



DAGSTUHL REPORTS

Volume 9, Issue 8, August 2019

Software Protection Decision Support and Evaluation Methodologies (Dagstuhl Seminar 19331) <i>Bjorn De Sutter, Christian Collberg, Mila Dalla Preda, and Brecht Wyseur</i>	1
Algorithms and Complexity for Continuous Problems (Dagstuhl Seminar 19341) <i>Dmitriy Bilyk, Aicke Hinrichs, Frances Y. Kuo, and Klaus Ritter</i>	26
Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction (Dagstuhl Seminar 19342) <i>Rolf Backofen, Yael Mandel-Gutfreund, Uwe Ohler, and Gabriele Varani</i>	49
Computational Proteomics (Dagstuhl Seminar 19351) <i>Nuno Bandeira and Lennart Martens</i>	70
Computation in Low-Dimensional Geometry and Topology (Dagstuhl Seminar 19352) <i>Maarten Löffler, Anna Lubiw, Saul Schleimer, and Erin Moriarty Wolf Chambers</i>	84

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

February, 2020

Bibliographic information published by the Deutsche Nationalbibliothek

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

License

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



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

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

The copyright is retained by the corresponding authors.

Aims and Scope

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

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

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

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

Editorial Board

- Gilles Barthe
- Bernd Becker
- Daniel Cremers
- Stephan Diehl
- Reiner Hähnle
- Lynda Hardman
- Oliver Kohlbacher
- Bernhard Mitschang
- Bernhard Nebel
- Albrecht Schmidt
- Wolfgang Schröder-Preikschat
- Raimund Seidel (*Editor-in-Chief*)
- Emanuel Thomé
- Heike Wehrheim
- Verena Wolf
- Martina Zitterbart

Editorial Office

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

Contact

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

<http://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.9.8.i

Software Protection Decision Support and Evaluation Methodologies

Edited by

Bjorn De Sutter¹, Christian Collberg², Mila Dalla Preda³, and Brecht Wyseur⁴

1 Ghent University, BE, bjorn.desutter@ugent.be

2 University of Arizona – Tucson, US, collberg@cs.arizona.edu

3 University of Verona, IT, mila.dallapreda@univr.it

4 Kudelski Group SA – Cheseaux, CH, brecht.wyseur@nagra.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19331 “Software Protection Decision Support and Evaluation Methodologies”. The seminar is situated in the domain of software protection against so-called man-at-the-end attacks, in which attackers have white-box access to the software that embeds valuable assets with security requirements such as confidentiality and integrity. The attackers try to compromise those by reverse-engineering the software and by tampering with it. Within this domain, the seminar focused mainly on three aspects: 1) how to evaluate newly proposed protections and attackers thereon; 2) how to create an appropriate benchmark suite to be used in such evaluations; 3) how to build decision support to aid users of protection tool with the selection of appropriate protections. The major outcomes are a structure for a white-paper on software protection evaluation methodologies, with some concrete input collected on the basis of four case studies explored during the seminar, and a plan for creating a software protection benchmark suite.

Seminar August 11–16, 2019 – <http://www.dagstuhl.de/19331>

2012 ACM Subject Classification Security and privacy → Software and application security

Keywords and phrases Benchmarks, Decision Support Systems, Evaluation Methodology, man-at-the-end attacks, metrics, predictive models, reverse engineering and tampering, software protection

Digital Object Identifier 10.4230/DagRep.9.8.1

1 Executive Summary

Christian Collberg

Mila Dalla Preda

Bjorn De Sutter

Brecht Wyseur

License  Creative Commons BY 3.0 Unported license
© Christian Collberg, Mila Dalla Preda, Bjorn De Sutter, and Brecht Wyseur

Overview and Motivation

The area of Man-At-The-End (MATE) software protection is an evolving battlefield on which attackers execute white-box attacks: They control the devices and environments and use a range of tools to inspect, analyze, and alter software and its assets. Their tools include disassemblers, code browsers, debuggers, emulators, instrumentation tools, fuzzers, symbolic execution engines, customized OS features, pattern matchers, etc.



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

Software Protection Decision Support and Evaluation Methodologies, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 1–25
Editors: Bjorn De Sutter, Christian Collberg, Mila Dalla Preda, and Brecht Wyseur



Dagstuhl Reports

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

To meet the security requirements of assets embedded in software, i.e., valuable data and code, many protections need to be composed. Those requirements include the confidentiality of secret keys and software IP (novel algorithms, novel deep learning models, ...), and the integrity of license checking code and anti-copy protections. Attackers attack them through reverse engineering and tampering, for which they use the aforementioned tools and for which they often can afford spending time and effort on executing many, highly complex and time-consuming, manual and automated analyses. The need for composing many protections follows from the fact that advanced attackers can use all the mentioned tools and try many different approaches. In other words, to be effective, the deployed protections need to protect against all possible attack vectors.

