve event completeness and accuracy.
3. Methods to deal with privacy and security are required.
4. Events may trigger the local processes to change; also affecting the global process, hence requiring management of ripple effects.

References

- 1 Deborah Estrin. *Small data, where $n = me$* . Communications of the ACM 57(4):32–34, 2014
- 2 Sohyung Cho, Shihab Asfour, Arzu Onar, Nandita Kaundinya. *Tool breakage detection using support vector machine learning in a milling process*. International Journal of Machine Tools and Manufacture 45(3):241–249, 2005

¹ <http://storm.apache.org/>

² <http://spark.apache.org/streaming/>

- 3 Kaiwen Zhang, Vinod Muthusamy, Hans-Arno Jacobsen. *Total order in content-based publish/subscribe systems*. IEEE 32nd International Conference on Distributed Computing Systems, 335–344, 2012
- 4 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, Roman Vitenberg. *Distributed event aggregation for content-based publish/subscribe systems*. 8th ACM International Conference on Distributed Event-based Systems, 95–106, 2014
- 5 Brian Babcock, Chris Olston. *Distributed top-k monitoring*. 2003 ACM SIGMOD International Conference on Management of Data, 28–39, 2003
- 6 Matteo Migliavacca, Ioannis Papagiannis, David M. Eyers, Brian Shand, Jean Bacon, Peter Pietzuch. *DEFCON: High-Performance Event Processing with Information Security*. 2010 USENIX Annual Technical Conference, 2010.
- 7 Wil M.P. van der Aalst, Niels Lohmann, Peter Massuthe, Christian Stahl, Karsten Wolf. *Multiparty Contracts: Agreeing and Implementing Interorganizational Processes*. The Computer Journal 53(1):90–106, 2010
- 8 Walid Fdhila, Conrad Indiono, Stefanie Rinderle-Ma, Manfred Reichert. *Dealing with change in process choreographies: Design and implementation of propagation algorithms*. Information Systems 49:1–24, 2015
- 9 Ho Soo Lee, Marshall I. Schor. *Match algorithms for generalized Rete networks*. Artificial Intelligence 54(3):249–274, 1992
- 10 Michael Stonebraker, Ugur Cetintemel, Stan Zdonik. *The 8 requirements of real-time stream processing*. ACM SIGMOD Record 34(4): 42–47, 2005
- 11 Tim Bass. *Intrusion detection systems and multisensor data fusion*. Communications of the ACM 43(4):99–105, 2000

3.10 Optimisation opportunities

Roman Vitenberg (University of Oslo, NO), Avigdor Gal (Technion – Haifa, IL), Alessandro Margara (University of Lugano, CH), Vinay Setty (MPI für Informatik – Saarbrücken, DE), Martin Ugarte (Free University of Brussels, BE), Matthias Weidlich (HU Berlin, DE), Lijie Wen (Tsinghua University Beijing, CN), and Kaiwen Zhang (TU München, DE)

License © Creative Commons BY 3.0 Unported license

© Roman Vitenberg, Avigdor Gal, Alessandro Margara, Vinay Setty, Martin Ugarte, Matthias Weidlich, Lijie Wen, and Kaiwen Zhang

The scope of optimisations in EP and BPM is large and diverse. A common use of EP is as a building block or component or a guiding paradigm within a BPM architecture. On the one hand, process improvements can be based on event data. On the other hand, application feedback and BPM insights can be exploited to optimise and configure implementation of the events processing component.

The state of the art covers a large area in a way that could be further systematised. Some representative works include [1, 18, 17, 6].

Challenge 1: Exploit BPM insight to improve event processing

A growing number of enterprises use complex event processing for monitoring and controlling their operations, while business process models are used to document working procedures. In a recent work [15], a method for optimising complex event processing using business process models was proposed, based on the extraction of behavioural constraints that are used, in turn, to rewrite patterns for event detection, and select and transform execution plans. This

work can be extended, for instance, by exploiting constraints derived from data dependencies between activities. For that, the use of declarative process languages, such as DECLARE, can be applied. Also, mining techniques can be used to identify common event patterns, and probabilistic methods can be used to quantify the importance of their appearance as a further optimisation method.

Challenge 2: Exploit process information to improve Quality of Service (QoS) in event processing/monitoring

CEP engines can monitor the execution of processes. In this context, we can envision the exploitation of process-related information to guide the CEP engine and optimise the QoS it provides. For instance, by knowing which information is more relevant for the stakeholder, the CEP engine can prioritise the execution of the rules that derive such information. Similarly, the CEP engine can allocate more resources to the processors that are responsible for extracting the knowledge that is more critical for the process at hand. Finally, in the case of bursts, the monitoring engine can also use process information to selectively discard the data that is less relevant for the process (selective load shedding) to bring back the monitoring system to its normal operating conditions with minimum impact on the quality of the results produced.

This involves the following open subchallenges:

- Extract/Infer Quality of Service (QoS) requirements from the process specifications
- Map the inferred QoS requirements to the CEP queries
- Adapt the CEP engine behaviour online to optimise QoS requirements

Challenge 3: Using Complex Event Processing (CEP) methods to better detect/predict/improve processes at run-time

An important aspect in Business Process Management is the ability to detect anomalies, predict delays and improve process execution policies in real-time. For example, in hospitals where real-time data is available, we would like to detect a re-route of a patient, her waiting time for the next event and the best order in which the patients are to be served. Current techniques for such run-time analysis rely on models (e.g. process models or machine learning models) that are fitted to historical data. However, the updated streams of incoming events provide additional information (beyond the current state of the system). For example, we may find out that a doctor did not arrive to work on the day of the surgery (current state assessment), and that all surgeries so far took longer than usual. Using a set of rules based on CEP, we could invoke an online mechanism (e.g. a heuristic) that would aid with detection prediction and improvement of running process instances [11].

Challenge 4: Optimising distributed EP and business process execution

A distributed orchestration engine for business process execution can be realised using a distributed content-based publish/subscribe system [8]. A service discovery mechanism is provided to automatically perform service composition. By supplying SLAs into the workflows, the architecture can be monitored and elastically provisioned. In this model, different execution engines can decide whether to host a needed activity or not, based on bandwidth, energy, and response time.

A related challenge is to optimise a distributed event-based system informed by business process execution. By identifying common flows, distributed operators can be placed in optimised paths. For instance, multiple closely related processes can be co-located on the same

machine. In addition, vertical scaling of overloaded CEP operators can be identified/predicted using workflow analysis.

Challenge 5: Use CEP techniques to diagnose/improve/optimize business process models in offline mode

CEP techniques are most suitable for monitoring, controlling and adapting of business process execution in online mode. This is very common in sensing or web service situations. Nonetheless, in industries, especially for manufacturing companies, the adaptation of business process models is not done directly in real time. It is the responsibility of business managers to diagnose/improve/optimize business process models manually in offline mode. However, CEP techniques can still help to provide insights into improvement chances of business process models upon the huge amount of events recorded during the execution of business process instances. E.g., if a former task A was executed by Tom and the latter task B was executed by Peter, the duration of the process instance could be very large. If so, it should be recommended that one resource assignment rule should be added to prevent Tom and Peter from executing A and B successively. The most important issue in this context is generating such recommended assignment rules in a systematic and automatic way. For related work, see [5, 7, 13].

Challenge 6: Distributed query optimisation in BPM, efficient implementation of non-functional properties in decentralised BPM

A number of core issues in the event processing area such as query optimisation, performance, and non-functional properties (reliability [14], adaptivity [9], prioritisation) are currently beyond the mainstream focus of BPM. All those issues could be useful for BPM, however. For example, CEP considers the fact that event streams originate at different sources and tries to place different operators on different nodes. Such a distributed query optimisation could improve the performance of BPM.

Challenge 7: Resource allocation for event-based architectures

A number of business processes involve running event-based architectures in the cloud. An important optimisation challenge in this context is an adaptive allocation of servers and bandwidth in the cloud that optimises QoS and reduces the use of resources and monetary cost. It was proposed in [12] how to perform such an allocation for a pub/sub architecture supporting social notifications from Spotify.

Challenge 8: Align aspects of declarative languages found in CEP and in BPM

The task of defining a unified framework is still a work in progress for both BPM and CEP (see [16] and [3], respectively). At the core of this task lies the problem of having an expressive declarative language that at the same time allows for efficient evaluation. Although the concept of evaluation is different in the two contexts (event detection in CEP and model compliance in BPM), there are certainly shared aspects that have yet to be understood. For example, the study of CEP has derived some formal models that allow to understand the efficient evaluation of operators (e.g. [2] and [4]). These operators are also present in BPM (e.g. the “followed by” operator), and therefore a knowledge transfer could lead to new optimisations in Business Processes. Conversely, BPM models constantly have to deal with the problem of making model definition accessible to end users, and therefore the community

has developed graphics representations (e.g. [10]) that could be used in CEP to allow for simple definition of patterns.

References

- 1 Irene Barba, Andreas Lanz, Barbara Weber, Manfred Reichert, and Carmelo Del Valle. *Optimized Time Management for Declarative Workflows*, pages 195–210. Springer Berlin Heidelberg, 2012.
- 2 Gianpaolo Cugola and Alessandro Margara. Tesla: A formally defined event specification language. In *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems*, pages 50–61. ACM, 2010.
- 3 Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv.*, 44(3):15:1–15:62, June 2012.
- 4 Yanlei Diao, Neil Immerman, and Daniel Gyllstrom. Sase+: an agile language for Kleene closure over event streams. 2007.
- 5 Dirk Fahland and Wil M. P. van der Aalst. Model repair – aligning process models to reality. *Inf. Syst.*, 47:220–243, 2015.
- 6 G. Hermosillo, L. Seinturier, and L. Duchien. Using complex event processing for dynamic business process adaptation. In *Services Computing (SCC), 2010 IEEE International Conference on*, pages 466–473, 2010.
- 7 T. Jin, J. Wang, Y. Yang, L. Wen, and K. Li. Refactor business process models with maximized parallelism. *IEEE Transactions on Services Computing*, 9(3):456–468, 2016.
- 8 Guoli Li, Vinod Muthusamy, and Hans-Arno Jacobsen. A distributed service-oriented architecture for business process execution. *ACM Trans. Web*, 4(1):2:1–2:33, January 2010.
- 9 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, and Roman Vitenberg. Distributed event aggregation for content-based publish/subscribe systems. In *Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, DEBS’14*, pages 95–106. ACM, 2014.
- 10 Maja Pesic, Helen Schonenberg, and Wil M. P. van der Aalst. Declare: Full support for loosely-structured processes. In *Proceedings of the 11th IEEE International Enterprise Distributed Object Computing Conference*, pages 287–, 2007.
- 11 Arik Senderovich, Andreas Rogge-Solti, Avigdor Gal, Jan Mendling, and Avishai Mandelbaum. The ROAD from sensor data to process instances via interaction mining. In *Advanced Information Systems Engineering – 28th International Conference, CAiSE 2016, Ljubljana, Slovenia, June 13-17, 2016. Proceedings*, pages 257–273, 2016.
- 12 Vinay Setty, Roman Vitenberg, Gunnar Kreitz, Guido Urdaneta, and Maarten van Steen. Cost-effective resource allocation for deploying pub/sub on cloud. In *IEEE 34th International Conference on Distributed Computing Systems, ICDCS 2014, Madrid, Spain, June 30 – July 3, 2014*, pages 555–566, 2014.
- 13 Rob J. B. Vanwersch, Khurram Shahzad, Irene Vanderfeesten, Kris Vanhaecht, Paul Grefen, Liliane Pintelon, Jan Mendling, Godefridus G. van Merode, and Hajo A. Reijers. A critical evaluation and framework of business process improvement methods. *Business & Information Systems Engineering*, 58(1):43–53, 2016.
- 14 Marco Volz, Boris Koldehofe, and Kurt Rothermel. Supporting strong reliability for distributed complex event processing systems. In *Proceedings of the 2011 IEEE International Conference on High Performance Computing and Communications, HPCC’11*, pages 477–486. IEEE Computer Society, 2011.
- 15 Matthias Weidlich, Holger Ziekow, Avigdor Gal, Jan Mendling, and Mathias Weske. Optimizing event pattern matching using business process models. *IEEE Trans. Knowl. Data Eng.*, 26(11):2759–2773, 2014.
- 16 Mathias Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer, 2007.

17 Gregor Zellner. A structured evaluation of business process improvement approaches. *Business Process Management Journal*, 17(2):203–237, 2011.

18 Haopeng Zhang, Yanlei Diao, and Neil Immerman. On complexity and optimization of expensive queries in complex event processing. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD’14, pages 217–228. ACM, 2014.

3.11 Event Data Quality

Kaiwen Zhang (TU München, DE), Alexander Artikis (NCSR Demokritos – Athens, GR), Anne Baumgraß (Synfoo – Potsdam, DE), Avigdor Gal (Technion – Haifa, IL), Mohammad Sadoghi Hamedani (Purdue University, US), and Stefan Schulte (TU Wien, AT)

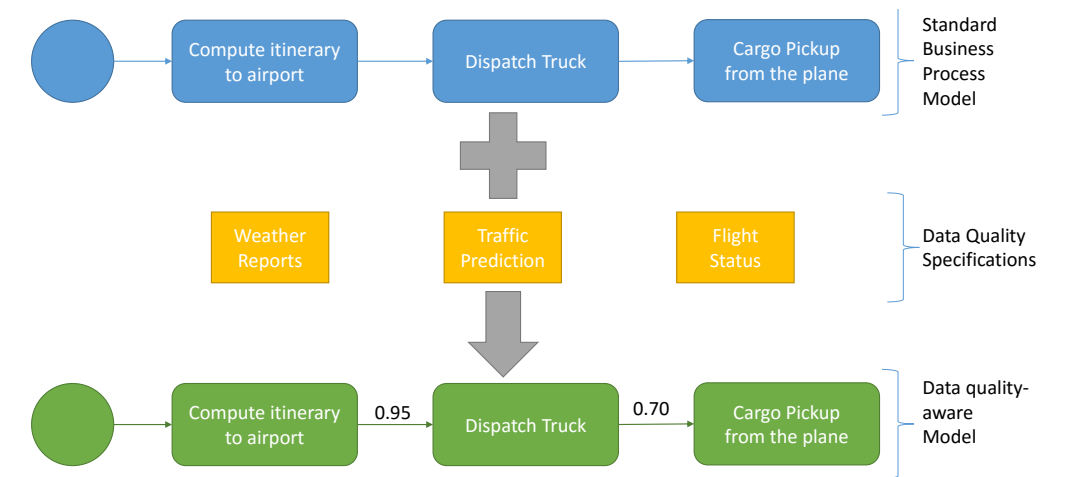
License © Creative Commons BY 3.0 Unported license
© Kaiwen Zhang, Alexander Artikis, Anne Baumgraß, Avigdor Gal, Mohammad Sadoghi Hamedani, and Stefan Schulte

Traditional business process modelling assumes perfect knowledge of all process executions, as well as perfect sources of event data. This creates a mismatch between the plan and the execution, which is affected by sources of uncertainties, such as unreliable sensors, traffic patterns, weather, etc.

We seek to capture the complexity of data quality, integrate it in business process modelling, and extract insights for concrete decision-making.

Figure 2 illustrates our top-down approach to enriching business process modelling using data quality. Two components must be supplied: the standard business process model, with perfect knowledge and quality, and the specifications for data quality. Combining these together form a data quality-aware model, which can then be analysed to gain additional insights on the business.

The first challenge is defining sources of uncertainties, assessing their quality, and capturing this information in a complete specification format. Sources include the reliability of the event streams (e.g. ordering, missing events, corrupted events, duplication, latency), consistency of correlated events (e.g. the same vehicle reported at two different locations simultaneously), integration issues when fusing various sets of data, or dealing with subjective processes



■ Figure 2 Event Data Quality.

(processes involving human factor). This type of information must be captured in a format which can be easily read and parsed. For instance, a possibility would be to provide contracts that each component abides to in XML.

The second challenge is to enrich the standard provided model with the data quality specifications into a transformed data quality-aware business process model. Currently, there are no such known models. We envision tools such as probabilistic and possibilistic graphical models to be adapted to represent processes with probabilistic transitions, depending on the level of uncertainty. Fuzzy set notation can be used to assign multiple properties to objects that handled by uncertain sources. For instance, a package inspected by customs could be 30% toy, 70% electronic, which would dictate its processing flow.

The third challenge is to study the quality-aware model and decision making. This includes worst-case/average-case analysis, model refinement, and improving the reliability of the event processing system employed (with the associated cost clearly defined).

3.12 From Event Streams to Process Models and Back

Holger Ziekow (FH Furtwangen, DE), Jean Bacon (University of Cambridge, GB), Claudio Di Ciccio (Wirtschaftsuniversität Wien, AT), David Eysers (University of Otago, NZ), Boris Koldehofe (TU Darmstadt, DE), Oliver Kopp (Universität Stuttgart, DE), Agnes Koschmider (KIT – Karlsruher Institut für Technologie, DE), Pnina Soffer (Haifa University, IL), Wei Song (Nanjing University of Science & Technology, CN), and Jan Sürmeli (HU Berlin, DE)

License © Creative Commons BY 3.0 Unported license

© Holger Ziekow, Jean Bacon, Claudio Di Ciccio, David Eysers, Boris Koldehofe, Oliver Kopp, Agnes Koschmider, Pnina Soffer, Wei Song, and Jan Sürmeli

Joint work of Jean Bacon, Claudio Di Ciccio, David Eysers, Annika Hinze, Arno Jacobsen, Boris Koldehofe, Oliver Kopp, Agnes Koschmider, Pnina Soffer, Wei Song, Jan Sürmeli, Lucineia Heloisa Thom, Lijie Wen, Holger Ziekow

Business process models typically capture processes at a high level of abstraction. In contrast, the logic for processing event streams is typically defined at a lower and more technical level (e.g. in the form of CEP rules). Specifically, process models operate on the granularity of activities within a process (e.g. the shipment of a good), whereas event processing operates on low-level input events that relate to the process (e.g. observing an RFID tag by a certain reader). We addressed the gap that exists between these abstraction levels of business process models and complex event processing rules by highlighting challenge areas where a combination of CEP and BPM could be beneficial. We identified the four challenge areas depicted in Fig. 3. Our discussions revealed the need for establishing a common ground of formalisms that unify common concepts from the worlds of CEP and BPM. This common ground was seen as a key enabler for all discussed challenge areas of combining CEP and BPM, and for bridging the gap between conceptual models and implementation.

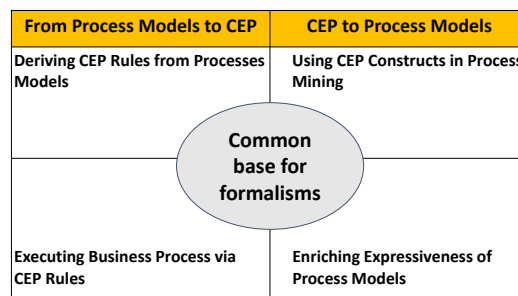
We identified the following main goals, key challenges and opportunities:

Deriving CEP Rules from Process Models (short-term):

Main goal: Automate generation of CEP rules for process monitoring and control

Key challenge: Process models do not provide all the information that may be exploited in rule generation

Opportunity: Augment process models with CEP constructs for more precise definitions of events and decision gates



■ **Figure 3** Key challenge areas.

Using CEP constructs for process mining (mid-term):

Main goal: Get from low level events (e.g. sensor data) to a process model

Key challenge: Applying process mining techniques on low-level events

Opportunity: Use CEP constructs to identify higher level activities from event logs

Enriching Expressiveness of Process Models (mid-term):

Main goal: Bridge the gap between conceptual models and implementation

Key challenge: Events that steer processes are too weakly defined in process models

Opportunity: Use CEP constructs to define events in process models more precisely

Executing Business Processes via CEP Rules (long-term):

Main goal: Seamless scalability, adaptiveness, context awareness, and distributed process execution

Key challenge: Transformation of imperative and conceptual process models into CEP rules

Opportunity: Support definitions of highly abstract activities in flexible processes with CEP constructs

Common basis for formalism (long-term):

Main goal: A meta-model for both processes and events

Key challenge: Find fundamental rules to express both control-flow and complex event patterns

Opportunity: ECA rules may inspire us

4 Tutorials

4.1 Managing Streams of Events: An Overview

Alessandro Margara (University of Lugano, CH)

License  Creative Commons BY 3.0 Unported license
© Alessandro Margara

Many modern software applications involve processing, analysing, and reacting to potentially large volumes of streaming data. Examples of such applications include monitoring systems, decision support systems, financial analysis tools, and traffic control systems. Designing and implementing these applications is difficult. Indeed, the streaming nature of the data demands for efficient algorithms, techniques, and infrastructures to analyse the incoming information on the fly and extract relevant knowledge from it. To solve this problem, several event and stream processing systems have been proposed in the last years, both from the academia and from the industry. These systems typically provide high level languages to specify how to interpret and transform the input data to produce the relevant results, and hide data distribution and communication for efficient distributed computation. This tutorial overviews the state-of-the-art proposals for stream and event processing. Given the heterogeneity of the proposed solutions, the tutorial introduces a number of models that isolate and analyse the various design choices behind such systems, and discusses the implications of such choices.

4.2 Business Process Management: Concepts, Models, Events

Mathias Weske (Hasso-Plattner-Institut – Potsdam, DE)

License  Creative Commons BY 3.0 Unported license
© Mathias Weske

Business process management covers a broad area, ranging from management topics like process improvement to the formal investigation of behavioural properties to implementation aspects related to process automation. In this talk, basic concepts in business process management are introduced. The BPM lifecycle organises major activities, including process design, automation, and enactment. Models play a key role in this field because they allow us to investigate processes, to argue about them, and to achieve a shared understanding. The syntax and semantics of process models are discussed using the BPMN industry standard. It is shown that events are a crucial aspect in the execution semantics of business processes. The talk concludes with a presentation of a process execution environment including a process engine that controls the execution of process instances.

4.3 Management, Utilisation, and Analysis of Instance Data in Distributed Process Settings

Stefanie Rinderle-Ma (Universität Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Stefanie Rinderle-Ma

Main reference W. Fdhila, C. Indiono, S. Rinderle-Ma, M. Reichert, “Dealing with change in process choreographies: Design and implementation of propagation algorithms”, *Inf. Syst.*, 49:1–24, 2015.

URL <http://dx.doi.org/10.1016/j.is.2014.10.004>

Distributed process settings are a means to describe, implement, and execute business networks in a process-oriented fashion. Describing and setting up such processes is already a challenging task. When enacting and executing them a multitude of additional challenges arise. This tutorial outlines a selection of challenges and possible solutions; they range from more basic and technical ones such as correlation to advanced utilisation and analysis of data emitted during choreography execution, for example, for distributed compliance checking and prediction of change effects in choreographies. The challenges and solutions are illustrated by means of use cases and projects from different domains such as manufacturing and energy.

4.4 Event Processing

Alejandro P. Buchmann (TU Darmstadt, DE)

License © Creative Commons BY 3.0 Unported license
© Alejandro P. Buchmann

Event processing and push-based systems are at the core of a broad range of applications. At the same time the development of application systems has lagged because of the difficulty for users to deal with a new programming paradigm and several open issues regarding non-functional properties ranging from security, privacy and trust to the management of large distributed and decentralised systems.

Starting from some examples of interesting applications we will identify open issues that must be overcome, discuss ideas for integrating event processing with other architectures and programming styles, and will look at event processing as a key element for self-adaptive systems.

5 Overview of Talks

5.1 Online Learning for Complex Event Recognition

Alexander Artikis (NCSR Demokritos – Athens, GR)

License © Creative Commons BY 3.0 Unported license
© Alexander Artikis

Joint work of A. Artikis, V. Micheloudakis, A. Skarlatidis, G. Paliouras

Complex event recognition is characterised by uncertainty and relational structure. Markov Logic Networks (MLN)s is a state-of-the-art Statistical Relational Learning framework that can naturally be applied to domains governed by these characteristics. We present OSL α – an online structure learner for MLNs that exploits an Event Calculus axiomatisation in order to constrain the space of possible structures. Learning MLNs from data streams is

challenging, as their relational structure increases the complexity of the learning process. In addition, due to the dynamic nature of event recognition applications, it is desirable to incrementally learn or revise the complex event definitions' structure and parameters. Our empirical analysis on real and synthetic data showed that OSL α learns complex event definitions orders of magnitude faster than OSL, the structure learning algorithm that it extends. Moreover, OSL α outperforms event recognition based on manual rules, and, in some cases, weighted manual rules.

References

- 1 V. Micheloudakis, A. Skarlatidis, G. Paliouras and A. Artikis. *OSLa: Online Structure Learning using Background Knowledge Axiomatization*. Proceedings of European Conference on Machine Learning (ECML), 2016

5.2 Smart Logistics in Practice – Using Event Processing for Comprehensive Transportation Monitoring

Anne Baumgraß (Synfioo – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license
© Anne Baumgraß

Joint work of Anne Baumgraß, Andreas Meyer, Marian Pufahl
URL <http://synfioo.com>

Synfioo predicts and visualises delay times in intermodal transports due to external disruptions such as strikes, tunnel closures and bad weather, using sources like social media, open and closed API, websites and telematics. Synfioo gives shippers, truckers and transport planners hours of advance warning, leading to significant savings in an industry where minor supply chain disruptions can cost millions.