As all protections come with overhead, and as many of them have downsides that complicate various aspects of the software development life cycle (SDLC), the users of a software protection tool cannot simply deploy all available protections. Instead, they have to select the protections and their parameters for every single asset in a program, taking into account non-functional requirements for the whole program and its SDLC.

The organizers of this workshop, and many experts in their network, consider the lack of automated decision support for selecting the best protections, and the lack of a generally accepted, broadly applicable methodology to evaluate and quantify the strength of a selected combination, the biggest challenges in the domain of software protection. As a result, the deployment of software protection is most often not trustworthy, error-prone, not measurable, and extremely expensive because experts are needed and they need a lot of time, increasing the time to market.

This situation is becoming ever more problematic. For example, connected intelligent vehicles are quickly being deployed in the market now and autonomous vehicles are going to be deployed in 3-5 years. Software protection evaluation and measurement research and development must match up that pace to provide enough technology support for controllable and scientific methods to manage the quality of automotive security as key part of vehicle reliability and safety. There is hence a huge need to make progress w.r.t. software protection decision support and evaluation methodologies, the topic of the proposed seminar.

Goals of the Seminar

Following a pre-seminar survey among the registered participants to focus the seminar and to select the highest priority objectives among the many possible ones, the primary goal of the seminar was determined to be the foundations of a white paper on software protection evaluation methodologies, to be used as a best practices guideline by researchers and practitioners when they evaluate (combinations of) defensive and/or offensive techniques in the domain of MATE software protection. This can also serve as a guideline to reviewers of submitted journal and conference papers in which novel techniques are proposed and evaluated. A secondary goal was the establishment of good benchmarking practices, including the choice of suitable benchmarks and the selection and generation thereof for use in future research in MATE software protection. A third goal was to collect feedback and ideas on how to push the state of the art in decision support systems.

Week Overview

Preparation

Prior to the seminar, the organizers set up a survey to collect the necessary information for a seminar bundle that provided background information about and to all participants. Moreover, they collected information regarding the potential outcomes that participants were most interested in, to which ones they could likely contribute, and which potential outcomes they considered most likely to make progress on. Furthermore, a reading list was presented to the participants with the goal of getting everyone on the same page as much and as soon as possible [1–8].

Whereas the schedule for the first two days was mostly fixed a priori, the schedule for later days was more dynamic, as it was adapted to the feedback obtained by the organizers during the early days, and to the outcomes of different sessions.

Monday

The first day was devoted to setting the scope of the seminar, and clarifying the seminar goals, strategy, and plan. In the morning, three overviews were presented of man-at-the-end software protection techniques in the scope of the seminar, as well as some attacks on them. These presentations focused on obfuscation vs. static analysis, (anti-)tampering in online games, and additional protections beyond the ones discussed in the first two presentations.

In the early afternoon, four deeper technical introductions were presented of four more concrete classes of defensive and corresponding offensive techniques that would serve as case studies throughout the seminar: 1) virtual machine obfuscation, 2) (anti-)disassembly, 3) trace semantics based attacks, and 4) data obfuscation. The strategy for the week was to brainstorm about these concrete techniques first, in particular on how the strength of these techniques are supposed to be evaluated, e.g., in papers that present novel (combinations of) techniques, or in penetration tests. Later, the concrete results for the individual case studies would then be generalized into best practices and guidelines for software protection evaluation methodologies.

Whereas the morning presentations and most of the case studies focused mostly on defensive techniques, three presentations in the afternoon provided complementary insights about offensive techniques, ranging from more academic semantics-based attack techniques, over an industrial case study of deobfuscation of compile-time obfuscation, and offensive techniques in binary analysis.

Thus, the scene was set in terms of both defensive and offensive techniques, and all participants to a large degree spoke the same language before starting the brainstorm sessions in the rest of the week.

Tuesday

Tuesday focused mostly on the seminar track of software protection evaluation methodologies.

In the early morning, additional input was provided on existing, already studied aspects relevant to such methodologies. This included software protection metrics, empirical experiments to assess protections, and security economics. These presentations provided useful hooks for the next session, which consisted of parallel, small break-out brainstorm sessions (three groups per case study) on the first two case studies. In these brainstorm sessions, the goal was to provide answers to questions such as the following:

- What would a document similar to the SIGPLAN empirical evaluation checklist look like for papers presenting new VM-based protections?
- Which requirements or recommendations can we put forward with respect to the protected objects (i.e., benchmarks) and their treatment (i.e., how they are created, compiled, ...) for the evaluation?
- What aspects of the attack models and which assumptions should be made explicit, which ones should be justified, e.g., regarding attacker goals and attacker activities.
- How should sensitivity to different inputs (e.g., random generator seeds, configuration options, features of code samples, ...) be evaluated and discussed?
- What threats to validity should be discussed?
- What aspects of the protection should be evaluated (potency, resilience, learnability, usability, stealth, renewability, different forms of costs, ...)?
- Under what conditions would you consider the protection to be “real world” applicable?
- What flaws (e.g., unrealistic assumptions) have you seen in existing papers that should be avoided?
- What are (minimal) requirements / recommendations regarding reproducibility?
- What pitfalls can you list that we should share with people?

After the independent brainstorms in small groups and following lunch, the three groups per case study came together to merge the results of their brainstorms, after which the merged results were shared in a plenary session.

Later in the afternoon, additional ideas were presented on topics relevant for software protection evaluation methodologies. The covered topics were benchmark generation, security activities in protected software product life cycles, the resilience of software integrity protection (work in progress), and a (unified) measure theory for potency. These topics were presented after the initial brainstorms not to bias those brainstorms. Their nature was more forward looking, covering a number of open challenges as well as potential directions for future research. They offered the speakers a sound board to get feedback and could serve as the starting point of informal discussions later in the seminar.

While the practice is discouraged by the Dagstuhl administration, we still decided to organize an evening session on Tuesday. Afterwards, we realized that this made the seminar a bit too dense, but it did serve the useful purpose of introducing the participants to the seminar track on decision support tools for software protection early enough in the seminar to allow enough time for informal discussions with and between researchers active on this topic during the remainder of the week. This was especially useful to allow those academic researchers to check the validity of some of their assumptions about real-world aspects with the present practitioners from industry and with researchers from other domains.

Besides an overview of an existing design and implementation of a software protection decision support system, a hands-on walk through of a practical attack on a virtual machine protection (as in one of the case studies) was presented, as well as some ideas to make such protection stronger.

Wednesday

Early on Wednesday morning, the focus shifted towards decision support tools, with three presentations by practitioners in companies that provide software protection solutions. These presentations focused on the support they provide to help their customers use their tools.

Later in the morning, case studies 3 and 4 were discussed in another round of parallel, small group break-out brainstorm sessions.

In the afternoon, the social outing took place, which consisted of a visit to Trier and a wine tasting at a winery where we also had dinner.¹

Thursday

On Thursday morning, another round of break-out sessions was organized to structure the outcomes of the first round. Based on inputs collected during the first three days, the organizers drafted a structure for a white paper on software protection methodologies. In 4 parallel sessions, the participants brainstormed on how to fit the results of the first round (i.e., bullet points with concrete guidelines and considerations for each case study) into that structure, and which parts of those results could be generalized beyond the individual case studies. In a plenary session, the results of these break-outs were then presented.

In addition, the specific topic of benchmarking was discussed, focusing on questions regarding the required features of benchmarks (e.g., should or should they not contain actual security-sensitive assets) as well as potential strategies to get from the situation today, in which very few benchmarks used in papers are available for reproducing the results, to a situation in which a standard set of benchmarks is available and effectively used in studies.

In the afternoon, several demonstrations of practical tools were given, including the already mentioned decision support system of which the concepts had been presented on Tuesday evening and the Binary Ninja disassembler that is rapidly gaining popularity. Two presentations were also given on usable security and challenges and capabilities of modern static analysis of obfuscated code. There provided additional insights useful for both designers of decision support tools and evaluation methodologies.

Friday

The last morning started off with a potpourri of interesting topics that did not fit well in the main tracks of general evaluation methodologies and decision support on the one hand, and benchmarking on the other. Given the availability of many experts in the domain of software protection, we decided that everyone that wanted to launch new ideas or collect feedback on them in the broad domain of the seminar should have that chance. So the day started with short presentations on the protection of machine learning as a specific new type of application, on security levels for white-box cryptography, and on hardware/software binding using DRAM.

Later in the morning, the seminar was wrapped up with a discussion of the outcomes so far, and an agreement on plans to continue the work on the software protection evaluation methodology white paper and the assembly of a benchmark collection.