The Synfioo GmbH results from the EU project “GET Service” for green logistics with 9 partners from industry and research (2012–2015). The technical background are both business process management and complex event processing. Furthermore, Synfioo implements machine learning techniques for its predictions.

5.3 Predictive Task Monitoring: Processing Flight Events to Foresee Diversions

Claudio Di Ciccio (Wirtschaftsuniversität Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Claudio Di Ciccio

Joint work of Claudio Di Ciccio, Han van der Aa, Cristina Cabanillas, Jan Mendling, Johannes Prescher, Anne Baumgraß

Main reference C. Di Ciccio, H. van der Aa, C. Cabanillas, J. Mendling, J. Prescher, “Detecting flight trajectory anomalies and predicting diversions in freight transportation,” *Decision Support Systems*, 88:1–17, 2016.

URL <http://dx.doi.org/10.1016/j.dss.2016.05.004>

Identifying flight diversions in a timely manner is a crucial aspect of efficient multi-modal transportation. When an aeroplane diverts, logistics providers must promptly adapt their transportation plans in order to ensure proper delivery despite such an unexpected event. In practice, the different parties in a logistics chain do not exchange real-time information related to flights. This calls for a means to detect diversions that just requires publicly available data,


thus being independent of the communication between different parties. The dependence on public data results in a challenge to detect anomalous behaviour without knowing the planned flight trajectory. Our work addresses this challenge by introducing a prediction model that processes events bearing only information on an aeroplane's position, velocity, and intended destination. This information is used to distinguish between regular and anomalous behaviour. When an aeroplane displays anomalous behaviour for an extended period of time, the model predicts a diversion. A quantitative evaluation shows that this approach is able to detect diverting aeroplanes with excellent precision and recall even without knowing planned trajectories as required by related research. By utilising the proposed prediction model, logistics companies gain a significant amount of response time for these cases.

References

- 1 Claudio Di Ciccio, Han van der Aa, Cristina Cabanillas, Jan Mendling, and Johannes Prescher. Detecting flight trajectory anomalies and predicting diversions in freight transportation. *Decision Support Systems*, 88:1–17, August 2016.
- 2 Cristina Cabanillas, Claudio Di Ciccio, Jan Mendling, and Anne Baumgrass. Predictive task monitoring for business processes. In Shazia Wasim Sadiq, Pnina Soffer, and Hagen Völzer, editors, *BPM*, volume 8659 of *Lecture Notes in Computer Science*, pages 424–432. Springer, September 2014.
- 3 Anne Baumgraß, Mirela Botezatu, Claudio Di Ciccio, Remco Dijkman, Paul Grefen, Marcin Hewelt, Jan Mendling, Andreas Meyer, Shaya Pourmirza, and Hagen Völzer. Towards a methodology for the engineering of event-driven process applications. In Manfred Reichert and A. Hajo Reijers, editors, *BPM Workshops*, volume 256 of *Lecture Notes in Business Information Processing*, pages 501–514. Springer International Publishing, 2016.
- 4 Anne Baumgrass, Cristina Cabanillas, and Claudio Di Ciccio. A conceptual architecture for an event-based information aggregation engine in smart logitics. In Jens Kolb, Henrik Leopold, and Jan Mendling, editors, *EMISA*, volume 248 of *Lecture Notes in Informatics (LNI)*, pages 109–123. GI, September 2015.

5.4 Smart Landscape: The Rugged Internet of Things

Annika M. Hinze (University of Waikato, NZ)

License  Creative Commons BY 3.0 Unported license
© Annika M. Hinze

Joint work of Annika Hinze, Chris Griffiths, Judy Bowen, Vimal Kumar, David Bainbridge

Different to smart city proposals, we target the harder problem of a ‘smart landscape’ in which disconnectedness due to remote locations and harsh operating conditions are common. This talk introduced the issues encountered when using IoT and event processing to help support safer working environments in hazardous work contexts. It particularly highlighted the need for identifying event patterns that are not known *a priori* as no historic event logs exist and dedicated observation of accidents is not a viable option. We have begun to explore the options of combining process approaches in combination with complex event processing techniques.

5.5 BPM in Cloud Architectures: Enabling the Internet of Things Through Effective Business Processes Management with Events

Hans-Arno Jacobsen (TU München, DE)

License © Creative Commons BY 3.0 Unported license

© Hans-Arno Jacobsen

Main reference G. Li, V. Muthusamy, H.-A. Jacobsen., “A Distributed Service Oriented Architecture for Business Process Execution”, *ACM Transactions on the Web*, 4(1)2:1–2:33, January 2010

URL <http://dx.doi.org/10.1145/1658373.1658375>

In today’s cloud-based enterprise systems, many business processes rely on service-level agreements (SLAs) to manage interactions with partners and suppliers. SLAs determine revenue, cost and customer satisfaction, but implementing and monitoring SLAs is often a manual and error-prone effort. Companies struggle with how to express, track, verify, manage, and enforce SLAs. This is further exasperated by our rapidly increasing reliance on “everything connected” to track supply and demand across global supply chains.

This talk presents a powerful enterprise process management architecture that manages SLAs across the entire supply chain and enterprise system life-cycle. Our approach leverages events available at every layer of the enterprise software systems stack to efficiently manage business process and interactions. Questions such as the following are addressed: Where is the value in real-time process monitoring and how does it work? Which technologies and design patterns are most effective for monitoring SLAs in real-time? What run-time adaptation and performance optimisations are practical to implement in business processes? What architecture can enable the above?

This talk is based on findings resulting from our PADRES Events & Services Bus (<http://padres.msrg.org>) and eQoSystem (<http://eQoSystem.msrg.org>) research projects.

5.6 D2Worm – A Management Infrastructure for Distributed Data-centric Workflows

Martin Jergler (TU München, DE)

License © Creative Commons BY 3.0 Unported license

© Martin Jergler

Joint work of Mohammad Sadoghi, Hans-Arno Jacobsen

Main reference M. Jergler, M. Sadoghi, H.-A. Jacobsen, “D2WORM: A Management Infrastructure for Distributed Data-centric Workflows”, in *Proc. of the 2015 ACM SIGMOD Int’l Conf. on Management of Data (SIGMOD’15)*, pp. 1427–1432, ACM, 2015.

URL <https://dx.doi.org/10.1145/2723372.2735362>

In this talk, we revisit D2WORM, a Distributed Data-centric Workflow Management system. D2WORM allows users to (1) graphically model data-centric workflows based on the Guard-Stage-Milestone (GSM) meta-model, (2) automatically compile the modelled workflow into several fine-grained workflow units (WFUs), and (3) deploy these WFUs on distributed infrastructures. A WFU is a system component that manages a subset of the workflow’s data model and, at the same time, represents part of the global control flow by evaluating conditions over the data. WFUs communicate with each other over a publish/subscribe messaging infrastructure that allows the architecture to scale from a single node to dozens of machines distributed over different data-centers.

5.7 Explicit Subscription for Enabling Event Buffering

Sankalita Mandal (Hasso-Plattner-Institut – Potsdam, DE) and Jan Sürmeli (HU Berlin, DE)

License © Creative Commons BY 3.0 Unported license
© Sankalita Mandal and Jan Sürmeli

Catching an event from inside a business process requires the prior subscription to some event producer such as a CEP-engine or a pub/sub broker. Between the moment of subscription and the moment of catching, multiple (possibly complex) events can occur. Current modelling languages for business processes only allow an abstract specification of event catching, but lack the notions to specify:

1. The subscription to specific, possibly complex, events from specific event producers.
2. The way multiple occurrences are handled and the order in which events are caught.

We propose to extend BPMN in order to overcome these shortcomings, centred on the explicit notion of a buffer: A buffer holds the (possibly complex) events that occur between subscription and catching, and allows the retrieval of its contents based on a buffer policy.

5.8 Associative Composition of Stream Processing Processes

Wolfgang Reisig (HU Berlin, DE)

License © Creative Commons BY 3.0 Unported license
© Wolfgang Reisig

Large stream based systems as well as (business) processes are usually composed of more elementary components. So, composition is a fundamental structuring principle of such systems and processes.

A system or a process is in general composed of many components. Examples include supply chains or systems with components that are composed along with an adapter. Composition is assumed to be associative in these cases, i.e. for components C_i it is assumed that $(C_1 + C_2) + C_3 = C_1 + (C_2 + C_3)$. This, however, requires a careful definition of components and their composition. We suggest quite a general such setting.

Based on the observation that a component frequently has two partner components in a composed system, we suggest the interface of a component to consist of two ports, with each port consisting of any set of labeled elements. Composition $C_1 + C_2$ is then defined by “gluing” elements of the right port of C_1 with equally labelled elements of the left port of C_2 .

As an application we consider a more general version of van der Aalst’ workflow nets, where in $C_1 + C_2$, some events of C_2 may already have occurred while some events of C_1 have not yet started. We show that this composition retains the important property of soundness.

5.9 RichNote: Adaptive Selection and Delivery of Rich Media Notifications to Mobile Users


Roman Vitenberg (University of Oslo, NO)

License  Creative Commons BY 3.0 Unported license
© Roman Vitenberg

In recent years, notification services for social networks, mobile apps, messaging systems and other electronic services have become truly ubiquitous. When a new content item becomes available, the service sends an instant notification to the user. When the content is produced in massive quantities, and it includes both large-size media and a lot of meta-information, it gives rise to a major challenge of selecting content to notify about and information to include in such notifications. We tackle three important challenges in realising rich notification delivery: (1) content and presentation utility modelling, (2) notification selection and (3) scheduling of delivery. We consider a number of progressive presentation levels for the content. Since utility is subjective and hard to model, we rely on real data and user surveys. We model the content utility by learning from large-scale real world data collected from the Spotify music streaming service. For the utility of the presentation levels we rely on user surveys. Blending these two techniques together, we derive utility of notifications with different presentation levels. We then model the selection and delivery of rich notifications as an optimisation problem with a goal to maximise the utility of notifications under resource budget constraints. We validate our system with large-scale simulations driven by the real-world de-identified traces obtained from Spotify. With the help of several baseline approaches we show that our solution is adaptive and resource efficient.

5.10 Real-time Explorative Event-based Systems

Mohammad Sadoghi Hamedani (Purdue University, US)

License  Creative Commons BY 3.0 Unported license
© Mohammad Sadoghi Hamedani

This talk covered formulating subscription as natural language, and subsequently, translating the imprecise natural language formulation to a precise query language (such as SQL or complex-events) followed by applying query/subscription expansion techniques to find all relevant (both historic/recent) publications in order to faithfully capture user's intention.

5.11 The ROAD from Sensor Data to Process Instances via Interaction Mining

Arik Senderovich (Technion – Haifa, IL)

License © Creative Commons BY 3.0 Unported license

© Arik Senderovich

Joint work of Arik Senderovich, Andreas Rogge-Solti, Avigdor Gal, Jan Mendling, Avishai Mandelbaum

Main reference A. Senderovich, A. Rogge-Solti, A. Gal, J. Mendling, A. Mandelbaum, “The ROAD from Sensor Data to Process Instances via Interaction Mining”, in Proc. of the 28th Int’l Conf. on Advanced Information Systems Engineering (CAiSE’16), LNCS, Vol. 9694, pp. 257–273, Springer, 2016.

URL http://dx.doi.org/10.1007/978-3-319-39696-5_16

Process mining is a rapidly developing field that aims at automated modelling of business processes based on data coming from event logs. In recent years, advances in tracking technologies, e.g., Real-Time Locating Systems (RTLS), put forward the ability to log business process events as location sensor data. To apply process mining techniques to such sensor data, one needs to overcome an abstraction gap, because location data recordings do not relate to the process directly. In this work, we solve the problem of mapping sensor data to event logs based on process knowledge. Specifically, we propose interactions as an intermediate knowledge layer between the sensor data and the event log.

We solve the mapping problem via optimal matching between interactions and process instances. An empirical evaluation of our approach shows its feasibility and provides insights into the relation between ambiguities and deviations from process knowledge, and accuracy of the resulting event log.

5.12 Cost-Effective Resource Allocation for Deploying Pub/Sub on Cloud

Vinay Setty (MPI für Informatik – Saarbrücken, DE), Guido Urdaneta, Gunnar Kreitz, and Maarten van Steen

License © Creative Commons BY 3.0 Unported license

© Vinay Setty, Guido Urdaneta, Gunnar Kreitz, and Maarten van Steen

Main reference V. Setty, R. Vitenberg, G. Kreitz, G. Urdaneta, M. van Steen, “Cost-Effective Resource Allocation for Deploying Pub/Sub on Cloud,” in Proc. of the IEEE 34th International Conference on Distributed Computing Systems (ICDCS’14), pp. 555–566, IEEE CS, 2014.

URL <http://dx.doi.org/10.1109/ICDCS.2014.63>

Publish/subscribe (pub/sub) is a popular communication paradigm in the design of large-scale distributed systems. A fundamental challenge in deploying pub/sub systems on a data center or a cloud infrastructure is efficient and cost-effective resource allocation that would allow delivery of notifications to all subscribers. In this work we addressed the answers to the following three fundamental questions: Given a pub/sub workload, (1) what is the minimum amount of resources needed to satisfy all the subscribers, (2) what is a cost-effective way to allocate resources for the given workload, and (3) what is the cost of hosting it on a public Infrastructure-as-a-Service (IaaS) provider like Amazon EC2.

We evaluate the solution experimentally using real traces from Spotify and Twitter along with a pricing model from Amazon. Using a variety of practical scenarios for each dataset, we also show that our solution scales well for millions of subscribers and runs fast.

5.13 Challenges of Data Integration in Cross-Organisational Processes

Sergey Smirnov (SAP SE – Walldorf, DE)

License © Creative Commons BY 3.0 Unported license
© Sergey Smirnov

Modern logistic companies enjoy a broad choice of IT products that can automate their business processes. The product examples include telematics solutions, software applications enabling communication within a company, and applications enabling integration with partners. However, small and medium enterprises often remain very conservative: they run their processes on disintegrated proprietary IT solutions, if any. This talk described the technical and organisational challenges that logistics companies face when they integrate their processes with other businesses.

5.14 Repairing Event Logs

Wei Song (Nanjing University of Science & Technology, CN) and Hans-Arno Jacobsen (TU München, DE)

License © Creative Commons BY 3.0 Unported license
© Wei Song and Hans-Arno Jacobsen
Main reference W. Song, X. Xia, H.-A. Jacobsen, P. Zhang, H. Hu, “Efficient alignment between event logs and process models”, *IEEE Transactions on Services Computing*, PP(99):1–1, 2016.
URL <http://dx.doi.org/10.1109/TSC.2016.2601094>

The aligning of event logs with process models is of great significance for process mining to enable conformance checking, process enhancement, performance analysis, and trace repairing. Since process models are increasingly complex and event logs may deviate from process models by exhibiting redundant, missing, and dislocated events, it is challenging to determine the optimal alignment for each event sequence in the log, as this problem is NP-hard. Existing approaches utilise the cost-based A^* algorithm to address this problem. However, scalability is often not considered, which is especially important when dealing with industrial-sized problems. In this paper, by taking advantage of the structural and behavioural features of process models, we present an efficient approach which leverages effective heuristics and trace replaying to significantly reduce the overall search space for seeking the optimal alignment. We employ real-world business processes and their traces to evaluate the proposed approach. Experimental results demonstrate that our approach works well in most cases, and that it outperforms the state-of-the-art approach by up to 5 orders of magnitude in runtime efficiency.

References

- 1 Arya Adriansyah, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. Conformance checking using cost-based fitness analysis. In *Proceedings of the 15th IEEE International Enterprise Distributed Object Computing Conference, EDOC'11, Helsinki, Finland, August 29 – September 2*, pages 55–64, 2011.
- 2 Massimiliano de Leoni, Fabrizio Maria Maggi, and Wil M. P. van der Aalst. Aligning event logs and declarative process models for conformance checking. In *Business Process Management – 10th International Conference, BPM'12, Tallinn, Estonia, September 3-6. Proceedings*, pages 82–97, 2012.
- 3 Massimiliano de Leoni and Wil M. P. van der Aalst. Aligning event logs and process models for multi-perspective conformance checking: An approach based on integer linear

- programming. In *Business Process Management – 11th International Conference, BPM’13, Beijing, China, August 26-30. Proceedings*, pages 113–129, 2013.
- 4 Jorge Munoz-Gama, Josep Carmona, and Wil M. P. van der Aalst. Conformance checking in the large: Partitioning and topology. In *Business Process Management – 11th International Conference, BPM’13, Beijing, China, August 26-30. Proceedings*, pages 130–145, 2013.
 - 5 Andreas Rogge-Solti, Ronny Mans, Wil M. P. van der Aalst, and Mathias Weske. Improving documentation by repairing event logs. In *The Practice of Enterprise Modeling – 6th IFIP WG 8.1 Working Conference, PoEM’13, Riga, Latvia, November 6-7, Proceedings*, pages 129–144, 2013.
 - 6 Wei Song, Xiaoxu Xia, Hans-Arno Jacobsen, Pengcheng Zhang, and Hao Hu. Heuristic recovery of missing events in process logs. In *IEEE International Conference on Web Services, ICWS’15, New York, USA, Jun 27-July 2*, pages 105–112, 2015.
 - 7 Wil M. P. van der Aalst, Arya Adriansyah, and Boudewijn F. van Dongen. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery*, 2(2):182–192, 2012.
 - 8 Jianmin Wang, Shaoxu Song, Xiaochen Zhu, and Xuemin Lin. Efficient recovery of missing events. *PVLDB*, 6(10):841–852, 2013.

5.15 Efficient Handling of Out-of-Order Events

Jan Sürmeli (HU Berlin, DE)

License © Creative Commons BY 3.0 Unported license
© Jan Sürmeli

Joint work of Dirk Fahland, Jan Sürmeli, Matthias Weidlich

A CEP system produces a stream of complex events from an input stream of events based on a query. The distribution of event producers raises the problem that the arrival order deviates from the production order of events: The input stream of a CEP system is thus not inherently ordered by time stamps. Purely adhering to the arrival order leads to a deviation between the computed and the desired output stream of complex events: Some complex events in the output stream are incorrect, others are missing. The problem is to achieve *correctness* and *completeness* while preserving low latency and feasibility for real time applications. One approach is that of *aggressive query evaluation*: Upon the detection of an out-of-order event, one invalidates and delivers the incorrect and missing complex events, respectively. Our research focuses on improving this approach by analysing the query at design time, and applying the analysis results to increase the performance at runtime.

5.16 Research Issues on the Extraction of Process Models from Natural Language Text

Lucinéia Heloisa Thom (Federal University of Rio Grande do Sul, BR) and Renato César Borges Ferreira

License © Creative Commons BY 3.0 Unported license

© Lucinéia Heloisa Thom and Renato César Borges Ferreira

Main reference R. C. Borger Ferreira, L. H. Thom, “Uma Abordagem para Gerar Texto Orientado a Processo a partir de Texto em Linguagem Natural”, in Proc. of the XII Brazilian Symp. on Information Systems, pp. 585–588. 2016.

URL <http://sbsi2016.ufsc.br/anais/Proceedings%20of%20the%20XII%20SBSI.pdf>

In organisations, business process modelling is very important to report, understand and automate processes. Organisations usually have unstructured documents related to their business processes, which can be very difficult to understand by process analysts and developers. The extraction of process models from natural language text may contribute to minimise the effort required during process modelling. This research proposes an approach to generate process-oriented text from text in natural language. In particular, to investigate the structure a text in natural language must present so that process models can be extracted from that. As practical result this proposes the development of an open-source tool to support: (i) the automatic selection of business process relevant information from text in natural language; (ii) the extraction of process models from business process-oriented text.

5.17 Scalable, Expressive Publish/Subscribe Systems

Kaiwen Zhang (TU München, DE), Hans-Arno Jacobsen (TU München, DE), Mohammad Sadoghi Hamedani (Purdue University, US), and Roman Vitenberg (University of Oslo, NO)

License © Creative Commons BY 3.0 Unported license

© Kaiwen Zhang, Hans-Arno Jacobsen, Mohammad Sadoghi Hamedani, and Roman Vitenberg

The publish/subscribe paradigm is known for its loosely coupled interactions and event filtering capabilities. Traditional applications using pub/sub systems require large-scale deployment and high event throughput. Thus, pub/sub has always put the emphasis on scalability and performance, to the detriment of filtering expressiveness and quality of service. The matching language is usually limited to topic-based or content-based event filtering and does not allow complex stream-based subscriptions to be expressed. Messages are delivered on a best-effort basis without any ordering or reliability guarantees.

Recently, modern pub/sub applications such as online games, social networks, and sensor networks, have specifications which extend beyond the basic semantics provided by standard systems. Installing additional services and event processing systems at the endpoints can overcome these limitations. However, we argue that such solutions are inefficient and put an avoidable strain on the pub/sub layer itself. Therefore, the focus of our work is to develop integrated solutions to extend pub/sub language expressiveness and quality of service, as well as demonstrate that this approach results in better performance from a holistic perspective. The methodology and technical insights from this line of work can contribute to the integration of event processing and business process management.

References

- 1 César Cañas, Kaiwen Zhang, Bettina Kemme, Jörg Kienzle, and Hans-Arno Jacobsen. Publish/subscribe network designs for multiplayer games. In Laurent Réveillère, Lucy

- Cherkasova, and François Taïani, editors, *Proceedings of the 15th International Middleware Conference, Bordeaux, France, December 8-12, 2014*, pages 241–252. ACM, 2014. doi:10.1145/2663165.2663337.
- 2 Christoph Doblander, Tanuj Ghinaiya, Kaiwen Zhang, and Hans-Arno Jacobsen. Shared dictionary compression in publish/subscribe systems. In *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*, DEBS'16, pages 117–124, New York, NY, USA, 2016. ACM. doi:10.1145/2933267.2933308.
 - 3 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, and Roman Vitenberg. Distributed event aggregation for content-based publish/subscribe systems. In Umesh Bellur and Ravi Kothari, editors, *The 8th ACM International Conference on Distributed Event-Based Systems, DEBS'14, Mumbai, India, May 26-29, 2014*, pages 95–106. ACM, 2014. doi:10.1145/2611286.2611302.
 - 4 Navneet Kumar Pandey, Kaiwen Zhang, Stéphane Weiss, Hans-Arno Jacobsen, and Roman Vitenberg. Minimizing the communication cost of aggregation in publish/subscribe systems. In *35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015, Columbus, OH, USA, June 29 – July 2, 2015*, pages 462–473. IEEE Computer Society, 2015. doi:10.1109/ICDCS.2015.54.
 - 5 Kaiwen Zhang, Vinod Muthusamy, and Hans-Arno Jacobsen. Total order in content-based publish/subscribe systems. In *2012 IEEE 32nd International Conference on Distributed Computing Systems, Macau, China, June 18-21, 2012*, pages 335–344. IEEE Computer Society, 2012. doi:10.1109/ICDCS.2012.17.

Participants

- Alexander Artikis
NCSR Demokritos – Athens, GR
- Jean Bacon
University of Cambridge, GB
- Anne Baumgraß
Synfoo – Potsdam, DE
- Alejandro P. Buchmann
TU Darmstadt, DE
- Claudio Di Ciccio
Wirtschaftsuniversität Wien, AT
- David Eysers
University of Otago, NZ
- Avigdor Gal
Technion – Haifa, IL
- Annika M. Hinze
University of Waikato, NZ
- Hans-Arno Jacobsen
TU München, DE
- Martin Jergler
TU München, DE
- Boris Koldehofe
TU Darmstadt, DE
- Oliver Kopp
Universität Stuttgart, DE
- Agnes Koschmider
KIT – Karlsruher Institut für
Technologie, DE
- Sankalita Mandal
Hasso-Plattner-Institut –
Potsdam, DE
- Alessandro Margara
University of Lugano, CH
- Ken Moody
University of Cambridge, GB
- Cesare Pautasso
University of Lugano, CH
- Wolfgang Reisig
HU Berlin, DE
- Stefanie Rinderle-Ma
Universität Wien, AT
- Mohammad Sadoghi
Hamedani
Purdue University, US
- Stefan Schulte
TU Wien, AT
- Arik Senderovich
Technion – Haifa, IL
- Vinay Setty
MPI für Informatik –
Saarbrücken, DE
- Jatinder Singh
University of Cambridge, GB
- Sergey Smirnov
SAP SE – Walldorf, DE
- Pnina Soffer
Haifa University, IL
- Wei Song
Nanjing University of Science &
Technology, CN
- Jan Sürmeli
HU Berlin, DE
- Lucinéia Heloisa Thom
Federal University of Rio Grande
do Sul, BR
- Martin Ugarte
Free University of Brussels, BE
- Stijn Vansummeren
Free University of Brussels, BE
- Roman Vitenberg
University of Oslo, NO
- Matthias Weidlich
HU Berlin, DE
- Lijie Wen
Tsinghua University Beijing, CN
- Mathias Weske
Hasso-Plattner-Institut –
Potsdam, DE
- Kaiwen Zhang
TU München, DE
- Holger Ziekow
FH Furtwangen, DE



Foundations of Secure Scaling

Edited by

Lejla Batina¹, Swarup Bhunia², and Patrick Schaumont³

¹ Radboud University Nijmegen, NL, lejla@cs.ru.nl

² University of Florida – Gainesville, US, swarup@ece.ufl.edu

³ Virginia Polytechnic Institute – Blacksburg, US, schaum@vt.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16342 “Foundations of Secure Scaling”. This seminar hosted researchers in secure electronic system design, spanning all abstraction levels from cryptographic engineering over chip design to system integration. We recognize that scaling is a fundamental force present at every abstraction level in electronic system design. While scaling is generally thought of as beneficial to the resulting implementations, this does not hold for secure electronic design. Indeed, the relations between scaling and the resulting security are poorly understood. This seminar facilitated the discussion between security experts at different abstraction levels in order to uncover the links between scaling and the resulting security.