References

- 1 S. Schrittwieser, S. Katzenbeisser, J. Kinder, G. Merzdovnik, and E. Weippl: Protecting software through obfuscation: Can it keep pace with progress in code analysis? *ACM Comput. Surv.*, **49**(1), 2016.
- 2 M. Ceccato, P. Tonella P, C. Basile, P. Falcarin, M. Torchiano, B. Coppens, and B. De Sutter: Understanding the behaviour of hackers while performing attack tasks in a professional setting and in a public challenge. *Empirical Software Engineering* 2018; **24**(1):240–286.

¹ For some reason, most of us don't remember the rest of the evening in enough detail to report on it reliably.

- 3 B. Cataldo, D. Canavese, L. Regano, P. Falcarin, and B. De Sutter: A Meta-model for Software Protections and Reverse Engineering Attacks. *Journal of Systems and Software* 150 (April): 3–21, 2019
- 4 B. Yadegari, B. Johannesmeyer, B. Whitely, and S. Debray: A generic approach to automatic deobfuscation of executable code. In: *Proc. IEEE Symposium on Security and Privacy*, pp. 674–691 (2015)
- 5 T. Blazytko, M. Contag, C. Aschermann, and T. Holz: Syntia: synthesizing the semantics of obfuscated code. *Proc. of the 26th USENIX Security Symposium (SEC'17)*, pp. 643–659. 2017
- 6 S. Banescu, C. Collberg, and A. Pretschner: Predicting the Resilience of Obfuscated Code Against Symbolic Execution Attacks via Machine Learning. *Proc. of the 26th USENIX Conference on Security Symposium (SEC'17)*, pp. 661-678, 2017
- 7 C. Basile et al.: D5.11 ASPIRE Framework Report. Technical Report ASPIRE project. <https://aspire-fp7.eu/sites/default/files/D5.11-ASPIRE-Framework-Report.pdf>
- 8 M. Ceccato et al.: D4.06 ASPIRE Security Evaluation Methodology – Security Evaluation. Technical Report ASPIRE project. <https://aspire-fp7.eu/sites/default/files/D4.06-ASPIRE-Security-Evaluation-Methodology.pdf>

2 Table of Contents

Executive Summary

<i>Christian Collberg, Mila Dalla Preda, Bjorn De Sutter, and Brecht Wyseur</i>	1
---	---

Overview of Talks


On the resilience of software integrity protection techniques (work in progress) <i>Mohsen Ahmadvand</i>	9
Automated Deobfuscation: A Tour on Semantic Attacks <i>Sébastien Bardin</i>	9
An Expert System for Software Protection <i>Cataldo Basile</i>	10
Hardening VM Semantics <i>Tim Blazytko and Moritz Contag</i>	10
An Introduction to Security Economics <i>Richard Clayton</i>	10
Introduction to the virtual machine obfuscation case study <i>Christian Collberg</i>	11
Software Protection Benchmark Generation <i>Christian Collberg</i>	11
Introduction to the (anti-)disassembly case study <i>Bart Coppens</i>	12
Securing workflows for industrial Use Cases <i>Jorge R. Cuéllar</i>	12
Introduction to the data obfuscation case study <i>Mila Dalla Preda</i>	13
Empirical Software Protection Experiments <i>Bjorn De Sutter</i>	13
Extra protections and attack in seminar scope <i>Bjorn De Sutter</i>	13
Introduction to the trace-semantics-based attack case study <i>Bjorn De Sutter</i>	14
Metrics for Software Protection Evaluation <i>Bjorn De Sutter</i>	14
A (unified) measure theory for potency? <i>Roberto Giacobazzi</i>	14
Security Activities in Protected SW Product Life Cycle <i>Yuan Xiang Gu</i>	15
Security Problems of AI/ML Applications <i>Yuan Xiang Gu, Mila Dalla Preda, and Roberto Giacobazzi</i>	15
(State of) The Art of War: Offensive Techniques in Binary Analysis <i>Christophe Hauser</i>	16

Hardware / Software Binding Using DRAM PUFs <i>Stefan Katzenbeisser</i>	16
Decision processes @ Guardsquare <i>Eric Lafortune</i>	17
Binary Ninja Demonstration <i>Peter Lafosse</i>	17
Usable Security <i>Katharina Pfeffer</i>	17
Case Study in Deobfuscation: Compile-Time Obfuscation <i>Rolf Rolles</i>	17
Protecting software through obfuscation: Can it keep pace with progress in code analysis? <i>Sebastian Schrittwieser, Stefan Katzenbeisser, Johannes Kinder, Georg Merzdovnik, and Edgar Weippl</i>	18
Software Protection, Cloakware Style <i>Bahman Sistany</i>	18
Security levels for white-box crypto <i>Atis Straujums</i>	19
Cheating in Online Games <i>Stijn Volckaert</i>	19
Modern Static Analysis of Obfuscated Code <i>John Wagner</i>	19
Kudelski Decision Support <i>Brecht Wyseur</i>	20
Seminar introduction <i>Brecht Wyseur</i>	20
Working groups on Software Evaluation Methodology White Paper	
Class 1: (Anti-) Disassembly	21
Class 2: Trace-based Attack Techniques	23
Participants	25

3 Overview of Talks

3.1 On the resilience of software integrity protection techniques (work in progress)

Mohsen Ahmadvand (TU München, DE)

License  Creative Commons BY 3.0 Unported license
© Mohsen Ahmadvand

In this talk we present our ongoing work on a methodology for measuring the resilience of software integrity protection techniques. Our methodology is comprised of catalogs of attacks, defences, and metrics. Attacks aid attackers to detect and/or to disable protections. Defences, on the other hand, hinder specific attacks and hence raise the bar. Metrics aim to capture the effectiveness of attacks or defences. We use a combination of empirical and analytical evaluations to measure the difficulty that is added by different combination of defences against attacks. Lastly, we present some preliminary results of machine learning based attacks on different composition of protections.

3.2 Automated Deobfuscation: A Tour on Semantic Attacks

Sébastien Bardin (CEA LIST, FR)

License  Creative Commons BY 3.0 Unported license
© Sébastien Bardin

Joint work of Sébastien Bardin, Richard Bonichon, Jean-Yves Marion

MATE attacks aim at taking advantage of a program once access to its executable code is granted. Typical goals include stealing critical assets (e.g., cryptographic keys or proprietary code) or software tampering (e.g., bypassing security checks). Obfuscation aims at defending against such attacks by turning the initial program into a very-hard-to-understand equivalent code. Obfuscation has thus become highly important in IP protection, leading to a arm race between obfuscation and deobfuscation techniques.

Recently, semantic analysis coming from source-level safety analysis have been proven to be highly efficient against standard code protections, leading Schrittwieser et al. asking whether “Obfuscation can keep pace with progress in code analysis”. Notably, Symbolic Execution combines both the standard advantages of semantic methods (automatic inference of values and triggers) with the robustness of dynamic analysis (allowing to bypass advanced protections such as packing and self-modification).

In this talk, we will review recent advances in automated semantic deobfuscation, with a special emphasis on Symbolic Execution and SMT solvers, together with an overview of their strengths, limitations and potential mitigation.

References

- 1 S. Banescu, C. Collberg, V. Ganesh, Z. Newsham, and A. Pretschner. Code obfuscation against symbolic execution attacks. In *Annual Conference on Computer Security Applications, ACSAC 2016*, 2016.
- 2 Sébastien Bardin, Robin David, and Jean-Yves Marion. Backward-bounded DSE: targeting infeasibility questions on obfuscated codes. In *2017 IEEE Symposium on Security and Privacy, SP*, 2017.

- 3 D. Brumley, C. Hartwig, Z. Liang, J. Newsome, D. Song, and H. Yin. Automatically identifying trigger-based behavior in malware. In Wenke Lee, Cliff Wang, and David Dagon, editors, *Botnet Detection: Countering the Largest Security Threat*, volume 36 of *Advances in Information Security*, pages 65–88. Springer, 2008.
- 4 J. Salwan, S. Bardin, and M.-L. Potet. Symbolic deobfuscation: from virtualized code back to the original. In *5th Conference on Detection of Intrusions and malware & Vulnerability Assessment (DIMVA)*, 2018.
- 5 S. Schrittwieser, S. Katzenbeisser, J. Kinder, G. Merzdovnik, and E. Weippl. Protecting software through obfuscation: Can it keep pace with progress in code analysis? *ACM Comput. Surv.*, 49(1), 2016.
- 6 B. Yadegari, B. Johannesmeyer, B. Whitely, and S. Debray. A generic approach to automatic deobfuscation of executable code. In *Symposium on Security and Privacy, SP*, 2015.

3.3 An Expert System for Software Protection


Cataldo Basile (Polytechnic University of Torino, IT)

License  Creative Commons BY 3.0 Unported license
© Cataldo Basile

This presentation presents the current status of the Decision Support System for Software Protection developed during the EC-funded ASPIRE project and now maintained by the Security Group of the Politecnico di Torino. Moreover, we present open issues and several hints for new research and collaborations to be discussed during this seminar. In addition to the presentation that discusses concepts of the decision support system, a live demonstration was presented as well.