Seminar August 21–26, 2016 – <http://www.dagstuhl.de/16342>

1998 ACM Subject Classification Hardware, Security/Cryptology, Verification/Logic

Keywords and phrases Cryptographic Engineering, Very Large Scale Integration, Secure Hardware Design, Technology Scaling, Complexity Scaling, Secure Evaluation

Digital Object Identifier 10.4230/DagRep.6.8.65

1 Executive Summary

Lejla Batina

Swarup Bhunia

Patrick Schaumont

License © Creative Commons BY 3.0 Unported license
© Lejla Batina, Swarup Bhunia, and Patrick Schaumont

In electronic system design, scaling is a fundamental force present at every abstraction level. Over time, chip feature sizes shrink; the length of cryptographic keys and the complexity of cryptographic algorithms grows; and the number of components integrated in a chip increases. While scaling is generally thought of as beneficial to the resulting implementations, this does not hold for secure electronic design. Larger and faster chips, for example, are not necessarily more secure. Indeed, the relations between scaling and the resulting security are poorly understood. This Dagstuhl Seminar hosted researchers in secure electronic system design, spanning all abstraction levels from cryptographic engineering over chip design to system integration.



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

Foundations of Secure Scaling, *Dagstuhl Reports*, Vol. 6, Issue 8, pp. 65–90

Editors: Lejla Batina, Swarup Bhunia, and Patrick Schaumont



Dagstuhl Reports

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

Discussion Topics

The mechanisms of secure scaling require investigation of the links between Cryptography, Technology, and Digital Integration. Cryptographers are concerned with novel and secure algorithms that remain secure even as cryptanalytic capabilities improve. Technologists are concerned with the next generation of transistors and their implementation into a reliable and stable process technology. Integrators are concerned with electronic design automation tools that can manage the rapidly increasing complexity of electronic design, and the are concerned with the integration of components on a complex system-on-chip.

Through its participants, the seminar offered a unique opportunity to discuss cross-cutting topics in Secure Scaling. The following list are examples of such cross-cutting topics.

- Scaling effects in Privacy and Security. The massive amount of connected devices will create significant challenges towards security and privacy. Major questions involve data ownership and key ownership and management.
- Power/Energy Efficient Crypto: Secure wireless devices and Secure RFID are two well known examples of applications that require security under severe power and/or energy constraints. Optimizing a cryptographic algorithm for power/energy efficiency needs to consider all abstraction levels of design.
- High-Performance Crypto: Information Technology is increasingly asymmetric, with larger, high-performance servers at one end, and a large population of tiny devices at the other side. Cryptographic designs must scale towards high-performance, high-throughput implementations while it must also accommodate small-footprint, low-latency designs.
- Secure Test: Complex chips utilize a number of testing strategies such as BIST and JTAG. When a chip includes a secure part, the test infrastructure carries a potential risk of abuse. Secure Test is a test strategy for complex chips that takes this risk fully into account.
- Complexity Management in Secure SoC: Managing and integrating a secure module into system-on-chip context is challenging and creates a hard verification problem that cuts through multiple traditional layers of design. Furthermore, managing multiple stakeholders in a single chip design is extremely challenging and may result in conflicting design requirements.
- Implementation Attacks: In modern cryptographic designs, side-channel analysis, fault-analysis and physical tampering are an integral part of the threat model. This requires design techniques that fully integrate countermeasures as part of the design process. In addition, the design of a countermeasure effective against most forms of tampering is an open research issue.
- Technology effects on implementation attacks. Better insight the internal operation of secure implementations at all abstraction levels leads to novel implementation attacks, that work at finer granularity, and that use novel source of leakage such as optical leakage.

The seminar supported participants in learning about the state-of-the-art developments in the three different domains covered in the workshop (Cryptography, Integration, and Technology). The seminar also supported the presentation of specific cross-cutting topics, as well as round-table (panel-style) discussions.

2 Table of Contents

Executive Summary	
<i>Lejla Batina, Swarup Bhunia, and Patrick Schaumont</i>	65
Organization	69
Overview of Talks	69
Privacy and Security Challenges of the Internet of Things	
<i>Bart Preneel</i>	69
Double Arbiter PUF and Security Evaluation Using Deep Learning	
<i>Kazuo Sakiyama</i>	70
IoT and Implementation Security	
<i>Thomas Eisenbarth</i>	71
A minimalistic perspective on Public Key Encryptions	
<i>Roy Debapriya Basu</i>	72
Secure System Design in the IoT Regime	
<i>Sandip Ray</i>	73
Where Security Meets Verification: From Microchip to Medicine	
<i>Swarup Bhunia</i>	74
Security Metric for IoT	
<i>Yier Jin</i>	75
Eliminating timing side-channels in cryptographic software	
<i>Peter Schwabe</i>	76
MAFIA: Micro-architecture Aware Fault Injection Attack	
<i>Bilgiday Yuce</i>	77
Optical Interaction through Chip Backside with Nanoscale Potential	
<i>Christian Boit</i>	79
How Secure are Modern FPGAs?	
<i>Shahin Tajik</i>	80
Detection and Prevention of Side-Channel Attacks	
<i>Naofumi Homma</i>	81
Implementation Security through Dynamic Reconfiguration	
<i>Nele Mentens</i>	82
Propagation of Glitches and Side-channel Attacks	
<i>Guido Bertoni</i>	83
Threshold Implementations	
<i>Svetla Nikova</i>	84
Secure Scaling, Scaling Securely	
<i>Francesco Regazzoni</i>	84
Scaling of Implementation Attacks	
<i>Georg Sigl</i>	85

Crypto, Integration, Technology: Good, Bad, Ugly?	
<i>Debdeep Mukhopadhyay</i>	87
Smart Card Secure Channel Protocol	
<i>Joan Daemen</i>	89
Participants	90

■ **Table 1** Schedule of talks.

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Topic	Cryptography	Integration	Technology	Cross-Cutting	
Chair	Lejla Batina	Swarup Bhunia	Ingrid Verbauwhede		
Speaker	Bart Preneel	Sandip Ray	Christian Boit	Guido Bertoni	Joan Daemen
	KU Leuven, BE	NXP, US	TU Berlin, DE	ST MicroElectronics, IT	ST Microelectronics, BE
Topic	IoT Privacy and Security	Secure SoC Design	Optical Interaction	Glitches and SCA	Smart Card Protocol
Speaker	Kazuo Sakiyama	Swarup Bhunia	Shahin Tajik	Svetla Nikova,	Patrick Schaumont
	UEC, JP	University of Florida, US	TU Berlin, DE	KU Leuven, BE	Virginia Tech, US
Topic	Double Arbiter PUF	Trojan Detection	Secure FPGA	Threshold Implemen.	The Stovepipe Model
Speaker	Thomas Eisenbarth	Yier Jin	Naofumi Homma	Francesco Regazzoni	
	WPI, US	U of Central Floriday, US	Tohoku University, JP	Alari, CH	
Topic	Side-channel Analysis	IoT Security Metric	EM Side-channels	Design Scaling	
Speaker	Roy D Basu	Peter Schwabe	Nele Mentens	Georg Sigl	
	IIT Kharagpur, IN	Radboud University, NL	KU Leuven, BE	TU Munich, DE	
Topic	Lightweight ECC	Timing Side-channels	Dynamic Reconfig.	Atatck Scaling	
Speaker		Bilgiday Yuce		Debdeep Mukhopadhyay	
		Virginia Tech, US		IIT Kharagpur, IN	
Topic		CPU FI Attacks		Physical Security	

3 Organization

During the first three day of the seminar, discussions highlighted each major design abstraction level, and its connection to security. In the next two days, we discussed cross-cutting issues related to secure scaling. Table 1 illustrates the schedule of talks over the three days.

After each set of talks, we organized a roundtable discussion to further elaborate on discussions raised during the presentations. To keep track of the talks, we maintained a wiki that collected all slides. The presentations are available on the Dagstuhl Wiki and the organizers will encourage the participants to publicly release their slides. We also used a note-taker for each talk, who kept track of the presentation and the questions. The note-takers supported the development of this report.

4 Overview of Talks

4.1 Privacy and Security Challenges of the Internet of Things

Bart Preneel (KU Leuven, BE)

License  Creative Commons BY 3.0 Unported license
© Bart Preneel

Notes taken by Joan Daemen.

The number of electronic devices connected to the internet has been growing exponentially and will continue to do so in the coming years. This is often called the Internet of Things (IoT). Many of these devices process and transmit personal information giving rise to privacy concerns. Additionally, the interconnectivity allows remotely monitoring and controlling things like medical sensors (some even implanted), cars (soon self-driving), aeroplanes, industrial infrastructures and power plants giving rise to security concerns. Moreover, during the last decades a number of organizations such as NSA, Google, Facebook have started using this infrastructure for mass surveillance for varying reasons including financial profit. Nowadays, we see a vast and complex ecosystem of companies trading in privacy-sensitive information of citizens for advertizing. Powerful data analysis techniques (Big Data) are applied to these data giving rise to an alarming level of knowledge present in this data.


Moreover, advances in genome decoding (and coding) technology will allow the commercial and political exploitation of our most personal data: our DNA. The large complexity of these ecosystems and the pace by which this evolves has lead to a lack of legal regulation. Moreover, the little regulation that there is, is not enforced. Citizens are stimulated to manage their personal data on remote servers owned and managed by corporate organizations. This is often referred to as the cloud.

There is no commonly agreed definition of privacy: it differs a lot per country and culture. Still, it is clear that from the privacy and security point of view, the situation is alarming and getting worse. Doom scenario's where evil forces abuse the information in the cloud for power abuse, terrorism or even world domination are not far fetched. Moreover, these systems have not been designed with security, safety or privacy in mind. Due to the complexity and evolving nature bugs and malware are introduced at a faster pace than their detection. Often products and apps are rolled out without any security protection and the idea is to add security later. In most cases, there is no incentive for implementing good security as the business model is often that the cost of fraud/breakdown is charged to the end customer. For some popular products such as Skype and Whatsapp we are given the impression that confidentiality is guaranteed by end to end encryption, but there is no good way to verify this and for Skype even evidence of the opposite. The Snowden revelations were an eye-opener on the amount and pervasiveness of the surveillance but after some initial indignation the world went back to business as usual.

The question relevant to this workshop is now: what can be our role in this as cryptographers and system designers and implementers? Our work is often used to protect the interest of the mentioned conglomerates rather than the interest of citizens. This is not a simple technical question but more of what we want our society to be and for that we should involve sociologists. But of course in any case the technical challenge of adding some security to these systems is formidable. The development cycle of hardware has become so complex with so many parties involved that the presence of trojans (hardware or software) is not improbable. What we can do is concentrate of sub-parts and try to do a good job there. Simplicity, open source and transparency are good guidelines in that. Education on security is also a good investment.

4.2 Double Arbiter PUF and Security Evaluation Using Deep Learning

Kazuo Sakiyama (The University of Electro-Communications, JP)

License  Creative Commons BY 3.0 Unported license
© Kazuo Sakiyama

Notes taken by Nele Mentens.

- Introduction
- Previous work of the presenter
 - Fault Sensitivity Analysis: fault analysis not requiring the ciphertext
 - RFID tag:
 - * complete analog/digital chip, largest crypto building block was Keccak, mutual authentication possible at a distance of 10 cm
 - * no difference in timing with or without Keccak because the analog part is dominant!
- Contributions of today's talk about PUFs:

- improvement of arbiter PUF (double arbiter PUF - DAPUF)
- Q-Class authentication (instead of using only 2 classes, so 0 or 1)
- using deep learning to evaluate the security
- Related work:
 - arbiter PUF -> ML attack -> N-XOR PUF -> double arbiter PUF -> ML attack on N-XOR PUF
- DAPUF details + results:
 - based on the idea of WDDL with complementary logic
 - can be extended to 3-1 and 4-1 DAPUF by XORing all comparisons
 - comparison of 2-XOR PUF and 2-1 DAPUF:
 - * uniqueness is improved compared to APUF
 - * randomness is not that good
 - comparison of 3-XOR PUF and 3-1 DAPUF:
 - * uniqueness and randomness good
 - * steadiness becomes more random for 3-1 and 4-1 DAPUF
- Q-Class authentication:
 - steadiness more random is good for ML resistance but not good for reliable authentication, that's why Q-class authentication is used, introducing multiple response classes
 - e.g. for some challenges all the responses are 0 (class 1), for some challenges all the responses are 1 (class 4), for some challenges a number of responses are 1 and a number of responses are 0 (class 2 and 3 depending on the percentage of 0/1)
 - 64 consecutive identical challenges are applied
 - the response to the verifier is the class number
 - experimental setting: use deep learning to build a clone and measure the responses, more secure if we go to 3-1 and 4-1 DAPUFs
- Q&A:
 - How to make sure that a fair comparison can be done between the two parts? Copy the lay-out + effort is made to make routing to the XOR balanced
 - Why XOR? XOR is chosen to introduce noise, because the FPGA does not generate enough noise
 - How is the training done for the ML attack? The class numbers are used
 - Why are the percentages of the 4 classes not exactly balanced? It turned out better for the ML attack resistance

4.3 IoT and Implementation Security

Thomas Eisenbarth (Worcester Polytechnic Institute, US)

License © Creative Commons BY 3.0 Unported license
© Thomas Eisenbarth

Notes taken by Georg Sigl.

Security and privacy are a major concern for the IoT: Computing is everywhere using as well as generating sensitive data. IoT creates tough power, size and cost constraints. Furthermore the attack surface increases due to the physical accessibility of the IoT devices. Spoofing of sensors will impact the physical world leading to safety and other risks. Designers have to find an optimum of cost, performance, and security. This leads to new solutions


like lightweight crypto and authenticated encryption. The main challenges on IoT are the communication interfaces, 30 years lifetime of devices, and physical attacks. The smart card industry is aware of physical attacks and has developed a certification process, which deals with that. This process is however very slow. Fail safe implementation techniques against SCA could speed up development and tools could be developed to apply them. Standardized tests like T-test or MIA should be used for security verification. Threshold implementations are a good standardized countermeasure which can be easily implemented even for lightweight crypto like Simon. How can we detect leakages in implementations? T-Test is a simple method developed by CRI. Thomas has improved the T-Test towards a paired T-Test, which eliminates common noise in pairs of measurements. But not only the IoT devices can be attacked. Attacks may be performed also against the cloud servers. An example are microarchitectural attacks: modern architectures introduce data dependent execution times due to performance optimizations. Cache attacks executed over the net will be a major threat for cloud systems.

Discussion

1. Discussion about the triangle security-cost-performance. There are other factors at system level which have to be taken into account. Examples are power consumption, latency requirements, or design time.
2. Is there a need for lightweight crypto? The use cases for lightweight crypto are very rare and may even become less if technology shrinks further. Only very low power applications with narrow communication distance can take advantage of lightweight implementations. The state cannot be reduced very much. Only implementation “tricks” help to reduce the area further. Another use case for lightweight crypto might be low latency. We currently have no good solutions for low latency cyphers in real time systems or for fast memory access.
3. What is the required randomness for threshold implementations? The required randomness is significant. Usually the area of the random number generators is not included in the area numbers. For Thomas’ Simon implementation the area requirement for the random number generation will be probably the same as for the crypto implementation. The generation of randomness for all kind of SCA protected implementations is not investigated sufficiently.
4. Do TI implementations help also against other attacks? Currently TI is dedicated to SCA only. Other more algorithmic countermeasures may be better against other attacks.

4.4 A minimalistic perspective on Public Key Encryptions

Roy Debapriya Basu (Indian Institute of Technology – Kharagpur, IN)

License  Creative Commons BY 3.0 Unported license
© Roy Debapriya Basu

Notes taken by Bilgiday Yuce.

Elliptic Curve Cryptography (ECC) is a promising alternative to RSA for resource-constrained systems. A lightweight (72 slices on a Spartan6 FPGA) and side-channel resistant ECC processor is proposed. The processor achieves low area by using One Instruction Set Computing (OISC) and utilizing the hardware macros on the FPGA. The processor uses one

instruction, SBN, to carry out most of the ECC operations. The execution time of Right Shift, Field Multiplication, and Shifting Key Register operations are not practically affordable when they are implemented with SBN instruction. Therefore, the processor includes low-area hardware accelerators for these operations. The processor achieves side-channel resistance with a low-cost operand swapping technique (IACR Report 925/2015).

Discussion

1. The processor has 5 variants of the SBN instruction, and it uses a 4-bit flag to choose between these variants. What would happen if we selected an undefined value of the 4-bit flag? What operation would be executed?
2. What are the basic motivations for OISC?
3. What is the area of OISC in comparison to NiosII, PicoBlaze, and MicroBlaze?
4. Are FPGAs suitable for IoT? Are not they more costly in comparison to ASIC?

4.5 Secure System Design in the IoT Regime

Sandip Ray (NXP Semiconductors – Austin, US)

License © Creative Commons BY 3.0 Unported license
© Sandip Ray

Notes taken by Yier Jin.

Intel starts the project titled "Secure, Intelligent, and Reliable Internet-of-Things (StaRT)". The key idea is how we can develop smart, reliable, trustworthy systems and applications with billions of (untrustworthy, potentially malicious) computing devices. The talk is related to this project.

We are (will be) in an IoT regime. A toy IoT example is a bad coffee detector. A general IoT hierarchical structure is presented where too many configurations of sensors, devices and gateways are available. From the system design perspective, every layer should have the computation capability so that the data does not need to travel all the way from end nodes to the cloud.

There are different stakeholders of the IoT landscape but their view/goals are not quite consistent for various reasons including lacking the standard. IoT market trends are introduced among them security is important. But security is not a stand-alone standard but within other metrics.

Assets in a smartphone is introduced. The attack surface of a smartphone is also introduced. Then the solution called Platform Security Assurance is introduced. One aspect is the linking between assets and accessing policies. For a more complexity fabrics example, to build a secure architecture requires the knowledge across all layers. A review to the security architecture reveals the complexity of such architectures.

The topic then moves to the unique features of IoT, and then to the post-quantum crypto. How can we configure the IoT so that it is resilient to post-quantum attacks in, say, 20 30 years.

Discussion

1. *Ingrid*: Coffee machine case. Local vs. cloud computing. *Sandip*: It depends on the data size to select powerful platform.

2. *Swarup/Patrick*: What we should focus at this moment? Shouldn't we handle the problem as a whole problem instead of individual problems? Why is it different from a traditional SoC design procedure? *Sandip*: Divide the problem into independent problems. Traditional metrics have specific guidelines. Things started changing since the smartphone where uncertainty raises. We are not sure the IP's specific role when designing the SoC for IoT since the functionalities vary.

4.6 Where Security Meets Verification: From Microchip to Medicine

Swarup Bhunia (University of Florida – Gainesville, US)

License  Creative Commons BY 3.0 Unported license
© Swarup Bhunia

Notes taken by Guido Bertoni.

A retrospective view: 1966 Apollo guidance computer versus today smartphone. In 50 years more powerful computers enabling new applications. From silicon micro to silicon nano electronics, next non-silicon technology.

Would the security be different from what we have in silicon techno? Could new techno provide better properties? And similar, better energy efficiency, reliability?

Question: Will the computer be a system of switches? Not known, there might be other paradigm New devices might have higher variation, good for TRNG or PUF. CMOS variation is usually very close to Gaussian, new techno might have other distributions. This could have an impact on the PUF construction.

Nano device could be very good to build memory but might have asymmetric access. Implication to side channel attack. Many new challenges and opportunity. Attacks on HW, from IC to IoT.

Define context of HW and the associated design flow. From design spec to IC fabrication, PCB fabrication to final customer. Many different points where an attack might take place. How to verify the design process?

Question: Are there evidence of inserted Trojan? Not official, there are some reports describing military chips with unexpected backdoors. Supply chain is mostly uncontrolled. PiRA, Puf in a package. Authentication method based on the resistance of PINs, like those in the GPIO.

Questions on how use cases and protocols. The physical method should be considered as a unique fingerprint of the device. Physical attack: modchip for game consoles. How to authenticate the PCB has not being modified?


Proposal: leverage JTAG in the field to check what JTAG sees in term of PCB and thus authenticate that PCB has not being altered. PCB integrity validation.

Microchip Trust Verification: recap on HW Trojan timeline. Bugs vs Malicious changes. Trust verification is quite different from traditional verification. Bigger challenge. Post production trust validation, could be destructive but only a fraction of chips might have Trojan so not useful. Non destructive, for test all of them. Statistical test (paper CHES 2009). Use of golden chip and adopt measurement of max freq and absorbed current.

New field: food and medicine counterfeiting detecting.

4.7 Security Metric for IoT

Yier Jin (*University of Central Florida, US*)

License  Creative Commons BY 3.0 Unported license
© Yier Jin

Notes taken by Thomas Eisenbarth.

Summary of Talk

Ethical hacking is still hacking, i.e. at best in a legal gray area

Audience comment: “That depends on the local legislation (e.g. jail-breaking a phone is legal in europe but not in the US).” Iot and CPS describe commercial vs. industrial (or Industrie 4.0) application but are technically the same thing.

Security aspects of them include:

- “Inconvenience level:” fridge sends spam
- “Privacy:” “big personal data” might be a problem in some scenarios (pacemaker)
- “Safety issue:” car IT is hacked
- “National security issue:” critical infrastructure

Examples from literature:

- physical devices (generator, a part of the infrastructure) is blown up remotely via network attack;
- a remotely controlled jeep drives into a ditch;
- a smart gun which fires at the wrong target;

Yier has also found many vulnerabilities for IoT devices:

- Nest Thermostat: jtag debug port was not shut down
- Smart home system: design is using DES, which can be brute-forced
- Smart band: in debug mode, security stack can be disabled.
- Roku device:
- F-Secure router can be upgraded to premium device, since HW and firmware are the same
- Smart Meter: remotely read data; replaces meter with fake one; interestingly, by replacing the smart part of the smart meter, the tamper evidence feature is also ‘hacked’

Take away:

- Device level hacking is a problem!
- Hackers can have a lot of patience: Reverse engineering is an option;

Solutions for IoT Security:

- Attack-oriented protections: check against a known list of attacks e.g. at: <http://www.hardwaresecurity.org/iot/database/>
- attacks sorted by level of system exploited
- Also provides a set of rules
- This is work in progress: other contributors welcome

Discussion

Q: What is TrapX? Attack or defense?

One device goes rogue and turns into WiFi router, forwards challenges and can monitor all communications.

Q: Hacking as described is like penetration testing? Do companies ask for this as a service?

Yes, they do. They also have bounty programs sometimes.

Q: Are any devices designed with security in mind?

They have security, but they do not properly implement, or they do not protect all levels. Example: firmware signing is becoming common, but secure boot is not: so one can replace verification code. Or they leave other obvious side channel, often simple ones.

Q: Do you follow a systematic approach for attack?

Yes, it is on the web site under Hands on Lab (see link above), but it is work in progress.

4.8 Eliminating timing side-channels in cryptographic software

Peter Schwabe (Radboud University Nijmegen, NL)

License  Creative Commons BY 3.0 Unported license
© Peter Schwabe

Notes taken by Swarup Bhunia.

Peter Schwabe presented his research findings on timing attacks in software, which focus on exploiting timing variations that depend on secret data. His attack models included cache attacks. He highlighted that timing attacks are serious concerns and among side-channel attacks this is only attack that can be remotely mounted on a hardware through a network. He used the square multiply example – a common operation in cryptography – to explain software timing attacks and its countermeasures through judicious low-cost software modifications. He highlighted that to prevent timing attacks, software need to follow only two rules.

1. Don't branch on secret data.
2. Don't access secret memory address.

He explained with examples how following these two simple rules through appropriate modification of a code can mitigate timing channels. For example a branch can be converted to arithmetic expression, which in specific cases can even make the code faster. He however mentioned that timing attack that exploits non-constant time integer arithmetic is not addressed by the proposed solution – e.g. for power PC processor. However, such attacks are not reported to happen in wild. He spent time to illustrate how load from and stores to addresses that depend on secret data, leak secret. A simple cache-timing attack does not reveal the secret address, it reveals just the cache line. He introduced constant-time equality comparison and used that to develop constant-time look-up table access from cache, which is effective for small tables. He mentioned that the effect of timing variance on oblivious RAM is worth investigating.