3.4 Hardening VM Semantics

Tim Blazytko (Ruhr-Universität Bochum, DE) and Moritz Contag (Ruhr-Universität Bochum, DE)

License  Creative Commons BY 3.0 Unported license
© Tim Blazytko and Moritz Contag

We discuss limitations of current VM-based obfuscation schemes and introduces automated attacks that reveal the core semantics of VM instructions. Afterwards, we propose hardening techniques which defeat the latter.

3.5 An Introduction to Security Economics

Richard Clayton (University of Cambridge, GB)

License  Creative Commons BY 3.0 Unported license
© Richard Clayton

This talk gave a very brief overview of the field of security economics as it has evolved over the past twenty years. Technical analysis of security failures allows us to work out which part of a system failed; security economics helps us understand why the system was built that

way in the first place – and hence how we can redesign it to be more resilient in the future. The key economic ideas are: Incentives and Liability – if Alice is being protected by Bob, then Bob will be far more motivated if he loses out, rather than Alice, should the system fail. Externalities and Negative externalities – does a system naturally move towards a monoculture; are you dumping costs onto other people? Moral Hazard – are you encouraging bad behaviour? Asymmetric Information – Akerlof’s “Market for Lemons” applies to security solutions just as much as to the market in used cars. Conflict Theory can be explained in relation to the Island of Anarchia whose flood defences are as good as the “least efforts” of the laziest family building their section of the sea wall; their ability to repel the Athenian Navy depends on the skill of their “best shot”, but their trade surplus as a group will depend on “sum of efforts”. This leads to an insight into software development – security depends on the worst effort of the sloppiest programmer (who writes a buffer overflow), on the best efforts of the security architect (so hire the best you can afford) and the sum of efforts of the testers (the more testing you do, the fewer bugs you should ship). The talk finished with some observations from a Security Economics perspective on CAPTCHAs, which have been broken by simple “hacks” rather than by advances in AI or graphics; on a Connect 4 competition – where it was realised that the aim was to win the competition rather than to play excellent Connect 4; and finally the story of “DVD Jon” whose DeCSS program “broke” the DVD Content Scrambling System at the end of the last century.

3.6 Introduction to the virtual machine obfuscation case study

Christian Collberg (University of Arizona – Tucson, US)

License  Creative Commons BY 3.0 Unported license
© Christian Collberg

In this talk I will give an overview of obfuscation by virtual machine generation. To virtualize a function F in order to protect some asset A , we 1) create a unique and random virtual instruction set I specific to F ; 2) translate the F into the virtual instruction set I (the bytecode array); and 3) construct an interpreter that can execute programs written in I . This interpreter consists of an execution stack, a dispatch unit which issues the next instruction, and one instruction handler per virtual instruction. We will discuss numerous ways to attack a virtual machine by reverse engineering the instruction set, the dispatch, or the instruction handlers. We will further discuss ways to protect the virtual machine against such attacks using diversification of the instruction set, obfuscating instruction handlers and dispatch units, or turning parts of virtual machines into dynamically generated code.

3.7 Software Protection Benchmark Generation

Christian Collberg (University of Arizona – Tucson, US)

License  Creative Commons BY 3.0 Unported license
© Christian Collberg

Researchers in software protection and malware analysis face a similar problem: what programs should they test their techniques on? Often, two different papers, solving similar problems, will perform evaluation on vastly different sets of benchmark programs. Hence, it becomes difficult to compare their results. Sometimes, researchers use performance

benchmarks as security benchmarks. This is problematic, since they were not designed with security in mind. For example, there is no specific “asset” to protect and hence no clear meaning of when an attack has succeeded. In this talk we will discuss a how to automatically generate security benchmark suites. The idea is that, to evaluate a new protection idea we should 1) generate random benchmark programs, 2) run those through the protection tool, and 3) attack these randomly generated challenges through a choice of reverse engineering tools. Generating random programs turns out to be a challenging problem. A random program P should, at minimum, fulfill the following requirements. P may contain different types of assets; P should have simple I/O behavior, making it easy to automate attacks; P should have “interesting” internal structure; P should terminate within a time bound; P should not be guessable from its I/O behavior; P should resemble a real program; Finally, it should be obvious when an attack has succeeded. In this talk we will present two types of random program generators we have constructed: namely generators of random hash functions, and generators of random CAPTCHA programs.