Audience showed tremendous interest in this topic. One question was on if we pre-load the table to cache, does it prevent the attack? Role of deliberate interrupt injection with potentially a fault attack in mounting timing attack was discussed and inferred as a topic which is worth looking further. Another question from the audience was: can you load part of the table (say half) to come to a good balance between performance and timing attack

resistance? The speaker agreed on that although pointed that compromise of security may not be acceptable in most applications.

Several other issues that were discussed are below. AES on composite field can potentially address the performance issue. Does Intel SGX architecture protect against timing attacks? They flush hashes between context switches. Yet it may not protect against timing attacks. Why do we see timing attacks in real world crypto algorithm every year? We see attacks against openssl. Do we need better compiler? Do we need better verification? Is speed really that important? Small hit in performance should be tolerated. Do we rethink cryptographic algorithm? Symmetric crypto which does not leak timing information. Everyone agreed that we need better hardware support – e.g. basic integer operation is constant time to deal with the attacks. Can we come up with new instruction like cache locking that helps in preventing time channels? The security-performance trade-off needs to be considered in this context.

4.9 MAFIA: Micro-architecture Aware Fault Injection Attack

Bilgiday Yuce (Virginia Polytechnic Institute – Blacksburg, US)

License  Creative Commons BY 3.0 Unported license
© Bilgiday Yuce

Notes taken by Hirokata Yoshida.

Abstract: Fault attacks are a serious thread to secure embedded software running on a wide spectrum embedded devices. In a fault attack, an adversary breaches the security by injecting faults into the underlying processor hardware and observing their effects in the output of the running software. For fault injection, the adversary temporarily alters the execution of instructions by running the processor beyond its nominal operating conditions. Therefore, an efficient fault attack requires hardware-level and software-level knowledge of the target system.

In this work, we propose an instruction fault sensitivity model that systematically captures the fault sensitivity of the processor pipeline for different instructions. This model enables us to gain insight into the most likely faults during the execution of an instruction, and to pinpoint the most sensitive points during the execution of a program. We also introduce a fault attack methodology called Microarchitecture Aware Fault Injection Attack (MAFIA), which makes use of the proposed model. In MAFIA, the adversary analyzes the executing of the target software on the target processor to design and implement a fault attack. The adversary starts with an algorithm-level analysis to determine high-level attack objectives. Then, the adversary examine the target software at the instruction-level to determine potential instructions for fault injection. Finally, the adversary analyzes the cycle-accurate execution of the target instructions with the help of the fault sensitivity model to determine best clock cycles and fault injection parameters to attack.

We demonstrated the efficiency of the proposed method on a LEON3 processor implemented on a Xilinx Spartan6 FPGA. As the target software, we attacked instruction duplication countermeasure, in which the sensitive instructions of a program duplicated and the consistency of the results of both copies are checked. It is assumed that such a countermeasure can only be broken by injecting multiple identical faults with expensive fault injection tools. We broke this countermeasure with single clock glitch injections by creating an instruction fault model for the LEON3 processor and using MAFIA.

The key conclusion is that one needs to consider both hardware and software layers to design efficient countermeasure against fault attacks targeting the embedded software. Currently, we are working on hardware/software methods to protect our processors from this kind of threats. As the future work, we are also planning to investigate the efficiency of MAFIA on the hardware duplication based countermeasures as well as on the multicore systems.

Presentation Fault attacks are an important class of hardware oriented attacks. Basically, they inject the faults into the operation of the device. They analyze the response of the device of this fault injection to breach the security of the device. What is software fault attack? He first analyzes algorithm. He makes the assumption on the faults. We call this fault model. Let's take a close look at fault injection process. It is executed sequence of the instruction. The attacker changes the operation condition of the device e.g. the attacker can change voltage of the device, clock signal of the device. This faults injection affects execution of instructions then instructions are propagated to software. The current fault model does not consider the hardware aspects of microprocessor. This creates gap between injected faults and assumed faults.

We propose instruction fault sensitivity model for a RISC pipeline. Using this kind of method, we can pinpoint what kind of fault we will inject into the device. Our target countermeasure is instruction duplication countermeasure. Basically, in this countermeasure, to protect any instruction, we execute instruction twice and then we compare the results of these two executions of the same instruction. and if result doesn't match, we alarm wrong signal. Taking the example of 7 stage 32-bit pipeline, how attack works is shown. The effect of data dependencies is studied and it is shown that this leads to additional opportunities for fault injection. Instruction fault sensitivity model of each instruction is explained for a specific cycle. Memory stage of the load instruction takes 7.5 nano seconds. fetch stage takes 3.5 ns. So we make such a table for each instruction of the processor. After getting this model at once, for a processor, we can use this model several times to apply different patterns. After our experiments of work., we achieve all of our scenario and we break this instruction duplication countermeasure.

The conclusions are as follows: If we want to design efficient prototypes against embedded software, we need to consider both hardware and software aspects of this attack. Considering these aspects, we will see existing countermeasures are vulnerable against this new type of attacks. We need to countermeasure if we want to protect our devices against this kind of attack. Currently our recent work is trying to find a better way of protecting processor against this kind of attacks.

Discussions

1. What if instructions and hardware are duplicated? → It depends on the resulting architecture. It depends on the all the instructions in the pipeline.
2. What about other faults like EM faults. In this case, you need to change the fault sensitivity model. This case is timing.
3. Examples you showed just try instructions in pipeline, for the other things, you could also check some there. May be we find a good cycle to attack.
4. What if I use identical instructions? You are saying that to protect two more instructions, the requirement is to protect any one of them. Do you think this is a viable countermeasure?

4.10 Optical Interaction through Chip Backside with Nanoscale Potential

Christian Boit (TU Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Christian Boit

Notes taken by Shahin Tajik.

Although there is some basic interaction and attack possibility through the frontside of the chip, optical interaction with the active devices is not so fruitful, because the metallization layers obstruct the optical paths. However, the light which is reflected or emitted from the backside of the chip faces no optical obstruction. The utilization of the flip-chips is another motivation for us to access the chip through the backside.


Photon emission is detectable through the backside during switching events of transistors. With the help of this technique the IC can be debugged. Furthermore, the signal flow can be followed on the chip with the help photon emission and Picosecond Image Circuit Analysis (PICA). On the other hand, we need to stimulate some areas with lasers to interact and debug the chip. In this case, part of the light will be reflected back and part the light will be absorbed. The latter can create voltage and current sources, which can lead to fault injection attacks and read-out of the data. Some techniques deploy wavelengths, which can just create heats but no electron-hole pairs in the silicon. This technique is called thermal laser stimulation (TLS). Using this technique one can read-out the stored values in the SRAMs. Note that there is no need for a clock signal. On the other hand, the reflected light will be modulated and can be used for the contactless probing of the signals. Different space charge layers on transistors are based on “on” or “off” state of the transistors and the reflected light has linear relations with voltage of the chip.

Photon interaction is bounded with the absorption of silicon for different wavelengths. Optical techniques have been taking advantage of the high infrared (IR) transmission for wavelengths $> 1 \mu m$. One of the hardest challenges for optical IC debug techniques is the ever increasing miniaturization. 10 nm and smaller technologies are the current feature sizes. Resolution R in the optical interaction is a function of wavelength λ : $R \propto \lambda / (2NA)$. NA is the Numerical Aperture (in air < 1) with $\lambda = 1 \mu m$. R is at best around 500nm if the chip is simply put into the optical path of the instrument. By introducing solid immersion lens (SIL) on back surface, the NA is increased by the index of refraction n_{SIL} . For silicon and $\lambda = 1 \mu m$, n is about 3.5, resulting in a maximum R of around 150 nm. The smallest technology announced by Intel, which can be probed is 10 nm. Note that the technology length is the feature size of the transistor’s gate and not the actual size of the transistor itself. The size of the FinFET technology is decreasing. In 2025 it will be 1.8 nm (pitch= 20nm). Slow decrease of the pitch helps us to still debug the chip.

If the resolution needs to increase further, the NA part of the equation cannot be enhanced more. Wavelength reduction remains the only possible approach. Below 10 nm we need to use shorter wavelengths. However, the absorption of silicon is problematic and most of the photons are absorbed by silicon substrate. We still can thin the substrate to 10 micrometer to reduce the absorption of the light. Moreover, in order to probe with visible light the silicon-based SIL should be replaced by GaP-based SILs.

4.11 How Secure are Modern FPGAs?

Shahin Tajik (TU Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Shahin Tajik

Notes taken by Bilgiday Yuce.

Most of the modern FPGAs keep their configuration in volatile, on-chip SRAM cells. Therefore, the configuration needs to be loaded into the on-chip SRAM cells from an off-chip Non-Volatile Memory (NVM) whenever the FPGA is powered on. In an untrusted environment, the transfer of configuration bitstream from off-chip NVM to on-chip SRAM cells may leak the design information. To mitigate this problem, FPGA vendors use bitstream encryption. In a trusted environment, the bitstream is encrypted and it is written into the NVM. Meanwhile, the encryption key is embedded into the FPGA. At each power-on of the FPGA in the untrusted environment, the encrypted bitstream is transferred to the FPGA from the NVM and it is decrypted using the embedded key. In this work, the target FPGA use soft PUFs for key storage and DPA-resistant decryptor for bitstream decryption. Although this approach provides protection against DPA and semi-invasive front-side attacks, it does not provide protection against semi-invasive back-side attacks such as Laser Voltage Probing (LVP) and Laser Voltage Imaging (LVI). LVP enables us to probe electrical signals of interests by just pointing the laser beam to the circuit node of interest. LVI allows us to create a 2D map/image of the electrical nodes operating at a specific frequency, and filtering out the remaining electrical nodes operating on a different frequency. This work provides two LVP-based attacks against FPGAs during the configuration. In the first attack, an adversary probes the contents of PUF response and key registers with LVI. Therefore, the adversary can extract these values from the FPGA and decrypt the bitstream. The second attack uses a combination of LVI and LVP to characterize the PUF that is used as key storage. After characterizing the PUF, the adversary can clone its functionality and retrieve the decryption key. These two attacks were demonstrated on Altera Cyclone IV FPGAs. An existing method for protecting devices against laser-based attacks is using Silicon light sensors. However, the light sensors are ineffective against the LVP and LVI techniques proposed in this work because the laser beam has a larger wavelength than the silicon bandgap. A possible countermeasure would be assigning random values to registers in the case of a reset event. Using a special coating on the backside of the FPGA can be another alternative protection. This work also proposes use of a ring oscillator network as a countermeasure against LVP. In the proposed method, a network of ring oscillators with virtually equal frequencies are deployed on the FPGA. Using LVP will then shift the frequency of the ring oscillator in the vicinity of the probed area. Therefore, the frequency deviation from the average frequency of the ring oscillator network can be used to detect LVP and raise an alarm signal. Preliminary results for this detection technique seems promising and a research is currently going on. As a result, two backside attacks based on back-side LVP and LVI are proposed to reverse engineer the key of the FPGAs. This shows the need for countermeasures for this kind of attacks.

Questions/Comments:

- Q: Why do PUFs are a better choice for key storage than BBRAMs or eFuses?
- Q: How to make sure PUF is really loaded correctly? Is it possible to change bitstream?
- Q: Would aging of PUF be a problem from the key storage point of view?
- Q: If you deploy ring-oscillator-based detector through the whole chip, they would cause a significant heating. Would it effect the measurements?
- Q: How big is the laser machinery used for LVI and LVP?

- Q: How many measurements are need to create LVI map of the electrical nodes of interest?
- Q: How quickly can you turn on/off the ring-oscillator-based detector?
- Q: How many ring oscillators did you use in ring-oscillator-based detector?

4.12 Detection and Prevention of Side-Channel Attacks

Naofumi Homma (Tohoku University, JP)

License © Creative Commons BY 3.0 Unported license
© Naofumi Homma

Notes taken by Debdeep Mukhopadhyay.

- Local EM attacks
 - Using Microprobes, observation of precise and local EM leakage
 - Beyond conventional leak assumptions
- Measurable leaks by microprobes
- Most of the countermeasures can be defeated because their leak assumptions are not met
- Countermeasures: Transistor level balancing, active shielding, special packaging
- Overhead, vulnerability still exists, use high resolution, or reverse-side attacks
- EM attack sensor-CHES 2014
- Idea: sense the presence of probing by observing electrical coupling, EM field variation LC oscillation frequency shifts due to Mutual Inductance, M.
- Based on this idea, proposal of Dual-coil sensor architecture
- No frequency reference needed. Observe variance of the coil oscillation frequency, f_{LC}
- Sensor core architecture uses two coils. The detection circuit subtracts the two coil oscillation frequencies.
- The experimental set up was elaborated fabricated in 0.18 μ CMOS.
- Detection, probe diameter 0.2 mm, 0.3 mm.
- Greater than 1% variation in f_{LC} can be detected. The detection range is 1 mm.
- Overhead: AES core 24.3 K, sensor 0.3 K, overhead is 1.2% Power: +9%, lesser than classical countermeasures

Limitations

- Attack may be to keep difference of LC oscillation frequencies during measurement, but attacker cannot see oscillation frequency.
- Detection vertical distance is 0.1 mm. Should be ok for front end attack. Conventional EMAs over chip package still possible. Combination with classical countermeasures important.
- Scaling on EM attack sensor: Other non-crypto algorithms: could compensate algorithm/gate-level countermeasures. Also applicable for other platforms, FPGAs, and even advanced CMOS technologies.
- Advanced CMOS: Oscillation frequency would increase. Magnetic flux passing through probe would decrease is the probe diameter is smaller.
- Consequence of smaller probes: Frequency shift amount would decrease: Digital counter may not detect the frequency shift.
- Improvement: Extend detection time. Extend detection process time to accumulate smaller shift amount differences.

- Time extension enables to detect small frequency shifts even probing from back side of LSI, around 0.5 mm.
- Overhead: Additional bits are required to counter and subtractor in addition to time delay, Delay: +12.1%, Area: +1.6%.
- Cancel out timing overhead by Simultaneous Operations of Crypto Core and Sensor
- Works because current flowing in crypto core minute and omnidirectional.
- Another idea for speed up: Frequency count by Time to Digital Converter (TDC)
- Conclusions: Sensing technology for side channel attacks
- Challenge for scaling: Application to other platforms, other technologies.

Questions

- Which Microprobes: 0.1 mm diameter hand made microprobes., Langer probes.
- Can be used for laser fault detection?
- Detection coils of the size does it depend on the probe size?
- EM attack sensor
- Scaling on attack sensor: What may happen in advanced CMOS technology

4.13 Implementation Security through Dynamic Reconfiguration

Nele Mentens (KU Leuven, BE)

License  Creative Commons BY 3.0 Unported license
© Nele Mentens

Notes taken by Patrick Schaumont.

Research Interests

- Efficient Crypto Coprocessor Design
- Design automation/ design space explo for crypto hardware
- Partial reconfiguration for security purposes

Implementation Security through dynamic reconfiguration

Why?

- Power Analysis attacks correlate power and secret data Fault Analysis attacks correlate faults and secret data
- Approach is to make the hardware dynamically reconfigurable without changing the IO behavior of the system

Randomly Reconfigurable Architectures

Symmetric Key Algorithms

- Randomized Pipelining within a round
- Randomize Pipelining within an SBOX

Move pipeline registers around to randomize power dissipation

Q(Patrick): Can we accumulate power over sufficient clock cycles to remove randomization effect?

Public Key Algorithms: Many possible randomization elements

- Parameters
- Circuit
- Order of Operations, randomized addition chains

Randomized ECC 25519 Q(Peter): For Montgomery representation? Yes

The design space for PK Algorithms is much larger, so you can use evolutionary algorithms and design automation to search feasible solutions

E.g. Randomized Addition Chains can be synthesized

Reconfigurable Technology. SRAM Configuraton Memory: Partial FPGA reconfiguration, such as in Xilinx, allows to reconfigure part of the FPGA. The new FPGA-SoC allows the FPGA to be reconfigured from within the processor. SRAM reconfiguration is coarse grain, with minimum width and height in the FPGA fabric. This is relatively slow because of serial (sequential) loading of partial bitstream.

CFGLUTs: 5-1 LUTs. Can be reconfigured directly by the user logic. This is very fast: 32 cycles to reconfigure one LUT. Routing cannot be reconfigured this way.

Virtual Reconfigurable Circuits: Circuits specifically designed for dynamically reconfigurable designs (Coarse grain reconfigurable asics)

Generation of New Configurations

- Offline generation: a fixed number of configurations stored in SRAM
- Online generation: (seems) not feasible at this moment for realistic designs. However, dedicated designs using CFGLUT may be feasible.
- Parametrizable bitstreams

Summary of Design Parameters

- Reconfiguration Time
- Reconfiguration Overhead
- Reconfiguration Granularity
- Reconfiguration Frequency
- Target Platform
- Number of Configuration Options

Q(Patrick): Connection to Whitebox Crypto? A: Whitebox does not work (or is obscure)

Q(Tahin): Can the attacker tamper with reconfiguration? Partial bitstream is well defined (in position), so could be tampered in principle.

Q(Roy): Can this be used for Trojans?

4.14 Propagation of Glitches and Side-channel Attacks

Guido Bertoni (ST Microelectronics – Agrate, IT)

License © Creative Commons BY 3.0 Unported license
© Guido Bertoni

Notes taken by Ingrid Verbauwhede.

Glitches represent a great danger for hardware implementations of cryptographic schemes. Their intrinsic random nature makes them difficult to tackle and their occurrence threatens side-channel protections. Although countermeasures aiming at structurally solving the problem already exist, they usually require some effort to be applied or introduce non-negligible

overhead in the design. Our work addresses the gap between such countermeasures and the naïve implementation of schemes being vulnerable in the presence of glitches. Our contribution is twofold: (1) we expand the mathematical framework proposed by Brzozowski and Ésik (FMSD 2003) by meaningfully adding the notion of information leakage, (2) thanks to which we define a formal methodology for the analysis of vulnerabilities in combinatorial circuits when glitches are taken into account.

4.15 Threshold Implementations

Svetla Nikova (KU Leuven, BE)

License  Creative Commons BY 3.0 Unported license
© Svetla Nikova

Notes taken by Chen-Mou Cheng.

Masking is not secure in CMOS because of glitches. TI: provably secure masking scheme based on secret sharing and multiparty computation. It was initially proposed for 1st order DPA but later on extended to any order. It is secure even in face of circuit glitches.

TI conditions: Correctness, non-completeness, and uniformity. TI techniques: It is important to decompose a complicated (high-degree) function to reduce its degree (and hence circuit complexity). There are automatic tools for decomposing and sharing. Other important optimizations include reusing and factorization.

AES: success; Keccak: efficient implementation proposed, ongoing work to solve uniformity issues. TI has also been applied to other ciphers such as Katan, Simon, Speck, ...

Higher order TI: any combination up to d component functions must be independent of at least one input share. Naturally, the number of shares increases. The cost for higher order TI is roughly linear in the degree for Katan-32 but can increase more rapidly for other, more complicated ciphers such as AES. Often time we can trade off between area and the amount of required randomness.

Discussions

Q: Is it possible to combine RNG & TI to make it more efficient?

A: It is possible to reuse randomness, e.g., in TI of Keccak. However, need to be careful about entropy & dependencies, etc.

C: Would be interesting to compare against glitch-free circuits in terms of area and randomness costs.

C: In practice, crosstalk can decrease the security of TI (and also other masking schemes).

4.16 Secure Scaling, Scaling Securely

Francesco Regazzoni (University of Lugano, CH)

License  Creative Commons BY 3.0 Unported license
© Francesco Regazzoni

Notes taken by Patrick Schaumont.

Secure Scaling, Scaling Securely

Handling the Scaling

- We need Design Automation. Early chips were designed by hand. Modern chips are designed with extensive use of design automation.
- We need Design Automation for Security. Security is considered at the end of the design. Cost and Time to Market are most important. Avoid Security Pitfalls. Handle the Complexity. And, most importantly, use standard design commodities (= tools).
- Automatic Application of Countermeasures. Input = Unprotected Algorithm + Countermeasure. Output = Algorithm where the countermeasure is applied. That does not mean that the result is a protected algorithm. It is only protected if the countermeasure is correct.
- Example: Software Automation with Compiler. Start from Software Implementation, do Information Leakage Analysis, Transformation Target Identification, do Code Transformation.
- Example: SC Leakage analysis by mutual information analysis. Then transform the instructions.
- Example 2: Protection using Custom Instructions. Implement Custom Instructions using protected logic.
- Example 3: Verification. Input: Algorithm, countermeasure. Output: Check that countermeasure is correctly applied.
- E.g. masking of an expression

$$s = p \text{ exor } (k \text{ exor } m)$$

→ compiler produces

$$s = (p \text{ exor } k) \text{ exor } m$$

which is wrong

- Example describes SLEUTH.
- Risks of Scaling
- Can fault attacks be applied to subthreshold technology?
- Subthreshold is used for low power.
- Fault attack: 0.8mV interval enables single-byte fault
- The interval depends on chip and temperature. In subthreshold, higher temperature gives lower threshold voltage.
- Scaling Securely
- Current photon based entropy source
- Initial Tests
- 512 x 128 single photon detectors Photon passes through semitransparent mirror.
- Parallel readout and Von Neumann Postprocessing
- Initial Tests pass the NIST tests when Von Neumann is included.

4.17 Scaling of Implementation Attacks

Georg Sigl (TU München, DE)

License  Creative Commons BY 3.0 Unported license
© Georg Sigl

Notes taken by Francesco Regazzoni.

Advanced EMA attacks Advanced Laser Attack Countermeasures

- Attack Setup: The station for mounting attacks has a fixed probe and a moving base to position the FPGA. The attack is carried out on the FPGA from the front.
- Advanced EMA attacks: The target is an assignment operation during an ECC computation. The values are stored in registers, which have a physical position on the die. If the probe is placed very close to the register storing value A, there will be a significant variation on the probe when there is a change in the value of A. If there is a dependency of the register usage and the secret information, this can be used for an attack.
- In the case you will have N iterations depending on the length of the secret, it is possible to split the traces collected into N portions, each of them corresponds to a calculation of a single bit. The traces are then assigned to two different sets to recover the secret key.
- The starting point of the attack is the identification of the hot spot for measurement (the points which allow to get the key). This cartography operation requires some time.
- Principal Component Analysis can improve the key recovery process. The PCA generates N principal components (where N is the length of the secret). If components clearly divide the traces into two clouds then the key is recovered easily. Reported results show that component 4 produces the best result.
- Single vs multiple probe (3 probes): with multiple probes, more PCA components contain information.
- With scaling of technology, EM side channel will be still feasible. If scaling means using better and more probes, EM is likely to be more dangerous.
- Advanced laser attack: A two laser setup has been internally developed, emitting beam for attacking front side and another beam for attacking back side. The attack was carried out on SPARTAN-3 at 90nm and SPARTAN-6 at 45nm. It is still possible to identify block RAMs and flip single bit at 45nm, but, comparing with the 90nm it is getting harder.
- An attack on AES with the infection countermeasure has been performed. To attack this countermeasure it is needed to shoot the laser at both computation units, or at the comparator. The target was SPARTAN-6. To identify the position where to shoot with the laser it is needed to generate a map with block RAM and flip-flop (which is design independent) and the flip flops used in the applications (which is design dependent). From over 80'000 shoots, 229 were exploitable faults (remember that even a single exploitable fault is sufficient).
- Attack summary: attacks are feasible also for smaller technologies, although shrink would make the shoot less precise. Also laser can scale by increasing the amount of lasers, e.g..
- As countermeasures two approaches can be applied: Attack detection (sensors, detectors, error detectors) and Attack prevention (Limiting the amount of repetition, add randomness, hiding). Both have to be taken into account vertically, spanning over cryptographic algorithms, System, Technology.
- Each countermeasure should address the following questions:
 - What is the coverage?
 - Which one is the best layer to implement it?
 - Is it possible to implement an orthogonal countermeasure?

4.18 Crypto, Integration, Technology: Good, Bad, Ugly?

Debdeep Mukhopadhyay (Indian Institute of Technology – Kharagpur, IN

License © Creative Commons BY 3.0 Unported license
© Debdeep Mukhopadhyay

Notes taken by Sahin Tajik.

Talk Abstract: The talk presents a glimpse on the effects on physical security as a result of integration, and technology scaling. Technology scaling and improvements in computer architecture have varying effects on side channels: 1) Good, when the attacks are hindered, 2) Bad, when the attacks are aided, and 3) Ugly, when it is difficult to characterise the consequences! The talk presents few case studies to illustrate the dependencies. Differential Fault Intensity Attacks (DFIA) is a menacing class of fault attack against cryptosystems. Fault attacks are typically countered using concepts of redundancies, which modern computer architectures seem to support owing their increases parallelism. However, malicious fault attackers with the capability of repeating fault injections with precise control can potentially defeat such classical countermeasures. In this context, Fault space transformation (FST) is proposed as a new class of countermeasures against these evolved fault attacks. The objective of FST is to reduce the probability of obtaining useful faults which can bypass standard countermeasures. Although, device scaling increases the failure rates, the chances of obtaining useful faults wrt. FST are further reduced, to make FST a promising fault attack countermeasure. The talk also shows an example of cache timing attack on a 128 bit cipher, known as Clefia. The cipher has small tables which reduces the chances of a cache attack, as the number of cache misses are a constant per encryption. However, modern computer architectures provide artefacts for parallel service of cache misses. The structure of block ciphers provide opportunity for out of order loading of tables in a round, but not across the rounds due to dependencies. This leads to timing variations because though the total number of cache misses are constant, the penalty due to cache misses in a round can be ameliorated compared to cache misses across rounds. Thus the distribution of cache misses also plays a role to determine the information leakage, and it was shown that modern computer architectures (after Intel Pentium-3) were prone to cache timing attacks on Clefia. Finally, the talk comments on the recently discovered bug on DRAM chips, typically observed post 42 nm technology. Repeated discharging and recharging of the cells of a row in a DRAM bank results in leakage of charge in adjacent rows. If repeated enough times, typically before the automatic refresh in adjacent rows, causes flips of bits. This example typically shows an instance where process integration leads to new avenues or sources of attacks. In conclusion, the talk tries to encourage the study of these effects to evolve systems which are more secured against the powerful side channel attacks by taking advantage of modern day architectures, while being aware of the vulnerabilities introduced by them.

- We would like to evaluate the physical security of crypto across integration and technology. We consider two cases: 1. cache attacks ,2. fault injection attacks.
- Cache memory leaks information based on a cache hit. In this case, the access time and power consumption will be less than the case, where there is a cache miss. As a result, the attacker can launch a cache attack by measuring the total time for the encryption. This technique has been used to attack a remote server. For instance, in the Bernstein's cache timing attack, we try to invoke the AES encryption by .xing part of the input, and randomize other parts of the inputs and obtain the total time for the encryption. By guessing the key and calculating the time correlations, we can break the AES. Smaller table

sizes make the cache attack harder. However, it has been shown that the cache attacks are possible even on the ciphers with small Sboxes. It has been shown that the CLEFIA cipher can be attacked. The processor's aim is to reduce the true miss penalty by using speculative loading, prefetching, out-of-order loading, parallelization and overlapping. However, the attack has been tested successfully on some platforms, such as Intel Core 2, Intel Atom, Pentium 4, Xeon – core 2 servers. However, the attack against Pentium 3 was unsuccessful.

- On the other hand, cryptographic algorithms can be analyzed using faults. Concurrent error detection, infection and data encoding can be the countermeasures against faults. However, naive redundancy can be broken by improving fault collision probability. A smaller fault space enhances the fault collision probability. A non-uniform probability distribution of the faults in the fault space also enhances the fault collision probability. With increase in the bias, the collision probability increases. Transforming the fault space implies that the adversary cannot beat the countermeasure by merely introducing the same fault twice. It is most unlikely that the transformed fault space will have a one-to-one correspondence in terms of the bias with the original one. Mathematically, the expected fault collision probability over all possible transformations is the same as for the uniform fault models.
- But what is the impact of technology scaling on the failure rates? Failure rate of a 65 nm device is 316% times more than the same at 180 nm. However, the growing parallelism can give rise to several levels of redundancy. Natural hardware faults can be detected by dual-modular-redundancy (DMR) and triple-modular-redundancy (TMR). Moreover, malicious faults can be detected by Fault Space Transformation. Fault Space Transformation can be applied on different cores exploiting the available redundancies.
- RowHammering: Repeated discharging and recharging of the cells of a row results in the leakage of charge in the adjacent rows. If it is repeated enough, typically before occurrence of the automatic refreshment in adjacent rows, causes flips of bits, which is known as RowHammer. This development appears to coincide with the upgrade to 42 nm productions for the DRAM chips. A small cell can hold a limited amount of charge, which makes it more susceptible to data loss. The close proximity of the cells introduces electromagnetic coupling effects. Higher variation in the process technology increases the number of the outlier cells that are susceptible to the cross talks. This is an example of converting a reliability issue to an attack.
- We conclude that the cryptographic techniques often need to be considered based on the underlying platforms: Improved architectures may give benefits or open new threats. We propose Fault Space Transformation as a novel fault tolerance technique for the block ciphers. Parallel architecture offers opportunities for redundancy based schemes. The utilization of Fault Space Transformation may be a good idea! Finally, technology scaling offers more variability and improved failure rates: The controlled usage of the faults is a major challenge and it could lead to suitable hammers to create faults and threaten actual systems.

4.19 Smart Card Secure Channel Protocol

Joan Daemen (*ST Microelectronics – Diegem, BE*)

License © Creative Commons BY 3.0 Unported license
© Joan Daemen

Notes taken by Patrick Schaumont.

Joan described a protocol that is used to implement the 'secure channel' of a smart card. A secure channel is the control/data link between the smart card and a central smart card management system (bank, network provider in case of SIM, ...). The entity on the card that implements the card end of the protocol is called "security domain". At the central system there is a hardware security module (HSM) that performs the central server side of the protocol.

The objective of the secure channel protocol is authorize commands send to the card, from the HSM to the security domain, and to authenticate responses returned from it, from the security domain to the HSM. Furthermore, the secure channel protocol enables the secure transfer of keys (e.g. installing a new vendor key). Some background information may be found from the GlobalPlatform Wiki (<https://sourceforge.net/p/globalplatform/wiki/Home/>)

The cryptographic primitive for authorization and authentication is a message authentication code (MAC). For authorization, the MACs are generated by the HSM and verified by the security domain; these MACs are called C-MAC. For response authentication, the MACs are generated by the security domain and verified by the HSM; these MACs are called R-MAC.

- The basic protocol for authorization (C-MAC) is a chain of MAC, where a C-MAC is repeatedly computed over new input data, which can be either a command or a response, and a secret MAC key. Hence, the complete chain of commands, or responses, is chained. Any tampering with the chain can be detected by the security domain. The computing of a C-MAC chain also enables to integrate non-functional context data, such as additional application identifiers and nonces.
- The basic mechanism for authentication is a single R-MAC, computed over the sequence of command-responses generated by the security domain. The R-MAC is verified by the HSM.

The protocol has several particular features compared to text-book challenge/response protocols.

- The protocol makes use of predictable nonce-counters, embedded in the MAC chain, which prevent replay of commands or responses. However, the protocol does not make use of random numbers and does not have freshness. Every C-MAC sequence is fully predictable. On the other hand, the R-MAC sequence contains freshness, since the command data (delivered through C-MAC) can include a challenge.
- The card has a key hierarchy which generates a sequence of 'working keys' for RMAC as well as for CMAC. These keys are derived from a master key using a one-way function, the card-id, and the C-MAC sequence counter. This means that the working keys are predictable, but also continuously updated. The working keys are stored in non-volatile memory and derived at runtime when needed. A new R-MAC working key can only be created when the C-MAC sequence was successfully completed.
- There is a ratification on the working keys, both R-MAC and C-MAC. The working keys are blocked when there is an excessive number of MAC failures (either R-MAC or C-MAC). Once the working key is blocked, the secure channel can no longer be used.

Participants

- Debapriya Basu Roy
Indian Institute of Technology –
Kharagpur, IN
- Lejla Batina
Radboud Univ. Nijmegen, NL
- Guido Bertoni
ST Microelectronics – Agrate, IT
- Swarup Bhunia
University of Florida –
Gainesville, US
- Christian Boit
TU Berlin, DE
- Chen-Mou Cheng
National Taiwan University –
Taipei, TW
- Joan Daemen
ST Microelectronics –
Diegem, BE
- Jia Di
University of Arkansas –
Fayetteville, US
- Thomas Eisenbarth
Worcester Polytechnic Inst., US
- Naofumi Homma
Tohoku University, JP
- Yier Jin
University of Central Florida –
Orlando, US
- Nele Mentens
KU Leuven, BE
- Debdeep Mukhopadhyay
Indian Institute of Technology –
Kharagpur, IN
- Ventzislav Nikov
NXP Semiconductors –
Leuven, BE
- Svetla Petkova-Nikova
KU Leuven, BE
- Bart Preneel
KU Leuven, BE
- Sandip Ray
NXP Semiconductors –
Austin, US
- Francesco Regazzoni
University of Lugano, CH
- Kazuo Sakiyama
The University of
Electro-Communications, JP
- Patrick Schaumont
Virginia Polytechnic Institute –
Blacksburg, US
- Peter Schwabe
Radboud Univ. Nijmegen, NL
- Georg Sigl
TU München, DE
- Shahin Tajik
TU Berlin, DE
- Ingrid Verbauwhede
KU Leuven, BE
- Hirotaka Yoshida
AIST – Tsukuba, JP
- Bilgiday Yuce
Virginia Polytechnic Institute –
Blacksburg, US



Next Generation Sequencing – Algorithms, and Software For Biomedical Applications

Edited by

Gene Myers¹, Mihai Pop², Knut Reinert³, and Tandy Warnow⁴

1 MPI – Dresden, DE, myers@mpi-cbg.de

2 University of Maryland – College Park, US, mpop@umd.edu

3 FU Berlin, DE, knut.reinert@fu-berlin.de

4 University of Illinois – Urbana-Champaign, US, warnow@illinois.edu

Abstract

Next Generation Sequencing (NGS) data have begun to appear in many applications that are clinically relevant, such as resequencing of cancer patients, disease-gene discovery and diagnostics for rare diseases, microbiome analyses, and gene expression profiling. The analysis of sequencing data is demanding because of the enormous data volume and the need for fast turnaround time, accuracy, reproducibility, and data security. This Dagstuhl Seminar aimed at a free and deep exchange of ideas and needs between the communities of algorithmicists and theoreticians and practitioners from the biomedical field. It identified several relevant fields such as data structures and algorithms for large data sets, hardware acceleration, new problems in the upcoming age of genomes, etc., which were discussed in breakout groups.

Seminar August 28 to September 2, 2016 – <http://www.dagstuhl.de/16351>

1998 ACM Subject Classification D.2.11 Software Architectures, D.2.13 Reusable Software, D.2.2 Design Tools and Techniques, E.1 Data Structures, J.3 Life and Medical Sciences

Keywords and phrases Cancer, DNA Sequence Assembly, Expression Profiles, Next Generation Sequencing, Sequence analysis, Software Engineering (Tools & Libraries)

Digital Object Identifier 10.4230/DagRep.6.8.91

Edited in cooperation with German Tischler


1 Executive summary

Gene Myers

Mihai Pop

Knut Reinert

Tandy Warnow

License  Creative Commons BY 3.0 Unported license
© Gene Myers, Mihai Pop, Knut Reinert, and Tandy Warnow

Motivation

In recent years, Next Generation Sequencing (NGS) data have begun to appear in many applications that are clinically relevant, such as resequencing of cancer patients, disease-gene discovery and diagnostics for rare diseases, microbiome analyses, and gene expression profiling, to name but a few. Other fields of biological research, such as phylogenomics, functional genomics, and metagenomics, are also making increasing use of the new sequencing technologies.



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

Next Generation Sequencing – Algorithms, and Software For Biomedical Applications, *Dagstuhl Reports*, Vol. 6, Issue 8, pp. 91–130

Editors: Gene Myers, Mihai Pop, Knut Reinert, and Tandy Warnow



Dagstuhl Reports

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

The analysis of sequencing data is demanding because of the enormous data volume and the need for fast turnaround time, accuracy, reproducibility, and data security. Addressing these issues requires expertise in a large variety of areas: algorithm design, high performance computing on big data (and hardware acceleration), statistical modeling and estimation, and specific domain knowledge for each medical problem. In this Dagstuhl Seminar we aimed at bringing together leading experts from both sides – computer scientists including theoreticians, algorithmicists and tool developers, as well as leading researchers who work primarily on the application side in the biomedical sector – to discuss the state-of-the art and to identify areas of research that might benefit from a joint effort of all the groups involved.

Goal of the seminar

The key goal of this seminar was a free and deep exchange of ideas and needs between the communities of algorithmicists and theoreticians and practitioners from the biomedical field. This exchange should have triggered discussions about the implications that new types of data or experimental protocols have on the needed algorithms or data structures.

Results

We started the seminar with a number of *challenge talks* to encourage discussion about the various topics introduced in the proposal. Before the seminar started we identified three areas the participants were most interested in, namely:

1. Data structures and algorithms for large data sets, hardware acceleration
2. New problems in the upcoming age of genomes
3. Challenges arising from new experimental frontiers and validation

For the first area Laurent Mouchard, Gene Myers, and Simon Gog presented results and challenges; for the second area Siavash Mirarab, Niko Beerenwinkel, Shibu Yooseph, and Kay Nieselt introduced some thoughts; and finally, for the last area, Jason Chin, Ewan Birney, Alice McHardy, and Pascal Costanza talked about challenges. For most of those talks the abstracts can be found below. Following this introductory phase, the participants organized themselves into various working groups the topics of which were relatively broad. Those first breakout groups were about

- Haplotype phasing
- Big data
- Pangenomics data representation
- Cancer genomics
- Metagenomics
- Assembly

The results of the groups were discussed in plenary sessions interleaved with some impromptu talks. As a result the participants split up into smaller, more focused breakout groups that were received very well. Indeed, some participants did already extend data formats for assembly or improved recent results on full text string indices.

Based on the initial feedback from the participants we think that the topic of the seminar was interesting and led to a lively exchange of ideas. We thus intend to revisit the field in the coming years in a Dagstuhl seminar again, most likely organized by different leaders of the field in order to account for these upcoming changes. In such a seminar we intend to encourage more people from clinical bioinformatics to join into the discussions.

2 Table of Contents

Executive summary

<i>Gene Myers, Mihai Pop, Knut Reinert, and Tandy Warnow</i>	91
--	----

Overview of Talks

Computational challenges in cancer genomics <i>Niko Beerenwinkel</i>	96
New advances in Sequencing Technology <i>Ewan Birney</i>	96
The art and science that we can learn from assembly graphs <i>Jason Chin</i>	96
Non-algorithmic aspects of sequencing software <i>Pascal Costanza</i>	97
Challenges in designing a library of practical compact data structures <i>Simon Gog</i>	98
Gene-Centric Assembly <i>Daniel H. Huson</i>	98
Computational Pan-Genomics: Status, Promises and Challenges <i>Tobias Marschall</i>	98
Challenges in organizing a metagenomic benchmarking challenge <i>Alice Carolyn McHardy</i>	99
Upcoming challenges in phylogenomics <i>Siavash Mirarab</i>	99
Recent advances and future challenges in BWT <i>Laurent Mouchard</i>	100
Examples where theory fails in practice and practice needs some theory <i>Gene Myers</i>	100
Pangenome Variant Calling <i>Veli Mäkinen</i>	101
Challenges of ancient genomics and pan-genomics <i>Kay Nieselt</i>	101
Data structures to employ embeddings of strings under edit distances to vectors under Hamming distance <i>S. Cenk Sahinalp</i>	102
Ensembles of HMMs <i>Tandy Warnow</i>	102
Three problems in metagenomics <i>Shibu Yooseph</i>	103

Working groups

Single-cell cancer genomics: variant calling & phylogeny

Niko Beerenwinkel, Mohammed El-Kebir, Gunnar W. Klau, Tobias Marschall, and S. Cenk Sahinalp 104

Cancer genomics

Mohammed El-Kebir, Niko Beerenwinkel, Christina Boucher, Anne-Katrin Emde, Birte Kehr, Gunnar W. Klau, Pietro Lio', Siavash Mirarab, Luay Nakhleh, Esko Ukkonen, and Tandy Warnow 106

Software libraries for indexing

Simon Gog, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Hannes Hauswedell, André Kahles, Ben Langmead, Laurent Mouchard, Gene Myers, Enno Ohlebusch, Simon J. Puglisi, Gunnar Rätsch, Knut Reinert, Bernhard Renard, Enrico Siragusa, German Tischler, and David Weese 109

Assembly

Gene Myers, Jason Chin, Richard Durbin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Oliver Kohlbacher, Veli Mäkinen, Alice Carolyn McHardy, Laurent Mouchard, Kay Nieselt, Adam M. Phillippy, Tobias Rausch, Peter F. Stadler, Granger Sutton, German Tischler, and David Weese 112

Big data

Gene Myers, Ewan Birney, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Richard Durbin, Simon Gog, Hannes Hauswedell, Birte Kehr, Ben Langmead, Laurent Mouchard, Enno Ohlebusch, Adam M. Phillippy, Mihai Pop, Simon J. Puglisi, Tobias Rausch, Karin Remington, S. Cenk Sahinalp, Peter F. Stadler, and German Tischler 114

Structural Variant Detection

Gene Myers, Jason Chin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Veli Mäkinen, Tobias Marschall, Adam M. Phillippy, Mihai Pop, Karin Remington, S. Cenk Sahinalp, and Granger Sutton 116

Visualization Group

Gene Myers, Jason Chin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Veli Mäkinen, Tobias Marschall, Adam M. Phillippy, Karin Remington, S. Cenk Sahinalp, and Granger Sutton 119

Metagenomics

Mihai Pop, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Simon Gog, Hannes Hauswedell, Daniel H. Huson, André Kahles, Pietro Lio', Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Enno Ohlebusch, Simon J. Puglisi, Gunnar Rätsch, Karin Remington, Bernhard Renard, Enrico Siragusa, Tandy Warnow, and Shibu Yooseph 120

Haplotype Phasing

Knut Reinert, Niko Beerenwinkel, Jason Chin, Richard Durbin, Mohammed El-Kebir, Anne-Katrin Emde, Gunnar W. Klau, Veli Mäkinen, Tobias Marschall, Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Bernhard Renard, Enrico Siragusa, Peter F. Stadler, Granger Sutton, Tandy Warnow, and David Weese . . . 125


Pan-Genomics
Knut Reinert, Jason Chin, Fabio Cunial, Simon Gog, André Kahles, Birte Kehr, Oliver Kohlbacher, Ben Langmead, Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Enno Ohlebusch, Adam M. Phillippy, Simon J. Puglisi, Gunnar Rätsch, Karin Remington, Bernhard Renard, Peter F. Stadler, Granger Sutton, German Tischler, and Shibu Yooseph 126

Participants 130

3 Overview of Talks

3.1 Computational challenges in cancer genomics


Niko Beerenwinkel (ETH Zürich – Basel, CH)

License  Creative Commons BY 3.0 Unported license
© Niko Beerenwinkel

Cancer genomics has seen tremendous advancements with the arrival of cost-effective high-throughput sequencing. These technologies allow for analyzing cancer samples in unprecedented detail. At the same time, the resulting sequencing data poses a range of new computational challenges in analyzing and interpreting the data. These challenges include (1) read mapping and mutation calling in mixed tumor samples, including low-frequency variants; (2) detection of complex genomic alterations, which are common in cancer genomes; (3) inferring the clonal structure of mixed tumor samples from bulk sequencing data; (4) reconstructing the evolutionary history of a tumor, i.e., solving the tumor phylogeny problem; (5) reconstructing tumor phylogenies from single-cell sequencing data, and (6) predicting cancer evolution by learning models from independent observations across tumor samples from different patients. Approaches to all of these challenges exist, but most are inherently difficult or even mathematically ill-posed. Progress with these challenges is likely to have an impact on cancer diagnostics and treatment.

3.2 New advances in Sequencing Technology

Ewan Birney (European Bioinformatics Institute – Cambridge, GB)


License  Creative Commons BY 3.0 Unported license
© Ewan Birney

I will present an overview of the features of new sequencing technology, in particular PacBio and Oxford Nanopore. Both produce long reads with somewhat higher error rates than Illumina short read sequencing. The error rate though is manageable as has been shown in particular with PacBio data. Both systems work asynchronously with individual reads being produced. In the case of Oxford Nanopore, the sequencing process can be stopped early in a read and a new read resampled in real-time. This provides new avenues of algorithms which combine decision making in real time with sampling management.

Note: I am a paid consultant to Oxford Nanopore, and thus I am very explicit about this conflict of interest

3.3 The art and science that we can learn from assembly graphs

Jason Chin (Pacific Biosciences – Menlo Park, US)

License  Creative Commons BY 3.0 Unported license
© Jason Chin

In an overlap-layout-consensus assembler, the assembly graph constructed from read overlaps is the major data structure for generating contigs. Repeat-induced ambiguities within the graph are typically removed by analyzing local neighboring subgraph, defined as a subgraph

of a selected node or an edges and its nearest neighbors, properties. Contigs are constructed after removing those ambiguities. However, the heuristic rules to remove the ambiguities may also remove useful information that can be used for improving genome assembly and understand local genome structure.

By analyzing non-local graph structures (e.g. the subgraph within certain distance from a vertex), we can recover such missing information and reveal important biological information within the data. For example, heterozygous variants between haplotypes within a diploid genome usually create “bubbles” in the assembly graph. Identifying and analyzing such bubbles can lead to a full haplotype resolved assembly. Local unresolved repeats also created local tangled sub-graphs which might break contigs. In such case, if we can still identify unique source and sink of the subgraph, we can generate the linkage information to connect contigs into as “extend contigs”. While some ambiguities remain in the tangled region, the extend contigs will contain all sequence information of the repeat regions.

Here are some related challenges for utilizing the assembly graph to extract more biological information:

1. Utilize the assembly graph information to define the “quality value” indicating uncertainties or errors at a given point of the contigs.
2. Understand whether there are systematic patterns of local repeats.
3. Develop algorithms for combining different data types at assembly graph level for scaffolding and resolving ambiguities.

3.4 Non-algorithmic aspects of sequencing software

Pascal Costanza (Intel Corporation, BE)

License  Creative Commons BY 3.0 Unported license
© Pascal Costanza

When composing tools to create sequencing pipelines, the most widely used approach to pass intermediate results from one tool to the next is through intermediate files. This limits the scalability of such pipelines when trying to take advantage of multiple cores due to Amdahl’s Law, since the file transfer from one tool to the next is a sequential bottleneck. We have shown in previous work that this limitation can be overcome by grouping several steps in a pipeline into a single tool and keeping all data in memory. Upcoming new memory technologies will make it more and more feasible to keep large amounts of data in memory - however, there is currently no good solution for allowing several tools written in different programming languages to equally take advantage of such in-memory representations of sequencing data and allow them to collaborate without going through the bottleneck of file transfers. Memory-mapped binary file formats that can be accessed through shared memory may be an answer, but there are open challenges that need to be addressed to make this practical.

3.5 Challenges in designing a library of practical compact data structures

Simon Gog (KIT – Karlsruher Institut für Technologie, DE)

License © Creative Commons BY 3.0 Unported license
© Simon Gog

In this talk we discuss the challenges in designing and maintaining a data structure library which enables researchers in Bioinformatic to build tools which can handle large datasets. Three of the main challenges in the moment is to improve the construction process of index structures, to identify primitives which allow the composition of structures which can deal with highly-repetitive data, and to add support for dynamic operations like inserting and deleting sequences.

We exemplify the impact of an improvement of a basic data structure to applications by the use of the partitioned Elias-Fano (PEF) encoding in Compressed Suffix Arrays. PEF was developed in the Information Retrieval field but we think that it has also impact on Bioinformatic applications.

We provide a tutorial for participant interested in space-efficient data structures here: <https://github.com/simongog/sigir16-topkcomplete>

3.6 Gene-Centric Assembly

Daniel H. Huson (Universität Tübingen, DE)

License © Creative Commons BY 3.0 Unported license
© Daniel H. Huson

Assembly of microbiome sequencing datasets is generally a difficult problem in practice. Some questions require the genes to be assembled, rather than genomes. Gene-centric assembly aims at assembling all reads that are recruited to a specific gene family. A first simple approach is to use BLASTX or DIAMOND (or an HMM-based approach) to align reads to references representing a given gene family and then to pass the reads to a full-featured assembler such as IDBA. We described so-called protein-reference guided assembly that aims at using protein alignments to detect DNA overlaps between reads recruited to a given gene family. Such an approach is implemented in the MEGAN software and this was briefly demonstrated.

3.7 Computational Pan-Genomics: Status, Promises and Challenges

Tobias Marschall (Universität des Saarlandes, DE)

License © Creative Commons BY 3.0 Unported license
© Tobias Marschall

Main reference The Computational Pan-Genomics Consortium, “Computational Pan-Genomics: Status, Promises and Challenges”, Briefings in Bioinformatics, pp. 1–18, 2016.

URL <http://dx.doi.org/10.1093/bib/bbw089>

Many disciplines, from human genetics and oncology to plant breeding, microbiology and virology, commonly face the challenge of analyzing rapidly increasing numbers of genomes. In case of Homo sapiens, the number of sequenced genomes will approach hundreds of thousands

in the next few years. Simply scaling up established bioinformatics pipelines will not be sufficient for leveraging the full potential of such rich genomic datasets. Instead, novel, qualitatively different computational methods and paradigms are needed. We will witness the rapid extension of computational pan-genomics, a new sub-area of research in computational biology. In this paper, we generalize existing definitions and understand a pan-genome as any collection of genomic sequences to be analyzed jointly or to be used as a reference. We examine already available approaches to construct and use pan-genomes, discuss the potential benefits of future technologies and methodologies, and review open challenges from the vantage point of the above-mentioned biological disciplines. As a prominent example for a computational paradigm shift, we particularly highlight the transition from the representation of reference genomes as strings to representations as graphs. We outline how this and other challenges from different application domains translate into common computational problems, point out relevant bioinformatics techniques and identify open problems in computer science. With this review, we aim to increase awareness that a joint approach to computational pan-genomics can help address many of the problems currently faced in various domains. (Abstract taken from DOI: 10.1093/bib/bbw089, CC-BY 3.0)

3.8 Challenges in organizing a metagenomic benchmarking challenge

Alice Carolyn McHardy (Helmholtz Zentrum – Braunschweig, DE)

License © Creative Commons BY 3.0 Unported license
© Alice Carolyn McHardy

The computational analysis of metagenomic NGS data sets is a rapidly evolving field. The Initiative for the Critical Assessment of Metagenome Interpretation (CAMI) aims to evaluate methods in metagenomics independently, comprehensively and without bias. The first CAMI challenge has been run in 2015. We find that the most important challenges of organizing such a challenge are to (i) engage both the method developer and the applied metagenomics fields, (ii) to decide on the nature of the benchmarking data sets, such that they are both realistic and interesting, (iii) to decide on the specific challenges and (iv) applied evaluation metrics, such that they both are informative for real world applications and accepted by the developer community, as well as (v) to ensure reproducibility of the tool submissions, data sets and the performance evaluation.