3.8 Introduction to the (anti-)disassembly case study

Bart Coppens (Ghent University, BE)

License  Creative Commons BY 3.0 Unported license
© Bart Coppens

This talk has the purpose of bringing the audience to the same level of background knowledge and terminology with regards to disassembly and anti-disassembly of binary programs. I start with techniques for unobfuscated programs. First, I talk about the very basic difference between linear sweep and recursive descent disassembly techniques. Then I explain how the resulting disassembly can be leveraged to reconstruct the original programs at ever-higher levels of abstractions. In particular, I talk about control flow recovery and how the quality of the disassembly can influence the resulting control flow graph. Next, I talk about anti-disassembly techniques, and how these can influence both the quality of the resulting disassembly as well as the quality of the reconstructed control flow graph. Finally, I talk about some advanced disassembly techniques whose purpose it is to deal correctly with binaries that have had such anti-disassembly techniques applied to them.

3.9 Securing workflows for industrial Use Cases

Jorge R. Cuéllar (Siemens AG – München, DE)

License  Creative Commons BY 3.0 Unported license
© Jorge R. Cuéllar

In industrial Use Cases it is often more important to secure the integrity of the process (manufacturing, testing, collaborative design, cloud applications, Industrial IoT-based Control Systems, Supply Chain, etc) than the confidentiality of the workflow itself. Different users (employees of different companies, notification bodies, governmental stakeholders or NGOs) collaborate, using smart devices like handhelds, to execute a workflow in a predefined form. The proposal combines the use of Petri-Nets for modelling the workflows and a logic (divided into 4 different layers: PKI, Trust reasoning, “Snippet”-reasoning, Accountability) that can be used to exchange information (tokens, similar to ACE/OAUTH tokens) and to reason locally

about the tokens in order to secure the integrity of the workflow. The PKI layer determines which certificates to verify for which secrets or public keys, the trust layer determines which parties may claim which assertions, the snippet layer verifies the single transactions of the Petri Net and the accountability layer provides a method for a judge to find which server is responsible for an incorrect decision.

3.10 Introduction to the data obfuscation case study

Mila Dalla Preda (University of Verona, IT)

License © Creative Commons BY 3.0 Unported license
© Mila Dalla Preda

Data are important components of programs and their values, evolution and structure provides important information for program understanding. With the term data obfuscation we refer to those protection techniques that target data. In particular, data obfuscation techniques often modify the encoding of data in order to prevent direct analysis and hide the content of data. In this introduction, we presented a short overview to ensure that all participants had a basic understanding of the subject.

3.11 Empirical Software Protection Experiments

Bjorn De Sutter (Ghent University, BE)

License © Creative Commons BY 3.0 Unported license
© Bjorn De Sutter

We present an overview of goals, pitfalls, issues and best practices in empirical experiments for determining the strength of software protections. This includes the technical aspects such as which protections are combined in the treatment of the objects, which tasks are given to the subjects, but also methodological aspects such as learning effect testing, statistical methods, threats to validity, etc.

3.12 Extra protections and attack in seminar scope

Bjorn De Sutter (Ghent University, BE)

License © Creative Commons BY 3.0 Unported license
© Bjorn De Sutter

We present an overview of attacks and protections in the scope of the seminar to complement the first two talks. This includes some of the tools that attackers use such as disassemblers, emulators, and debuggers, as well as a short overview of automated methods for deobfuscation. It also includes protections against all kinds of attacker-activities, such as anti-tampering, anti-debugging, anti-taint-tracking, etc.

3.13 Introduction to the trace-semantics-based attack case study

Bjorn De Sutter (Ghent University, BE)

License  Creative Commons BY 3.0 Unported license
© Bjorn De Sutter

As an introduction to the breakout sessions on evaluation techniques, case study 3 comprises two attack techniques: generic deobfuscation by Yadegari et al. and Syntia by Blazytko et al. An overview of these techniques is presented, and some potential issues with the evaluation are enumerated.

References

- 1 B. Yadegari, B. Johannesmeyer, B. Whitely and S. Debray. *A Generic Approach to Automatic Deobfuscation of Executable Code*. 2015 IEEE Symposium on Security and Privacy, San Jose, CA, 2015, pp. 674-691.
- 2 Tim Blazytko, Moritz Contag, Cornelius Aschermann, and Thorsten Holz. *Syntia: synthesizing the semantics of obfuscated code*. In Proceedings of the 26th USENIX Conference on Security Symposium (SEC'17), Engin Kirda and Thomas Ristenpart (Eds.). USENIX Association, Berkeley, CA, USA, pp. 643-659.