3.9 Upcoming challenges in phylogenomics

Siavash Mirarab (University of California at San Diego, US)

License © Creative Commons BY 3.0 Unported license
© Siavash Mirarab

A major challenge in reconstructing evolutionary histories (i.e., phylogenies) is accounting for the potential discordance between histories of individual genes (i.e., gene trees) and the species as a whole (i.e., the species tree). Reconstructing phylogenies from genome-scale data has the promise to address this long-standing challenge in phylogenetics. However, several new challenges are presented when genome-wide data are used for phylogeny inference. At the highest level, the definition of a gene and a species becomes important and non-trivial. Scalable methods for species delimitation and for selecting recombination-free regions of the

genome are needed; moreover, we need to better understand impacts of recombination on phylogeny estimation, both at the gene and the species level. Simultaneous modeling of multiple causes of discordance between gene trees and the species tree is also challenging, both from theoretical and practical perspectives. When models that incorporate multiple causes of discordance are designed, inference under them often becomes an intractable computational problem. This has limited the best of existing methods that handle multiple causes of discordance to no more than tens of species. Finally, testing the accuracy of genome-scale phylogenies and interpreting the results generated by various methods requires care; improved methods for assessing support will be needed.

3.10 Recent advances and future challenges in BWT

Laurent Mouchard (University of Rouen, FR)

License  Creative Commons BY 3.0 Unported license
© Laurent Mouchard


Given a text T , the Burrows-Wheeler Transform of T is the last column of the conceptual matrix where rows are alphabetically ordered cyclic shifts of T . $BWT("BANANA\$")=ANNB\AA . This reversible transform, that does not compress text has a tendency of aggregating similar individual letters. It has been used as a preprocessing tool for compressors such as bzip2 for example. There exists a function, named LF (Last-First) that can be used for recovering the original text T when one has only access to $BWT(T)$. This transform and the corresponding data structure has been used in the context of Next-Generation Sequencing for preprocessing the reference sequences in order to speed up the detection of starting positions of myriads of short fragments (reads) in the reference sequences. Some technical aspects, such as time and space complexity are addressed. Several recent advances are presented:

- Dynamic and relative BWT
- Role of BWT in the context of FM-indices
- BWT of a set of highly similar sequences
- BWT construction using external memory
- Merging BWTs

A brief overview of future challenges is presented that paves the way for interactions/discussions during the Seminar.

3.11 Examples where theory fails in practice and practice needs some theory

Gene Myers (MPI – Dresden, DE)

License  Creative Commons BY 3.0 Unported license
© Gene Myers

We present a number of examples in the area of noisy, long read DNA reconstruction (assembly) where theory fails in practice:

- BWT's are theoretically superior, but k-mer sort and merge provides faster read mapping and overlap.
- Current multi-alignment heuristics are too slow and unable to separate polyploid genomes.

- We suggest that graphs are not good representations for pan genomics as they give unintuitive representation of next reversals, transpositions, and inversions.

And where practice needs some theory:

- A clear elucidation of the differences between deBruijn and string graphs is needed, along with an understanding of the limitation of each.
- CIGAR notation for alignments is space inefficient for noisy reads, and current formats are not designed for simplicity of adoption and machine reading.
- We suggest that assembly benchmarking would be significantly more informative if simulated data and theoretical sound metrics were used.
- HPC middle-ware is cumbersome and not tailored to bioinformatics.
- Good visualization and editing tools for assemblies still do not exist.

3.12 Pangenome Variant Calling

Veli Mäkinen (University of Helsinki, FI)

License © Creative Commons BY 3.0 Unported license
© Veli Mäkinen

Detection of genomic variants is commonly conducted by aligning a set of reads sequenced from an individual to the reference genome of the species and analyzing the resulting read pileup. Typically, this process finds a subset of variants already reported in databases and some novel mutations characteristic to the sequenced individual. Most of the effort in the literature has been put to the alignment problem on a single reference sequence, although our gathered knowledge on species such as human is pan-genomic: We know most of the common variations in addition to the reference sequence. There have been some efforts to exploit pan-genome indexing, where the most widely adopted approach is to build an index structure on a set of reference sequences containing observed variation combinations.

The enhancement in alignment accuracy when using pan-genome indexing has been demonstrated in experiments, but surprisingly the above multiple references pan-genome indexing approach has never been tested on its final goal, that is, in enhancing variation detection. This is the focus of this article: We study a generic approach to add variation detection support on top of the multiple references pan-genomic indexing approach. Namely, we study the read pileup on a multiple alignment of reference genomes, and propose a heaviest path algorithm to extract a new recombined reference sequence. This recombined reference sequence can then be indexed using any standard read alignment and variation detection workflow. We demonstrate that the approach actually enhances variation detection on realistic data sets.

This is joint work with Daniel Valenzuela, Niko Välimäki, and Esa Pitkänen.

3.13 Challenges of ancient genomics and pan-genomics

Kay Nieselt (Universität Tübingen, DE)


License © Creative Commons BY 3.0 Unported license
© Kay Nieselt

The advent of next-generation sequencing and recently developed enrichment techniques utilizing tailored baits to capture ancient DNA fragments have made it possible to reconstruct

and compare whole genomes of extinct organisms. Computational paleogenomics deals with the reconstruction and analysis of ancient genomes. Ancient DNA has a number of characteristics, such as short fragment lengths (mean length less than 150bp), and damaged bases, which need to be considered when reconstructing the genome, calling SNPs, comparing genomes or reconstructing phylogenies. In each of these four areas I propose several, partly related challenges. The first challenge addresses the question how to optimally reconstruct the genome from short read data. Typically mapping against a modern reference genome is performed, while de novo assembly is rarely possible. Could hybrid solutions be devised? SNP calling from assembled genomes poses a second problem, since often these assembled genomes suffer from low coverage. The third and fourth challenge address the question how to compare ancient and modern genomes. Since one needs a common coordinate system, the question is how to compute whole-genome alignments (WGA) from ancient as well as modern genomes. Or should one rather refrain from WGAs at all? Finally, in the context of phylogeny reconstruction a number of questions remain largely unsolved. One challenge in this area is to compute a lower bound of genome coverage for which a phylogenetic tree can still be reliably built. And finally relating also to the third challenge is the more general question whether phylogenetic trees consisting of modern as well as ancient genomes should be built from WGAs or with alignment-free methods?

3.14 Data structures to employ embeddings of strings under edit distances to vectors under Hamming distance

S. Cenk Sahinalp (Simon Fraser University – Burnaby, CA)

License  Creative Commons BY 3.0 Unported license
© S. Cenk Sahinalp

When comparing or aligning sequences, mismatches are much easier to handle than indels. Recent results in parsing (genomic) strings through random walks based on shared random bits result in a conceptually simple way to embed strings under edit distance to Hamming vectors, approximately preserving their pairwise distances. Such an embedding simplifies the problem of (pairwise or multiple) sequence alignment problem, even though the distortion (in the distance) they imply are higher than what could be tolerated in real world applications.

3.15 Ensembles of HMMs

Tandy Warnow (University of Illinois – Urbana-Champaign, US)

License  Creative Commons BY 3.0 Unported license
© Tandy Warnow

Profile HMMs are a major tool in bioinformatics analyses and are used for multiple purposes, including the representation of multiple sequence alignments, the detection of homology, protein classification, metagenomic taxon identification, protein structure and function prediction, etc. Yet a single profile HMM is not always suitable for representing a large collection of diverse sequences. In this talk, I will present some approaches to representing a collection of aligned sequences using an ensemble of profile HMMs instead of a single profile HMM. These approaches are able to improve phylogenetic placement, large-scale multiple sequence alignment, protein family classification, and metagenomic taxon identification. Not

only do these methods improve on accuracy (precision and recall) compared to methods based on single HMMs, they also provide improved accuracy compared to leading alternative methods. The relevant methods are SEPP (cf. [1]), TIPP (cf. [2]), UPP (cf. [3]), and HIPPI (cf. [4]). The talk is available at <http://tandy.cs.illinois.edu/warnow-dagstuhl.pdf>. The software base is available at <https://github.com/smirarab/sepp> (Siavash Mirarab github page).

References

- 1 Siavash Mirarab, Nam-phuong Nguyen and Tandy Warnow . *SATé-Enabled Phylogenetic Placement*. Proceedings PSB 2012, pp. 247–258, World Scientific, 2012
- 2 Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop and Tandy Warnow. *TIPP:Taxonomic Identification and Phylogenetic Profiling*. Bioinformatics, Oxford Journals, 2014
- 3 Nam-phuong Nguyen, Siavash Mirarab, Keerthana Kumar and Tandy Warnow . *Ultra-large alignments using Phylogeny-aware Profiles*. Genome Biology, 16:124, BioMed Central, 2015
- 4 Nam-phuong Nguyen, Michael Nute, Siavash Mirarab and Tandy Warnow. *HIPPI: Highly accurate protein family classification with ensembles of HMMs*. To appear, BMC Genomics

3.16 Three problems in metagenomics

Shibu Yooseph (University of Central Florida – Orlando, US)

License © Creative Commons BY 3.0 Unported license
© Shibu Yooseph

Metagenomics is a cultivation independent paradigm that has enabled detailed studies of microbial communities. Sequence data generated from a metagenome sample can be used to make inferences about the taxonomy, genome composition, and metabolic potential of the constituent microbes in the sampled community. However, the nature and volume of data generated by currently used sequencing technologies also pose computational challenges that require the development of efficient algorithms to effectively analyze these data. Here we discuss three computational problems in metagenomics to highlight these challenges and opportunities. First, to improve annotation of databases containing partial protein sequences, we describe approaches that have higher sensitivity than commonly used homology detection methods like BLAST. The higher sensitivity is obtained by combining database sequence searches together with the assembly of relevant overlapping database sequences to improve homology detection. Second, we describe the computational problem of identifying the host bacterial or archaeal sequences of a given set of viral metagenome sequences, and bottlenecks with current approaches. Third, we consider the problem of developing a unified framework for the estimation of both species abundance curves and metagenome coverage from a set of metagenomic reads.

4 Working groups

4.1 Single-cell cancer genomics: variant calling & phylogeny

Niko Beerenwinkel (ETH Zürich – Basel, CH), Mohammed El-Kebir (Brown University – Providence, US), Gunnar W. Klau (CWI – Amsterdam, NL), Tobias Marschall (Universität des Saarlandes, DE), and S. Cenk Sahinalp (Simon Fraser University – Burnaby, CA)

License © Creative Commons BY 3.0 Unported license
© Niko Beerenwinkel, Mohammed El-Kebir, Gunnar W. Klau, Tobias Marschall, and S. Cenk Sahinalp

4.1.1 Topics

- Variant calling in single-cell tumor sequencing
 - Single-nucleotide variants (SNV)
 - Copy-number variants (CNV)
 - Structural variants (SV)
 - Phylogeny inference given single-cell tumor sequencing data

4.1.2 Background

- Intra-tumor heterogeneity:
 - Tumor is heterogeneous composed of different cell populations with different somatic mutations.
 - With bulk sequencing the observations are a composite signal from different cell populations => requiring deconvolution
 - This is not the case with single-cell sequencing (SCS) where the observations are from a single cell
- There are specific errors with SCS due to the whole-genome amplification (WGA) step
 - High false negative rate in SNV calling due to allele drop-out in the WGA step
 - * Used to be ~40%; now improved to ~10%
 - Elevated false positive rate in SNV calling due to WGA step
 - Non-uniform read coverage
 - More GC-bias
- Single-cell sequencing is becoming more affordable.
 - Right now about 50 cells are sequenced
 - Most SCS studies are done using whole-exome sequencing (non-uniform read coverage is an even bigger issue in this case)

4.1.2.1 Questions

- Has reproducibility of single-cell sequencing been studied?
 - Nick Navin studied this in healthy cells

4.1.3 SNV calling

4.1.3.1 Issues

- Noisy data with high FP and FN rate (see Background).

4.1.3.2 Approaches

- Use SNV callers that were designed for bulk-sequencing (GATK, MuTect, ...)
- New SNV caller specific for SCS data: Monovar
 - Accounts for allele drop-out and elevated FP rate
 - Uses dynamic programming to compute posterior probabilities and to call SNVs for each cell with max posterior probability.
- Phylogeny inference under the infinite sites assumption to clean up noisy observations: SCITE and OncoNEM.

4.1.3.3 Opportunities

- Do SNV calling by considering all sequenced single-cells of a tumor simultaneously.
 - Monovar is considering cells one by one (with respect to the normal), i.e. assuming independence of cells
- Do SNV calling jointly with phylogeny inference
- Do SNV calling by integrating bulk-sequencing samples.

4.1.4 CNV calling

4.1.4.1 Issues

- Non-uniform read coverage
- Most SCS data is whole-exome only

4.1.4.2 Approaches

- Ginkgo

4.1.4.3 Opportunities

- Joint inference of all cells simultaneously
- In the context of a phylogeny?

4.1.5 Phylogeny inference

4.1.5.1 Motivation

Why do we care about the tree?

- To quantify heterogeneity
- To study the evolutionary process in cancer: is it a burst or is it gradual?
 - Neutral evolution model: (Big Bang): star phylogeny
 - Clonal expansion model: non-star phylogeny
 - These hypotheses can be tested.
- To study the trees of a cohort of patients where we have phenotype and treatment information.
 - Can we find patterns in the trees related to a phenotype?
- To study metastases and migration of tumor cells
 - Where do tumor cells that circulate in the blood come from?
 - Oliver raises the point that in melanoma the metastasis are different from the primary tumor.

4.1.5.2 Approaches

- SCITE
- OncoNEM

4.1.6 Ideas

- Combining bulk and single-cell sequencing
 - How many single cells do you need to sequence in order to detect all relevant clones?
 - * Should we sequence all billion cells of a tumor?
 - * This is a sampling question and it depends on the tumor being well-mixed, and whether there are selective advantages.
- Can we get time-series data?
 - Liver cancer is a candidate:
 - * It's not surgically removed and thus time-series samples can be obtained by a needle while the patient is under treatment
 - * Niko says this is painful and thus hard to get such samples, but Oliver may have access to such samples.
 - Leukemia
- What is a good generative model for the somatic mutational process in cancer?
 - This will allow us to validate variant calling and phylogeny inference methods.
 - Niko suggests that HMMs are enough
 - Tandy prefers phylo-HMMs [refs] or tree-based HMMs [refs]
- Philosophical discussion about Bayesian approaches
 - Niko: The following is a misconception: Bayesian computations are expensive, and likelihood computations are cheap.
 - * In some cases sampling from the posterior is hard to achieve
 - * It takes a long time for the MCMC chain to mix
 - Max Likelihood approaches: You can do anything to optimize the objective function.
 - Bayesian inference: Any sampling schemes that construct a proper Markov chain are fine. Some converge faster than others. Anything goes.
 - Bayesian: How to summarize your posterior?
 - How to communicate the uncertainty?

4.2 Cancer genomics

Mohammed El-Kebir (Brown University – Providence, US), Niko Beerenwinkel (ETH Zürich – Basel, CH), Christina Boucher (Colorado State University – Fort Collins, US), Anne-Katrin Emde, Birte Kehr, Gunnar W. Klau (CWI – Amsterdam, NL), Pietro Lio' (University of Cambridge, GB), Siavash Mirarab, Luay Nakhleh, Esko Ukkonen (University of Helsinki, FI), and Tandy Warnow (University of Illinois – Urbana-Champaign, US)

License © Creative Commons BY 3.0 Unported license

© Mohammed El-Kebir, Niko Beerenwinkel, Christina Boucher, Anne-Katrin Emde, Birte Kehr, Gunnar W. Klau, Pietro Lio', Siavash Mirarab, Luay Nakhleh, Esko Ukkonen, and Tandy Warnow

Cancer is a disease caused by somatic mutations that accrue in a population of cells during the lifetime of an individual [6]. This process can be described by a phylogenetic tree and results in different subpopulations of cells, or *clones*, each with different complements of somatic mutations. A clone is composed of all cells that share the same most recent common ancestor,

or equivalently all the leaves that occur in a subtree of the phylogeny. This definition of a clone is elusive: at one extreme all tumor cells form a clone, whereas at the other extreme each tumor cell is a clone. The desired resolution is not clear and depends on the specific application.

Here, we discuss recent trends in computational cancer genomics and identify topics of interest with open computational challenges.

4.2.1 Bulk vs. Single-cell Sequencing

Most cancer sequencing studies are performed using bulk-sequencing technology, where the observations are composite signals from a mixture of cells with different somatic mutations. In contrast, with single-cell sequencing (SCS) the observations are from individual tumor cells. However, there are specific errors that occur during the whole-genome amplification (WGA) step, including segmental drop out where not all copies of a genomic segment are amplified. The used sequencing approach has thus implications in variant calling and phylogeny inference, and requires tailored methods and error models as we will discuss in the following.

4.2.2 Variant Calling

Somatic variants differ in size and include single-nucleotide variants (SNVs) that affect individual genomic positions, copy-number variants (CNVs) that affect larger genomic regions and more complex structural variants (SVs) that do not necessarily change the copy number such as inversions. Calling somatic SNVs and CNVs in tumor bulk-sequencing samples with respect to a matched normal samples requires dealing with mixed samples, where variants do not necessarily occur in all cells. This topic has been studied extensively in the literature but it is not solved and remains a hard problem. Here, we focus on variant calling in SCS data where we have to account for SCS-specific errors.

In the context of SNV calling, allele drop-out leads to elevated false positive and false negative rates. For instance, not observing any reads with an SNV does not mean that the SNV is not present in a tumor cell as the segment containing the SNV could simply have failed to amplify in the WGA step. Recently, the method Monovar has been proposed for calling SNVs that accounts for errors specific to SCS [10].

Typically, read depth is used to infer copy-number values for genomic segments. However, due to allele drop-out, read depth is non-uniform in SCS data even for healthy cells that are heterozygous diploid. This effect is even more pronounced in whole-exome sequencing data where only 3% of the genome is sequenced. There are thus several opportunities in calling SNVs and CNVs for SCS data. For instance, considering all tumor cells simultaneously could improve consistency in the calls. Moreover, joint phylogeny inference and calling may further improve the accuracy.

4.2.3 Structural Variation Calling

Calling of structural variants (events > 50 bp) poses some additional challenges. Short read technologies are inherently limited in their capability to detect SVs, especially when events are complex and involve repetitive sequences. Moreover, wet-lab validation of such events can be difficult and even for germline SVs no comprehensive ground truth data sets are available. Therefore, biological formation mechanisms are far from being fully understood.

Most algorithms that act on short read data use a combination of read-pair, split-read, and read-depth signals and also several local assembly approaches have been developed. However, different tools can lead to very different SV call sets [9]. And even when tools

agree, combining the different types of signals into robust variant-allele fraction estimates is non-trivial [3].

Long-range technologies, including long reads, synthetic long reads and optical mapping, have the potential of resolving SVs better, especially when combined with short read data and when used to infer SVs simultaneously or iteratively with CNVs [1]. Another potential opportunity might be combining long read data with improved single-cell technology to monitor accumulation of variants over time, which could lead to a deeper understanding of SV formation.

While exploring these opportunities, better visualization tools (such as e.g. GenomeRibbon¹) and more consistent file formats for SV calls will be needed in order to make the calls more accessible and easier to handle.

4.2.4 Phylogeny Inference

Inferring tumor phylogenies allows one to study and test the applicability of different modes of evolution in human cancers such as the clonal expansion model [5] or the Big Bang model [8]. Moreover, studying phylogenetic trees of a cohort of patients where we have phenotypic and treatment information allows one to identify patterns that are related to specific phenotypes or treatment.

There are several challenges in phylogeny inference depending on the used sequencing strategy. SCITE [4] infers phylogenetic trees using SNVs under the infinite sites assumption and uses a likelihood model to account for elevated FP and FN rates in SCS data. Studying whether the infinite sites assumption is a reasonable assumption, especially in the context of copy-number variants, is an interesting open question. With bulk-sequencing data, tree inference methods must account for mixed samples and simultaneously solve a deconvolution and tree inference problem [2]. Since variant allele frequencies of SNVs are confounded by CNVs, it is thus essential to jointly consider SNVs and CNVs – and ideally all types of variants including SVs – when inferring phylogenetic trees given bulk-sequencing data.

4.2.5 Integrative Analysis

Integrated analysis of different molecular profiles obtained from the same tumor allows one to comprehensively study a tumor. For instance combined expression (RNA-seq) and mutation (DNA-seq) data may improve variant calling and could allow one to study the effect of somatic variants on expression, including alternative splicing and gene fusions. Moreover, combining different sequencing strategies could mitigate the challenges associated to the individual strategies and lead to an over-all better understanding of the somatic variants present in a tumor and their evolutionary history.

All of the above mentioned challenges require the development of novel methods. However, subsequent validation of such methods is difficult as no ground truth is available. Hence, it is also essential to formulate good generative models that comprehensively capture the somatic mutational process in cancer. Such models are currently missing, and we propose to consider and adapt existing models used in species evolution [7].

References

- 1 Xiang Chen, Pankaj Gupta, Jianmin Wang, Joy Nakitandwe, Kathryn Roberts, James D. Dalton, Matthew Parker, Samir Patel, Linda Holmfeldt, Debbie Payne, et al.

¹ <http://www.genomeribbon.com>

- CONSERTING: integrating copy-number analysis with structural-variation detection. *Nature Methods*, 12:527–530, 2015.
- 2 Mohammed El-Kebir, Gryte Satas, Layla Oesper and Benjamin J. Raphael *Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures*. *Cell Systems*, 3(1):43–53, July 2016.
 - 3 Xian Fan, Wanding Zhou, Zechen Chong, Luay Nakhleh and Ken Chen . Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics*, 15(1):1–12, 2014.
 - 4 Katharina Jahn, Jack Kuipers and Niko Beerenwinkel. *Tree inference for single-cell data*. *Genome biology*, 17(1):86, May 2016.
 - 5 Serena Nik-Zainal et al. *The life history of 21 breast cancers*. *Cell*, 149(5):994–1007, May 2012.
 - 6 P C Nowell . *The clonal evolution of tumor cell populations*. *Science*, 194(4260):23–8, Oct 1976.
 - 7 Adam Siepel and David Haussler . *Combining phylogenetic and hidden Markov models in biosequence analysis*. *Journal of Computational Biology*, 11(2-3):413–428, 2004.
 - 8 Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A. Graham, Matthew P. Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata and Christina Curtis. *A Big Bang model of human colorectal tumor growth*. *Nature Genetics*, 47(3):209–216, March 2015.
 - 9 Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, The 1000 genomes project consortium, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, 2015.
 - 10 Hamim Zafar, Yong Wang, Luay Nakhleh, Nicholas Navin and Ken Chen . *Monovar: single-nucleotide variant detection in single cells*. *Nature methods*, 13(6):505–507, May 2016.

4.3 Software libraries for indexing

Simon Gog (KIT – Karlsruher Institut für Technologie, DE), Pascal Costanza (Intel Corporation, BE), Anthony J. Cox (Illumina – United Kingdom, GB), Fabio Cunial (MPI – Dresden, DE), Hannes Hauswedell (FU Berlin, DE), André Kahles (ETH Zürich, CH), Ben Langmead (Johns Hopkins University – Baltimore, US), Laurent Mouchard (University of Rouen, FR), Gene Myers (MPI – Dresden, DE), Enno Ohlebusch (Universität Ulm, DE), Simon J. Puglisi (University of Helsinki, FI), Gunnar Rätsch (ETH Zürich, CH), Knut Reinert (FU Berlin, DE), Bernhard Renard (Robert Koch Institut – Berlin, DE), Enrico Siragusa (IBM TJ Watson Research Center – Yorktown Heights, US), German Tischler (MPI – Dresden, DE), and David Weese (SAP Innovation Center – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license

© Simon Gog, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Hannes Hauswedell, André Kahles, Ben Langmead, Laurent Mouchard, Gene Myers, Enno Ohlebusch, Simon J. Puglisi, Gunnar Rätsch, Knut Reinert, Bernhard Renard, Enrico Siragusa, German Tischler, and David Weese

4.3.1 Topics

- Recent work on bidirectional BWT
- Future plans to build Seqan on top of sdsl-lite
- Understanding why people use/don't use Seqan or other libraries?

4.3.2 Areas of interest – why are we here?

- Andre: interested in what is out there
- Enno: pan genome, compressed de Bruijn graphs, understanding what else people want
- Laurent: dynamic data structures, pangenomes
- Ben: latest on Seqan, sds, pangenomes
- Tony: indexing variant sets
- Enrico: SDSL, pangenomes, good implementations of assembly graphs
- David: k-mer index
- German: indexing large data structures, semi-external methods
- Pascal: efficiency of implementation
- Knut: practicalities of integrating libs

4.3.3 Knut – recent work on 2-directional BWT

- Two FM indexes – fwd and reverse, back in 1 = fwd in other
- Need cardinality of intervals
- Originally by Lam, improved by Ohlebusch, Gog + others Prefix sums would speed up but need to store them Present work shows – can do this with only extra bit per BWT character
- Simon believes can get rid of bit too
- bitwise ops Implemented in Seqan 7.44s → 4.79s over wavelet tree on DNA Space e.g. 88Mb → 131Mb
- See preprint for more detail: <https://arxiv.org/abs/1608.02413>
- German – implement fwd and reverse complement in 1 index, searches both dirs in 1 search, as done by Heng Li
- Ben – may still need rev index to do approx matching/branching? German believes not the case
- Discussion around how Knut's data structure would go into sds-lite
- Simon – self-contained data structures – support structures – augment self-contained ds (and has pointer to it), eg add rank support
- further discussion

4.3.4 reuse vs rewrite

- Solution based on minimal perfect hash functions? (from Veli Makinen) Simon – MPHFs popular in information retrieval but not in bioinformatics
- Discussion on space usage

4.3.5 Latest and greatest

- German (libmaus2 on github)
 - Huffman and RLE
 - own alg for indexing DNA, scales up to NCBI ref db, 1.5Tb inc fwd and r/c on github
 - GPL, mandated by Sanger (not ideal for integ with Seqan)
 - alg published but not exptl results
- Ben
 - Bowtie uses own implementation of FM index
 - Index building – Burkhardt + Karkkainen, parallelized
 - Bowtie2 uses Lam's bidirectional index

- Simon on SDSL
 - support for small and big alphabets (1 million) – latter needed for IR apps
 - inspiration from Pizza/Chili – generic imps of common components
 - bit vectors, rank/select
 - 8 flavours of wavelet tree! (choice depends on alphabet size among other factors)
 - 2 page cheat sheet describes everything
 - parameterize wavelet tree by bit vector – many combinations
 - configured at compile time in the manner of C++ STL
 - not only performance advantages, but also Pizza/Chili was hard to configure at index construction time via the API
 - Polymorphic construct() function builds anything
- Ben – avoid file copy by memory mapping?
- Simon – code to do this on branch right now
- relative not absolute pointers important
- David – is there abstract interface to string so that eg string non-contiguous in memory could be used?
- Large page sizes Configure OS for large page sizes as recommended by SNAP developers
- Kurt on SeqAN
 - Seqan BAM → SAM is 2x faster than htlib
 - Compile-time parameterization by alphabet type and index type (eg q-gram index, FM index)
 - Generic iterator interface for tree traversal
 - Compile-time generic programming module for dynamic programming (192 flavours!)
 - Multiple genomes stored using journaled string tree
 - * Q: can you index this? A: Yes/no!
 - * 15% overhead (of JST) for 1 string but 50-fold speedup for 130
 - * easy to add or delete a sequence
 - Working with Intel to add vectorization to core lib
 - Multithreading
- for SeqAn 3.0
 - Separate apps from core in build system
 - C++14, C++17 features where poss
 - multithread/SIMD of core components
 - external libraries eg sds, maybe graph libs
- Q: CRAM support? A: a lot of overhead to fully implement the CRAM spec Q: use htlib for CRAM? A: probably a lot of overhead in converting internal structs
- pragma simd to force vectorization, was in Intel compiler only but is now in OpenMP
- Recommendations on Cilk vs openMP vs TBB [can someone else summarize please]

Optimizing vectorized code:

 - vector reports compiler switch to see what is vectorized...
 - and these reports can be embedded in source code ...
 - Vtune is GUI for this
 - or use gdb or Intel's equiv to look at assembly language directly

4.4 Assembly

Gene Myers (MPI – Dresden, DE), Jason Chin (Pacific Biosciences – Menlo Park, US), Richard Durbin (Wellcome Trust Sanger Institute – Cambridge, GB), Mohammed El-Kebir (Brown University – Providence, US), Anne-Katrin Emde (New York Genome Center, US), Birte Kehr (deCode Genetics – Reykjavik, IS), Oliver Kohlbacher (Universität Tübingen, DE), Veli Mäkinen (University of Helsinki, FI), Alice Carolyn McHardy (Helmholtz Zentrum – Braunschweig, DE), Laurent Mouchard (University of Rouen, FR), Kay Nieselt (Universität Tübingen, DE), Adam M. Phillippy (National Institutes of Health – Rockville, US), Tobias Rausch (EMBL – Heidelberg, DE), Peter F. Stadler (Universität Leipzig, DE), Granger Sutton (The J. Craig Venter Institute – Rockville, US), German Tischler (MPI – Dresden, DE), and David Weese (SAP Innovation Center – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license

© Gene Myers, Jason Chin, Richard Durbin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Oliver Kohlbacher, Veli Mäkinen, Alice Carolyn McHardy, Laurent Mouchard, Kay Nieselt, Adam M. Phillippy, Tobias Rausch, Peter F. Stadler, Granger Sutton, German Tischler, and David Weese

4.4.1 Topics

Proposed discussion topics

- Assembly data format
 - Beyond linear representation
 - Capture ambiguity and quality
- Emerging technologies
 - Optimal/economic integration
 - Genome finishing
- Pre-assembly QC
 - Estimating ploidy, genome size, repeat content from raw reads
 - Error correction
- Population assembly, cancer assembly, metagenome assembly (other group)
- Local assembly for variant detection
- Assembly and graph visualization
- High-performance computing
- Provocation: Why isn't assembly solved? What's missing to solve it?

4.4.2 Assembly data format

A unified data format is needed that captures the full information (and ambiguity) of an assembly

Chin, Durbin, and Myers proposed an extension of the GFA format

- Vertices are segments, edges are overlaps
- Describes consensus and multi-alignments
- Consistent with SAM notations
- Segments can be with or without pieces (specify coordinate + alignment info (cigar or trace))
- Edge types: dovetail, branch, contain
- Expressive enough to describe the full assembly (graph + segments + pieces + alignments)
- History tracking through SAM header

Major proposed changes to the current GFA spec

- 'Pieces' as subcomponents of segments
- 'Branches' as any local alignment between segments
- Optional alignment formats (cigar or trace)
- Object-size prolog (debated)

Questions/Remarks

- Best way to encode haplotypes? With local alignments?
- Enough to have just one link type?
- What about scaffolds with ambiguous gap sizes?
- Provide validator/convertor for new format
- Develop binary version of format
- JSON or SAM style?
- Is there a way to represent collections?
 - All segments of a given chromosome (e.g. from Hi-C clustering)
 - All segments of a given organism (e.g. from metagenomic binning)
- Converters to common/alternate graph formats needed
- Are 'general' edges too flexible?
 - Can now represent all local alignments between segments
 - What does this graph structure represent? How it is visualized?
- NCBI is interested in an assembly submission format. What are their needs/requirements?

Proposed spec GFA 2.0 is here: <https://github.com/thegenemyers/GFA-spec>

4.4.3 Emerging technologies

Technologies for building great assemblies: what's new?

4.4.3.1 Technology list

- Long reads (PacBio, Nanopore)
- Short reads (Illumina)
- Synthetic long reads (Illumina TSLR)
- Linked reads (10x Genomics)
- In vitro Hi-C (Dovetail)
- In vivo Hi-C (PhaseGenomics)
- Optical Maps (BioNano)

4.4.3.2 Considerations

- Different technologies offer different resolution and accuracy
- Economics of best reconstruction (What kind of assembly do you need?)
- Contig vs. scaffold size (PacBio vs. 10x)
- New scaffolding opportunities with chromatin interaction frequency (Hi-C)
- New phasing options (10x, PacBio, Hi-C)
- Complementarity. Where do technologies break? (e.g. PacBio vs. BioNano)
- Iterative improvement and validation using multiple techs
 - E.g. PacBio → BioNano → Hi-C gives chromosome-scale assemblies

4.4.4 Pre-assembly QC

Is it my genome, my data, or my assembler that is causing the problem?

4.4.4.1 Suggestions

- Illumina
 - Compute k-mer frequency to estimate haploid and diploid coverage, repetitiveness, and genome size
 - Might be difficult for Hi-C due to non-uniform coverage
- PacBio
 - Count overlaps, rather than k-mers, to estimate coverage, repeats, and genome size
 - Reads can be too noisy for k-mer based approach

4.5 Big data

Gene Myers (MPI – Dresden, DE), Ewan Birney (European Bioinformatics Institute – Cambridge, GB), Pascal Costanza (Intel Corporation, BE), Anthony J. Cox (Illumina – United Kingdom, GB), Fabio Cunial (MPI – Dresden, DE), Richard Durbin (Wellcome Trust Sanger Institute – Cambridge, GB), Simon Gog (KIT – Karlsruher Institut für Technologie, DE), Hannes Hauswedell (FU Berlin, DE), Birte Kehr (deCode Genetics – Reykjavik, IS), Ben Langmead (Johns Hopkins University – Baltimore, US), Laurent Mouchard (University of Rouen, FR), Enno Ohlebusch (Universität Ulm, DE), Adam M. Phillippy (National Institutes of Health – Rockville, US), Mihai Pop (University of Maryland – College Park, US), Simon J. Puglisi (University of Helsinki, FI), Tobias Rausch (EMBL – Heidelberg, DE), Karin Remington (Computationality, US), S. Cenk Sahinalp (Simon Fraser University – Burnaby, CA), Peter F. Stadler (Universität Leipzig, DE), and German Tischler (MPI – Dresden, DE)

License © Creative Commons BY 3.0 Unported license

© Gene Myers, Ewan Birney, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Richard Durbin, Simon Gog, Hannes Hauswedell, Birte Kehr, Ben Langmead, Laurent Mouchard, Enno Ohlebusch, Adam M. Phillippy, Mihai Pop, Simon J. Puglisi, Tobias Rausch, Karin Remington, S. Cenk Sahinalp, Peter F. Stadler, and German Tischler

Public archives of DNA sequencing data are filled with valuable datasets contributed by projects large and small across the world. They are also growing extremely rapidly; the Sequence Read Archive, for example, has a doubling time of about 18 months. In the days before next-generation sequencing dominated the field, public databases were easy for everyday scientists to query. Today, these databases contain petabytes of data. While simply storing this data has recently become practical and sustainable – thanks in part to improved compression – the task of querying these databases, or even a large fraction thereof, is now very challenging. It's not possible for a typical biological researcher to rapidly query the large archives like the Sequence Read Archive or the European Nucleotide Archive.

We feel that an important focus of Computational Genomics research should be on tackling the problem of indexing very large amounts of sequencing data. Advances in this field would have two major benefits: it would make it easier for typical scientists to query these archives, and it would create an important incentive for producers of “private” sequencing data (e.g. clinical samples) to eventually release them to the public in some form. We also

note that the Global Alliance for Genomics and Health (GA4GH) have proposed mechanisms for allowing limited querying of private sequencing data spread across many loci.

To enable fast queries over archives, the pivotal need is for data structures capable of answering queries that take query sequences and return information about whether and where that queries occur in the raw data. This kind of query is in the spirit of local alignment; while other queries could certainly be useful, we focus on this kind here because it can serve as a building block for many others. We suggest that such a data structure should exist separately from the raw sequencing data; in other words, the raw data would still be stored and made available in an un-indexed form, which, while it does not allow fast queries, does allow a wide variety of methods to be applied. The core data structures we suggest for further investigation are

1. those based on the Burrows-Wheeler Transform (BWT) or FM Index,
2. those based on the de Bruijn graphs,
3. those based on multi-vantage-point trees, and
4. those based on sketching schemes or other schemes that reduce the key space by replacing sequences with representatives that are “nearby” in, say, edit distance space.

These data structures are primarily responsible for finding whether and where sequences occur, but they must be augmented to make it possible to determine which particular archived samples the sequences occur in. This is related to the “document listing problem,” and also related to the “colored” de Bruijn graph.

We briefly attempt to estimate the size required by a colored de Bruijn graph data structure built over a very large archive of sequencing data. We assume that the k -mer length is long enough that the number of distinct keys is governed by the amount of data rather than by the limited number of k -mers. We assume the number of distinct k -mers occurring in the archived data is 10^{12} , and that the number of bits required to associate metadata (i.e. the “color” bits) with each k -mer is about 10^8 per key. This leads to an estimate of about 10^{20} total bits, with additional space needed to store the keys themselves. However, there are many opportunities for compression, since

1. overall, the total collection of bit vectors is sparse; mostly 0s, given that most k -mers are absent from most datasets,
2. if the positions of the bit vectors correspond to samples then there is a dependence structure among the columns; since some samples are biologically similar, we expect them to be similar in k -mer composition, and
3. the bit vectors themselves are dependent since two k -mers that overlap by $k-1$ positions are likely to occur in similar patterns of database samples.

The assumption that only 10^{12} keys are needed needs further discussion. The number is almost certainly much larger in practice when real sequencing data is used. This is because of sequencing errors, which give rise to a very large number of k -mers that are made unique (or nearly so) by the random sequencing errors they overlap. We suggest that some degree of “smoothing” or error correction is needed to reduce the size of the key space prior to building the data structure.

4.6 Structural Variant Detection

Gene Myers (MPI – Dresden, DE), Jason Chin (Pacific Biosciences – Menlo Park, US), Mohammed El-Kebir (Brown University – Providence, US), Anne-Katrin Emde (New York Genome Center, US), Birte Kehr (deCode Genetics – Reykjavik, IS), Veli Mäkinen (University of Helsinki, FI), Tobias Marschall (Universität des Saarlandes, DE), Adam M. Phillippy (National Institutes of Health – Rockville, US), Mihai Pop (University of Maryland – College Park, US), Karin Remington (Computationality, US), S. Cenk Sahinalp (Simon Fraser University – Burnaby, CA), and Granger Sutton (The J. Craig Venter Institute – Rockville, US)

License © Creative Commons BY 3.0 Unported license

© Gene Myers, Jason Chin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Veli Mäkinen, Tobias Marschall, Adam M. Phillippy, Mihai Pop, Karin Remington, S. Cenk Sahinalp, and Granger Sutton

4.6.1 Structural Variant Calling

4.6.1.1 Basics

- Definition of SV: variant >50bp
- Types of sequencing-based signals/approaches:
 - Split reads (SR)
 - Read pairs (RP)
 - Read depths (RD)
 - Assemblies
- Challenges for SV calling
 - need for improved SV detection methods
 - need for improved annotation/resources for SVs
 - need for improved file formats and visualization tools
 - lack of biological understanding

4.6.1.2 SV detection methods/algorithms

- filtering of FPs, especially de-novo SVs; LD can help in population data
- current methods have limitations:
 - mostly limited to Illumina PE data, need to integrate technologies
 - most established methods focus on accessible genome (unique reads) but SVs are often in repeats
 - need for better quality scores (both for mapped reads as input into SV calling, and for called SVs)
 - problem of false positives, low validation rates esp. for de-novo SVs
- SV validation difficult
 - intersection of tools as proxy for precision, but shared artifacts as well as true unique calls
 - need for benchmarking data, SV gold standard
 - wetlab validation not straight-forward (cloning vector approaches? longer read techs)
 - SVs that agree with protein (mass spec)
 - SVs are often complex: mini events around SV breakpoints, homologies
- better merging/overlapping of SV sets, removing redundancy
 - resolution differs by method (assembly/split-read > read pair > read depth)
 - merging/comparing is a difficult problem (reciprocal overlap not sufficient)

- germline SV filtering:
 - family structure helps: jointly assess parents with child
 - population structure/haplotype information: LD

4.6.1.3 Need for improved annotation/resources for SVs

- need for useful databases of SVs
 - need for a dbSNP for SVs
 - DGV problematic
 - 1000G is not a great resource yet
- need comprehensively characterized genomes, SV goldstandard
 - currently being analyzed: 3 trios with 10x and strand-seq eventually
- need for improved annotation/resources for SVs
 - prediction of functional impact of SVs

4.6.1.4 Need for improved file formats and visualization tools

- SV visualization tools
 - IGV (limitations): Jason Chin working with IGV folks to add signs for large insertions
 - genomerebbon.com
 - Circos (but static, for complex events)
 - SVviz
- formats: VCF, bedpe, separate format for genotyping

4.6.1.5 Lack of biological understanding

- SV type classification not straight-forward
- mechanisms of SV formation not well-understood
- can knowledge about biological mechanism aid methods for detection?

4.6.1.6 Papers of interest

- Paper from 1000G SV group (mechanisms of SV formation) [1]
- Resolving the complexity of the human genome using single-molecule sequencing [2]
- GoNL de-novo SV [3]
- Veli: paper on merging SVs (tandem repeat regions, deletions, only pairwise)
- BreakDown for SV VAF estimation in cancer [4]

4.6.2 Somatic SV calling

4.6.2.1 What is different in cancer, what is different in single-cells?

Differences of the problem:

- Heterogeneity needs to be taken into account
 - Lower support for variants when analyzing mixtures of subclones (purity)
 - Intersecting variants from different subclones
- Non-uniform coverage in single-cell sequencing
- Comparative approaches: tumor vs normal

Methodological difference:

- Often the same methods as for germline SV detection

- Different post-processing
- Instead of population-wide calling, tumor/normal joint calling

4.6.2.2 Purity estimation / allelic frequency / proportionality problem

- Depends on copy number (purity and copy number can be traded for each other)
- Approach: Joint probability distribution from tumor and normal e.g. 0 in normal, >0 in tumor (HitSeq 2016)

4.6.2.3 Integration of copy-number estimation (RD) and adjacencies (SR + RP) in tools

- JABBA (unpublished): both predicted at the same time
- CONSERING: both predicted iteratively

4.6.2.4 How can you validate calls? How can we get ground truth?

- Idea (Mohammed): Construct tumor phylogeny from SNVs and from SVs separately and compare the result.
- Simulation impossible as long as we don't understand what is going on.
- Single-cell data solves clonality problem, long reads resolve complex events

4.6.2.5 Can we disentangle complex events, can we define atomic event and can we resolve an order in which events have occurred?

- Biologically, events are often more complex than atomic operation (not just simple deletions, insertions, inversions, ...). Fuzzy definition of breakpoints doesn't help.
- Ideal definition: Assuming that we have all cancer cells sequenced, an event is a change between two adjacent cells.
- Question: Is there a clear pattern of these events so that we can define a set of atomic operations?

References

- 1 Alexej Abyzov, Shantao Li, Daniel Rhee Kim, Marghoob Mohiyuddin, Adrian M. Stütz, Nicholas F. Parrish, Ximeng Jasmine Mu, Wyatt Clark, Ken Chen, Matthew Hurles, Jan O. Korbel, Hugo Y. K. Lam, Charles Lee and Mark B. Gerstein. *Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms*. Nature Communications 6, 7256, 2015
- 2 Mark J. P. Chaisson, John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M. Landolin, John A. Stamatoyannopoulos, Michael W. Hunkapiller, Jonas Korlach and Evan E. Eichler. *Resolving the complexity of the human genome using single-molecule sequencing*. Nature, 517, pp. 608–611, 2015
- 3 Wigard P. Kloosterman, Laurent C. Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y. Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H. Moed, Vyacheslav Koval, Ivo Renkens, Markus J. van Roosmalen, Pascal Arp, Lennart C. Karssen, Bradley P. Coe, Robert E. Handsaker, Eka D. Suchiman, Edwin Cuppen, Djie Tjwan Thung, Mitch McVey, Michael C. Wendl, Genome of the Netherlands Consortium, André Uitterlinden, Cornelia M. van Duijn, Morris A. Swertz, Cisca Wijmenga, GertJan B. van Ommen, P. Eline Slagboom, Dorret I. Boomsma, Alexander Schönhuth, Evan E. Eichler, Paul I.W. de Bakker, Kai Ye and Victor Guryev. *Characteristics of de novo structural changes in the human genome*. Genome Research, 25, pp. 792–801, 2015.

- 4 Xian Fan, Wanding Zhou, Zechen Chong, Luay Nakhleh and Ken Chen. *Characteristics of de novo structural changes in the human genome*. BMC Bioinformatics, 15:299, BioMed Central, 2014

4.7 Visualization Group

Gene Myers (MPI – Dresden, DE), Jason Chin (Pacific Biosciences – Menlo Park, US), Mohammed El-Kebir (Brown University – Providence, US), Anne-Katrin Emde (New York Genome Center, US), Birte Kehr (deCode Genetics – Reykjavik, IS), Veli Mäkinen (University of Helsinki, FI), Tobias Marschall (Universität des Saarlandes, DE), Adam M. Phillippy (National Institutes of Health – Rockville, US), Karin Remington (Computationality, US), S. Cenk Sahinalp (Simon Fraser University – Burnaby, CA), and Granger Sutton (The J. Craig Venter Institute – Rockville, US)

License © Creative Commons BY 3.0 Unported license

© Gene Myers, Jason Chin, Mohammed El-Kebir, Anne-Katrin Emde, Birte Kehr, Veli Mäkinen, Tobias Marschall, Adam M. Phillippy, Karin Remington, S. Cenk Sahinalp, and Granger Sutton

God created visualization and he saw it was good.

We focused on assembly/pan-genome visualization. The first question is defining the purpose – what do we want to visualize and what are the question we want to answer with them.

One observation is that in pan-genomes there are chunks of conserved regions interspersed by highly variable regions. We don't have a good way of visualizing the highly variable region, or interpreting its content in the context of its neighborhood. Some relevant questions may be: are there genes disrupted by this region?; are there specific variants? etc.

These problems are much easier to conceptualize in the context of pan-genomes rather than metagenomic assembly graphs. In assembly graphs complexity due to repeats and errors cannot be easily distinguished from actual biological signals (translocations, strain variants).

Finding the tangles in the graph may be attempted by using the SPQR tree datastructure that hierarchically decomposes a bi-connected graph into tri-connected components (the tangles/variants). In the pan-genome setting this may be achieved with simpler algorithms.

We discussed the Pan-Tetris paradigm (cf. [1]) that is gene-centric and also models the ordering of genes. The visual representation makes it easy to 'combine' tracks representing orthologous genes that may have been mis-aligned in the multiple alignment or have mis-annotated. An important functionality not present is ability to use this information to edit and update the underlying genome alignment or annotation.

In terms of updates we discussed the importance of consistency checks and version tracking to prevent and enable recovery from errors.

We also discussed the need for hierarchical visualizations (SPQR trees for example can provide such a mechanism for assembly graphs) going from the large structure down to the base level.

References

- 1 André Hennig, Jörg Bernhardt and Kay Nieselt *Pan-Tetris: an interactive visualisation for Pan-genomes*. BMC-Bioinformatics, 16(Suppl 11), BioMed Central, 2015

4.8 Metagenomics

Mihai Pop (University of Maryland – College Park, US), Pascal Costanza (Intel Corporation, BE), Anthony J. Cox (Illumina – United Kingdom, GB), Fabio Cunial (MPI – Dresden, DE), Simon Gog (KIT – Karlsruher Institut für Technologie, DE), Hannes Hauswedell (FU Berlin, DE), Daniel H. Huson (Universität Tübingen, DE), André Kahles (ETH Zürich, CH), Pietro Lio' (University of Cambridge, GB), Alice Carolyn McHardy (Helmholtz Zentrum – Braunschweig, DE), Siavash Mirarab (University of California at San Diego, US), Kay Nieselt (Universität Tübingen, DE), Enno Ohlebusch (Universität Ulm, DE), Simon J. Puglisi (University of Helsinki, FI), Gunnar Rätsch (ETH Zürich, CH), Karin Remington (Computationality, US), Bernhard Renard (Robert Koch Institut – Berlin, DE), Enrico Siragusa (IBM TJ Watson Research Center – Yorktown Heights, US), Tandy Warnow (University of Illinois – Urbana-Champaign, US), and Shibu Yooseph (University of Central Florida – Orlando, US)

License © Creative Commons BY 3.0 Unported license

© Mihai Pop, Pascal Costanza, Anthony J. Cox, Fabio Cunial, Simon Gog, Hannes Hauswedell, Daniel H. Huson, André Kahles, Pietro Lio', Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Enno Ohlebusch, Simon J. Puglisi, Gunnar Rätsch, Karin Remington, Bernhard Renard, Enrico Siragusa, Tandy Warnow, and Shibu Yooseph

4.8.1 Topics Proposed for Discussion

Originally we proposed the following topics for additional discussion.

- Taxonomic analysis
- Analyses of viruses, fungi, Eukaryotes ...
- Functional analysis
- (Metagenome) Assembly
- Strain reconstruction
- How do you want to benchmark
- Integration of -omics data

In the end, the discussion focused on the many challenges posed by the first set of topics and we did not discuss issues surrounding the integration of multiple types of omics data. Also, issues related to databases were found to be central to multiple of the topics. A summary of the discussions is provided below.

4.8.2 Taxonomic issues

While bacterial taxonomy was historically based on morphology, taxonomic schemes have largely moved to including the use of molecular sequence data to organize bacteria (and archaea). However, given that events like lateral gene transfer are common among bacterial groups, it is often also problematic to represent bacterial evolution and relationships using a tree structure. In addition, higher resolution of bacterial groups are complicated by the current lack of formal definitions (i.e. with mathematical utility) for concepts like “bacterial strains”. Due to these difficulties, the NCBI database has stopped tracking the concept of “strain” altogether. We discussed whether this is a problem – after all the strain sequences would have the nucleotide sequence ID as an identifier making it unnecessary to “pollute” the taxonomy database as well. We decided that in some cases having a definition of strain may be useful. One option is to create a domain-specific strain labeling scheme, e.g., *Staphylococcus aureus* MEC+ for a *S. aureus* strain containing the MEC cassette. For consistency, this annotation should be defined computationally and may only make sense within a specific domain. As such a same strain may acquire different “names” in the context of antibiotic

resistance, as opposed to its annotation in the context of bioremediation, for example. A translation between the many possible naming schemes could also be defined computationally, or better yet, the sequence of the strain could represent the ultimate identifier linking the different databases.

In the discussion of taxonomy we also noted that official naming rules are unnecessarily strict – an organism must be isolated, assayed for a number of biological attributes, given a name that follows valid Latin grammar, and submitted and accepted to the International Journal of Systemic and Evolutionary Microbiology (a fun read on the topic is at <http://biorxiv.org/content/biorxiv/early/2016/01/19/037325.full.pdf>).

While computational representations for a taxonomy (e.g., the `names.dmp`, `nodes.dmp` representation of the NCBI taxonomy) are preferable for computational analyses, we discussed the need for real names for taxonomy labels as these names are meaningful for biologists. A number of taxonomies (RDP, GreenGenes, SILVA) impose arbitrary restriction on the number of levels of a taxonomy simply due to computational convenience – it is easy to parse names into levels by splitting strings (the textual representation of a path in a tree) only if the number of levels is consistent. We argued for a more expressive computational representation, coupled with an optional textual representation of the paths which would only be used for display to end-users.

4.8.3 Viral metagenomics

Viruses also present additional set of taxonomy related challenges. Viruses constitute a large group of diverse entities that have thus far been largely uncharacterized and have defied a coherent unified taxonomic classification scheme. In addition, there are no good “markers” that define a virus (unlike bacteria where a number of housekeeping genes are universally found). Even composition-based or taxonomic methods may be insufficient as viruses frequently copy their genes from their host.

Viral diversity is estimated to be very high and viruses are important players in many environments (e.g. unknown viruses might be causing diseases in human). Currently there are many practical hurdles to their study and viral diversity is thought to be vastly under-represented in reference databases. It is not uncommon to identify long DNA segments in metagenomic mixtures that do not have good homology to any organisms, even distantly related. An example is the recent discovery of a new phage in HMP data (the crAss Phage [1]), phage that is found in about 50% of the human population yet its proteins only bear very distant similarities with other phage-like proteins. This problem is compounded by the fact that in many cases the abundance of a virus within a sample is very low (e.g., 1 out of every 20 million reads), making it impossible to estimate the parameters of statistical models for organism identification, or requiring substantial computational resources in order to process the large volumes of data necessary to ensure sufficient coverage. Different participants reported that in their experience, the targeted enrichment of viruses do not work well. RNA viruses pose further difficulties, for instance, in getting starting material for sequencing. There are iterative approaches to close in on the virus sequence using a combination of assembly and reference search per step. However, finding a virus in a sample does not always necessarily mean it is relevant for a given diseases phenotype.

Another opportunity for interesting research is creating host-virus linkages (phage-bacterium, defining host range for eukaryotic viruses, etc.).

4.8.4 Database issues

4.8.4.1 Correctness of reference databases

An additional challenge is posed by the databases themselves. In many cases the sequences deposited in public databases are mis-annotated, contain contaminants (e.g., mycoplasma in the human genome [2]), or represent “enriched metagenomes” rather than isolates (e.g., the sequences from this paper [3] which are deposited as isolates).

This state of affairs allows the opportunity for interesting research projects, such as the automatic identification of errors or contamination in public datasets. These could be phrased as outlier detection at different levels of resolution: – inconsistent taxonomic placement of an entire assembly – inconsistent taxonomic placement of individual contigs or genomic regions – outliers in terms of nucleotide composition, – discordant sequences not found in other genomes with the same label. A challenge in doing such analyses is avoiding the mis-classification of “true outliers”, such as ribosomal RNA operons or mobile elements that are genuine biological phenomena and not simply mis-annotations.

4.8.4.2 Databases of metagenomic datasets

There was also a discussion on publicly available resources for comprehensive collections of metagenome datasets. Resources include EBI, NCBI’s SRA, HMP DACC, CAMI, MetaHIT, iMicrobe, MG-RAST, etc. When analyzing new datasets, it is also important to be able to leverage existing datasets and inferences as much as possible. Frameworks and representation schemes that do not require expensive recomputations should be developed. On this front, challenges and opportunities include [a] the efficient representation of a comprehensive metagenome database (ideas from data representation for pan-genomics data could be applicable here), [b] index construction from large reference datasets, with an updateable index being highly desirable (for instance, allowing for easy addition and removal of strains depending on task), [c] adaptable alignment strategies to allow for more variability and recombinants in viral genomes (for instance, mapping against a pan genomic viral database); ability to deal with different use cases (viral with high variability, fungi with lower variability and different index), [e] mapping against pan-genome graphs, [f] approaches for build these pan-genome graphs (streaming/online concepts, succinct data structures), and [g] need for visualization and higher level access.

4.8.5 Quantitative metagenomics

While the metagenomic paradigm has been very useful in understanding the taxonomic and functional make-up of microbial communities, it is also important to understand the limitations of the framework used by most metagenomic studies. From data generated by whole genome shotgun sequencing, and in the absence of any calibration information related to quantitation (like spike-in of known amounts or qPCR data), the inferred taxonomic or functional profiles reflect relative abundances, and not absolute abundances, of taxonomic or functional groups. Thus it is not possible to assess microbial load (i.e. total number of microbes per unit volume or per unit mass) from these data, impacting our ability to answer important questions in many areas of microbiome research including for several biomedical applications.

The importance of data transformation of read count data was also discussed in the context of metagenome comparisons and identification of differentially abundant groups. Data transformations depend on the downstream analyses and include corrections for gene

copy number (e.g.: 16S gene copy numbers) and genome sizes. However the choice of the transformation function(s) is not always clear. Any comparison between metagenomic datasets also has to consider possible study specific biases associated with sample preparation and sequencing. To be able to compare metagenomic samples, it is important to define dissimilarity measures between samples. These measures may be based on taxonomic or functional profiles, or even derived directly from the underlying sequence similarities of raw reads. For many microbiome evaluations (like in the case of human microbiome comparisons), it is important to understand the distributions (of taxonomic and functional profiles) of reference population groups (like “healthy” individuals).

4.8.6 Functional annotation

Functional analysis of metagenomic data typically involves the annotation of predicted genes using resources like KEGG and MetaCyc/BioCyc that organize protein function at various levels including pathways. Analysis approaches then typically compute abundances of functional groups at these different levels. However, it is not obvious whether concepts like pathway abundances for metagenome data have any biological interpretation. Even the simpler computational challenge of pathway detection (in a given taxon in a metagenomic dataset) is complicated by the observation that, in pathway organization, genes are often assigned to multiple pathways, and these dependencies need to be explicitly modeled. Methods for increasing identification and resolution of functions of genes in metagenomic datasets were also discussed, including the use of co-localization information on the genome as a marker for membership in the same pathway or functional unit. Challenges in annotation of genes related to properties like Antimicrobial Resistance, Virulence, and Pathogenicity were also discussed. When annotating a gene in a metagenomic dataset for one of these properties, it is not always possible to make the inference based on a simple homology search against reference databases (like CARD, VFDB etc.). The conditions and constraints that lead to these properties are rather complex, and cannot be modeled by a simple criterion involving presence/absence of a particular gene. We need new approaches to express such constraints and conditions, and new ways to search against such complex rule sets.

4.8.7 Metagenomic assembly

Obtaining high quality genome assemblies from metagenomic datasets remains a challenge, with reconstruction of strains being a particularly challenging problem currently. Current binning tools are quite good for strain reconstruction in cases when there are not too many closely related strains. For real metagenomic data, however, misassembly is almost indistinguishable from novel strains occurring in species with high-recombination rates (read evidence is necessary to confirm the recombination event). Techniques for parsing strains from pangenome-based formalisms may be worth exploring in this context. Binning and assembly techniques based on co-occurrence patterns of contigs across multiple samples were also discussed.

4.8.8 Benchmarking and validation

Benchmarking of methods for metagenome analyses were also discussed. Common tasks to benchmark include assembly, taxonomic binning (bins represent sequences originating from one taxon; this corresponds to individual strains at the lowest level of the taxonomy), and taxonomic profiling (estimation of frequencies of different taxa in a sample). Benchmark dataset design criteria and evaluations are important components of any benchmarking. For

strain analyses, goals would be to reconstruct genomes of individual strains and to distinguish between closely related strains, representing different degrees of evolutionary relatedness in a data set: [a] given a mixture, identify species represented by an individual strain, [b] given a mixture, distinguish between two or more strains of the same species that are present simultaneously without recombination, and [c] simulate recombination within the sample.

Ways to generate benchmark datasets include [a] simulating mixing of sequences starting from isolate sequences, [b] creating mock communities by “mixing DNA together” and sequencing the mock community, and [c] using real data sets, and hold back reference genome sequences of community members as standard of truth that have been isolated and sequenced in addition from the same community.

Benchmark dataset generation should model different variables (sequencing technology, read length, coverage, insert size, etc.). Inclusion of plasmids and other kinds of non-bacterial material into the benchmark set would enable the creation of more realistic datasets.

References

- 1 Bas E. Dutilh, Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, Daan R. Speth, Victor Seguritan, Ramy K. Aziz, Ben Felts, Elizabeth A. Dinsdale, John L. Mokili and Robert A. Edwards. *A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes*. Nature Communications, 5, 4498, 2014
- 2 William B Langdon. *Mycoplasma contamination in the 1000 Genomes Project*. BioData Mining, 7:3, BioMed Central, 2014
- 3 Mette T Christiansen, Amanda C Brown, Samit Kundu, Helena J Tutill, Rachel Williams, Julianne R Brown, Jolyon Holdstock, Martin J Holland, Simon Stevenson, Jayshree Dave, CY William Tong, Katja Einer-Jensen, Daniel P Depledge and Judith Breuer. *Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples*. BMC Infectious Diseases, 14:591, BioMed Central, 2014

4.9 Haplotype Phasing

Knut Reinert (FU Berlin, DE), Niko Beerenwinkel (ETH Zürich – Basel, CH), Jason Chin (Pacific Biosciences – Menlo Park, US), Richard Durbin (Wellcome Trust Sanger Institute – Cambridge, GB), Mohammed El-Kebir (Brown University – Providence, US), Anne-Katrin Emde (New York Genome Center, US), Gunnar W. Klau (CWI – Amsterdam, NL), Veli Mäkinen (University of Helsinki, FI), Tobias Marschall (Universität des Saarlandes, DE), Alice Carolyn McHardy (Helmholtz Zentrum – Braunschweig, DE), Siavash Mirarab (University of California at San Diego, US), Kay Nieselt (Universität Tübingen, DE), Bernhard Renard (Robert Koch Institut – Berlin, DE), Enrico Siragusa (IBM TJ Watson Research Center – Yorktown Heights, US), Peter F. Stadler (Universität Leipzig, DE), Granger Sutton (The J. Craig Venter Institute – Rockville, US), Tandy Warnow (University of Illinois – Urbana-Champaign, US), and David Weese (SAP Innovation Center – Potsdam, DE)

License © Creative Commons BY 3.0 Unported license

© Knut Reinert, Niko Beerenwinkel, Jason Chin, Richard Durbin, Mohammed El-Kebir, Anne-Katrin Emde, Gunnar W. Klau, Veli Mäkinen, Tobias Marschall, Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Bernhard Renard, Enrico Siragusa, Peter F. Stadler, Granger Sutton, Tandy Warnow, and David Weese

4.9.1 Problem definition

Haplotype phasing describes the problem of reconstructing the individual haplotypes of a polyploid organism.

Different cases can be distinguished which alter the computational problem.

- Diploid genome
- Polyploid genome
- unknown ploidy (RNA-viruses, but also repeats in assembly, metagenomics clonotypes or strains)

4.9.2 Approaches

There are in general 3 different approaches to solve the problem:

- Read-based: The Next Generation Sequencing (NGS) reads obtained from sequencing machines can stem from any of the organisms genomic strands. From which is not known. Hence we have to infer an assignment of each read to a reconstructed haplotype. This can be done via the help of an MSA or without (alignment free). The problem is in general harder if the ploidy is unknown.
- Information from other experiments (arrays, etc.)
- Population approaches (trios (or available pedigree) or other groups of related individuals)

4.9.3 Discussion points

- Is it solved given that we have long reads (or will have cheap long reads some time in the future)? ⇒ No. It helps of course, but depends on SNP frequency, error rate, sequencing depth
- Depending on the problem (e.g. potato has high SNP frequency) various technologies can be applied (long or short reads)
- The problem can be simplified or solved using approaches from molecular biology (e.g. separating haplotypes by microfluidics, inbreeding)

- The problem is the complement problem to error correction (either its a sequencing error or a SNP)
- It is confounded by possible other error sources (sequencing errors, MSA errors)
- Ploidy > 4 cannot needs several SNPs to be resolved. Ploidy (from a computational perspective) not a constant number either globally or locally.

4.9.4 Challenges

- How to integrate other data (Hi-C), how to joint phasing (analysis of multiple samples)?
- Can scaffold information and haplotype information be integrated?
- How to estimate uncertainty (conflict between probabilistic methods and optimization)
- De novo / reference-free haplotyping (for bad or non existing reference genomes, see [1] that gives a partition of the reads which could help assembly
- Simulators should be adapted to take into biological parameters into account
- Can visualization of phased haplotypes (on the population scale) [2] help for optimization of haplotypes?

References

- 1 Mikko Rautiainen, Leena Salmela and Veli Mäkinen. *Identification of Variant Compositions in Related Strains Without Reference*. Algorithms for Computational Biology, LNCS volume 9702, pp. 158-170, 2016, Springer Verlag
- 2 Günter Jäger, Alexander Peltzer and Kay Nieselt. *inPHAP: Interactive visualization of genotype and phased haplotype data*. BMC Bioinformatics, 2014, BioMed Central

4.10 Pan-Genomics

Knut Reinert (FU Berlin, DE), Jason Chin (Pacific Biosciences – Menlo Park, US), Fabio Cunial (MPI – Dresden, DE), Simon Gog (KIT – Karlsruher Institut für Technologie, DE), André Kahles (ETH Zürich, CH), Birte Kehr (deCode Genetics – Reykjavik, IS), Oliver Kohlbacher (Universität Tübingen, DE), Ben Langmead (Johns Hopkins University – Baltimore, US), Alice Carolyn McHardy (Helmholtz Zentrum – Braunschweig, DE), Siavash Mirarab (University of California at San Diego, US), Kay Nieselt (Universität Tübingen, DE), Enno Ohlebusch (Universität Ulm, DE), Adam M. Phillippy (National Institutes of Health – Rockville, US), Simon J. Puglisi (University of Helsinki, FI), Gunnar Rätsch (ETH Zürich, CH), Karin Remington (Computationality, US), Bernhard Renard (Robert Koch Institut – Berlin, DE), Peter F. Stadler (Universität Leipzig, DE), Granger Sutton (The J. Craig Venter Institute – Rockville, US), German Tischler (MPI – Dresden, DE), and Shibu Yooseph (University of Central Florida – Orlando, US)

License © Creative Commons BY 3.0 Unported license

© Knut Reinert, Jason Chin, Fabio Cunial, Simon Gog, André Kahles, Birte Kehr, Oliver Kohlbacher, Ben Langmead, Alice Carolyn McHardy, Siavash Mirarab, Kay Nieselt, Enno Ohlebusch, Adam M. Phillippy, Simon J. Puglisi, Gunnar Rätsch, Karin Remington, Bernhard Renard, Peter F. Stadler, Granger Sutton, German Tischler, and Shibu Yooseph

4.10.1 Topics

Interesting topics to be studied in this field:

1. Coordinates, data structure / graph
2. Annotation

3. Query support / questions
4. Tools /standards : transition of existing tools to the pan-genome
5. How to update an existing pan-genome with a new member
6. Taxonomic / evolutionary scale
7. Storage formats

4.10.2 Data structures

Use cases for coordinate system approaches:

1. E. coli's genomes are much more dynamic than for example human genomes
2. One approach computing an explicit coordinate system is the SuperGenome [1]: Based on a WGA, the SuperGenome is defined by the concatenation of all locally collinear blocks computed from the WGA. By this the coordinate system of the SuperGenome is derived from the alignment coordinates of all concatenated blocks. Furthermore, it defines an injective mapping of each individual genome into the global coordinate system defined by the SuperGenome.

Issues when working with a pan-genome defined by a global coordinate system is the ability to map between different coordinate systems and to update it (see below).

It was also suggested to define no coordinate system, but to construct the pan-genome just consisting of blocks. The question arises how to define the blocks.

Pan-genomes defined as a graph structure:

1. Though the pan-genome graph may be cyclic, no path of an individual genome in such a graph should contain a loop.
2. Should the graph be stored / transformed into a DAG? Because many operations on a DAG are much easier than on a cycle graph.
3. One version of storing the graph is storing all paths.
 - a. Even if we store the graph, what would be an efficient way to encode the set of all observed paths? Sparse bitvectors?
4. Provocative question: What could be the reason to store the graph structure? For visualisation for example. Adjacency of genes.
 - a. Why not a context-free grammar? There are efficient data structures for that.
5. What is different between different subsets of genomes within the pan-genome. What is common? What is proximal?
6. Transition probabilities on arcs?

4.10.3 Annotation

One issue is: given an annotation (in gtf format say), how is the annotation transferred to the pan-genome? Difficult to do on the graph, rather straightforward in a coordinate system

In the graph: annotations should be placed onto paths rather than on the whole graph by itself. Should the annotation also be stored together with the paths? Or independently stored? Or better to have a data structure that is able to quickly identify say conserved annotations. One other possibility is to define blocks by annotations.

4.10.4 Queries

Alignment queries:

1. Align a read to the pan-genome. Exact matching against a graph is possible, papers have also studied the efficiency of this process. Return:
 - a. just matches observed in the input sequences;
 - b. also combinations that are never observed in the input sequences.

Analysis/domain queries:

1. Gene finding: In the graph, for example identify “genes”, say, which in the graph would correspond to some kind of “bundles” (linear pieces) with a limited “thickness”.
2. Co-occurrences of alleles
3. Searching for certain graph structures
4. Use expression data for example to find genes
5. Find “motifs”.
6. GWAS-like comparisons: How does one group compare against another with respect to the graph structure?

String queries:

1. Compute the probability that a given short string S occurs in a genome described by the pan-genome.
 - a. What happens if S contains flexible, rigid gaps?
 - b. What happens if S is a PWM (e.g. as done during motif finding)?
2. Compute the probability of finding pattern T in a genome described by the pan-genome, given the occurrence of pattern S .
 - a. Probability that two patterns never occur together?
 - b. Probability that two patterns overlap?
3. Compute the probability of a new genome given a pan-genome (e.g. to decide whether a genome is more similar to one pan-genome than to another).
4. Given a short string S , return the set of all possible variants of S in a genome represented by the pan-genome. This allows to display e.g. all known variants of a gene or of a regulatory region simultaneously.
 - a. Compute the probability that a variant S' of S , rather than S itself, occurs in a genome.
5. Given a new genome S and a threshold k , return the k paths Q in the pan-genome with highest conditional probability $P(Q|S)$. Such paths could be used to input data that is missing from S , and to annotate S with a candidate recombination structure.
6. Given a new genome S , compute the average length of a recombination fragment, the probability of a recombination at a given position or inside a given interval, over all possible parses of S according to the pan-genome.
7. Given a threshold k , find a set V of k variation loci which maximize the number of variation loci not in V that can be predicted using V . This query, called tag selection, recurs in the design of SNP arrays.
8. Given a regulatory region in a genome represented by the pan-genome, and given a binding model for transcription factors, compute a probability distribution over all configurations of transcription factors bound to the regulatory region.
 - a. Compute a probability distribution over all possible translated sequences of a gene.
 - b. What is the probability that a given region in the pan-genome accomplishes a given function? (e.g. that it is an enhancer of gene expression)

9. Given a short string S , compute the set of all strings that can replace S and still produce valid genomes according to the pan-genome. Such “synonyms” might correspond e.g. to all possible variants of a given transcription factor binding site along a genome, which allow the same TF to regulate simultaneously multiple genes in different ways.
10. Given a short string S , let its context be the set of left- and right- extensions of S of a given length k . Compute the contingency table of two strings S and T , based on their sets of contexts in the pan-genome.
11. Assume that some paths in the pan-genome are marked. Compute some notion of “most discriminant features” between the marked and the unmarked paths.

Update queries to the data structure:

1. add a new unary path;
2. add a new genome;
3. “merge” N pan-genome data structures;
4. “concatenate” pan-genome structures created separately for distinct regions of the genome (e.g. for specific genes or linkage disequilibrium regions).
5. how frequent are updates in practice? and of which type?
6. One reason not to update is the change of the coordinate system.

Probably better to merge the different pan-genomes into one graph structure, because comparing different graphs tend to be very hard.

4.10.5 Tools / standards

1. Toolkit VG used on graph structures by the Durbin group (read alignment, construct the graph, ...). URL: <https://github.com/vgteam/vg>
2. Apart from that: how to devise a gene finder, and other tools on top of primary constructions
3. Comparing two pan-genomes, how to do that? Examples: two virus populations, two subpopulations of humans, two snapshots of a bacterium, cross compare tumors.

4.10.6 Taxonomic breadth

When does the pan-genomic concept break down? And when does it cease to help for e.g. the querying tasks?

References

- 1 A. Herbig, G. Jäger, F. Battke and K. Nieselt. *GenomeRing: alignment visualization based on SuperGenome coordinates*. Bioinformatics, 28:12, pp. i7–i15, Oxford Journals, 2012.

Participants

- Niko Beerenwinkel
ETH Zürich – Basel, CH
- Ewan Birney
European Bioinformatics
Institute – Cambridge, GB
- Christina Boucher
Colorado State University –
Fort Collins, US
- Jason Chin
Pacific Biosciences –
Menlo Park, US
- Pascal Costanza
Intel Corporation, BE
- Anthony J. Cox
Illumina – United Kingdom, GB
- Fabio Cunial
MPI – Dresden, DE
- Richard Durbin
Wellcome Trust Sanger Institute –
Cambridge, GB
- Mohammed El-Kebir
Brown Univ. – Providence, US
- Anne-Katrin Emde
New York Genome Center, US
- Simon Gog
KIT – Karlsruher Institut für
Technologie, DE
- Hannes Hauswedell
FU Berlin, DE
- Daniel H. Huson
Universität Tübingen, DE
- André Kahles
ETH Zürich, CH
- Birte Kehr
deCode Genetics – Reykjavik, IS
- Gunnar W. Klau
CWI – Amsterdam, NL
- Oliver Kohlbacher
Universität Tübingen, DE
- Ben Langmead
Johns Hopkins University –
Baltimore, US
- Pietro Lio'
University of Cambridge, GB
- Veli Mäkinen
University of Helsinki, FI
- Tobias Marschall
Universität des Saarlandes, DE
- Alice Carolyn McHardy
Helmholtz Zentrum –
Braunschweig, DE
- Siavash Mirarab
University of California at San
Diego, US
- Laurent Mouchard
University of Rouen, FR
- Gene Myers
MPI – Dresden, DE
- Luay Nakhleh
Rice University – Houston, US
- Kay Nieselt
Universität Tübingen, DE
- Enno Ohlebusch
Universität Ulm, DE
- Adam M. Phillippy
National Institutes of Health –
Rockville, US
- Mihai Pop
University of Maryland – College
Park, US
- Simon J. Puglisi
University of Helsinki, FI
- Gunnar Rätsch
ETH Zürich, CH
- Tobias Rausch
EMBL – Heidelberg, DE
- Knut Reinert
FU Berlin, DE
- Karin Remington
Computationality, US
- Bernhard Renard
Robert Koch Institut –
Berlin, DE
- S. Cenk Sahinalp
Simon Fraser University –
Burnaby, CA
- Enrico Siragusa
IBM TJ Watson Res. Center –
Yorktown Heights, US
- Peter F. Stadler
Universität Leipzig, DE
- Granger Sutton
The J. Craig Venter Institute –
Rockville, US
- German Tischler
MPI – Dresden, DE
- Esko Ukkonen
University of Helsinki, FI
- Tandy Warnow
University of Illinois –
Urbana-Champaign, US
- David Weese
SAP Innovation Center –
Potsdam, DE
- Shibu Yooseph
University of Central Florida –
Orlando, US