3.14 Metrics for Software Protection Evaluation


Bjorn De Sutter (Ghent University, BE)

License  Creative Commons BY 3.0 Unported license
© Bjorn De Sutter

We present an overview of opportunities and challenges for using quantitative metrics for evaluation of software protections. This includes a discussion of some existing metrics from the domain of software engineering, pitfalls in using them on protected software. We also discuss the relation between attacker effort of individual steps of an attack path and features such as potency and resilience.

3.15 A (unified) measure theory for potency?

Roberto Giacobazzi (University of Verona, IT)

License  Creative Commons BY 3.0 Unported license
© Roberto Giacobazzi

We observe that there exists a potentially infinite set of metrics for measuring the potency of code obfuscation. Any code attack can be encoded into an (abstract) interpreter-based attack. Because there are infinitely many interpreters, this implies that there are infinitely many metrics, one for each attack. This means that looking at standard SW engineering-like metrics is clueless in this field. All these metrics have anyway some common aspects: They are not measure of complexity in the sense of Blum's and they all try to measure the level of uncertainty that an attacker (i.e., an abstract interpreter) gets out of the performed analysis. I think that we should shift the measure of potency from SW engineering-like metrics to entropy. The presentation shows the main challenges in this direction.

3.16 Security Activities in Protected SW Product Life Cycle

Yuan Xiang Gu (Irdeto – Ottawa, CA)

License © Creative Commons BY 3.0 Unported license
© Yuan Xiang Gu

This talk is to aim a better understanding what the best industrial security practices may be looking for from this seminar. First, we discuss three aspects of economics of security for a protected SW product:

- 1) Challenges to design a secure System
- 2) Do security right at early stage
- 3) SW security debt

And then, we clarify 9 kinds of security activities during a protected SW product life cycle from early stages of requirements, architecture design and implementation design, to code stages of implementation, testing and assurance, to the post stages of deployment, monitoring, update and renew. By these discussions, it is very clear that during different development and maintenance stages of a protected SW product, security activities require different kinds of security guidelines and evaluation approaches and supports to make right decision. Also, based on our experience from our own practices in past more than 20 years, we present some real constrains which security technologies, methods and approaches should be compliant to and suitable for real adaption and uses.

3.17 Security Problems of AI/ML Applications

Yuan Xiang Gu (Irdeto – Ottawa, CA), Mila Dalla Preda (University of Verona, IT), and Roberto Giacobazzi (University of Verona, IT)

License © Creative Commons BY 3.0 Unported license
© Yuan Xiang Gu, Mila Dalla Preda, and Roberto Giacobazzi

This presentation is to introduce a new research subject on AI/ML security. AI/ML technology is getting much more adaptations for many applications in past 10 years. Recently, more and more high-stake applications start to adapt AI/ML as well. There are a couple of driving forces for AI/ML's recent success, but security is not a real important play factor yet. On the other hand, researchers are focusing on very narrow AI/ML security on the adversarial ML problem that is a special and serious issue. Our recently research shows that AI/ML security has much broader security scope well beyond adversarial ML problem. Moreover, we suggest that software protection can be adapted to address many security problems of AI/ML application systems. This talk just gives some highlights of our findings and would like to raise awareness by sharing them with a list of open questions and suggestions for how to move this new research forward.

3.18 (State of) The Art of War: Offensive Techniques in Binary Analysis

Christophe Hauser (USC – Marina del Rey, US)

License © Creative Commons BY 3.0 Unported license
© Christophe Hauser

Joint work of Christophe Hauser, Audrey Dutcher, Siji Feng, John Grosen, Christopher Kruegel, Mario Polino, Christopher Salls, Yan Shoshitaishvili, Nick Stephens, Giovanni Vigna, and Ruoyu Wang

Main reference Yan Shoshitaishvili, Ruoyu Wang, Christopher Salls, Nick Stephens, Mario Polino, Andrew Dutcher, John Grosen, Siji Feng, Christophe Hauser, Christopher Krügel, Giovanni Vigna: “SOK: (State of) The Art of War: Offensive Techniques in Binary Analysis”, in Proc. of the IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016, pp. 138–157, IEEE Computer Society, 2016.

URL <https://doi.org/10.1109/SP.2016.17>

Finding and exploiting vulnerabilities in binary code is a challenging task. The lack of high-level, semantically rich information about data structures and control constructs makes the analysis of program properties harder to scale. However, the importance of binary analysis is on the rise. In many situations binary analysis is the only possible way to prove (or disprove) properties about the code that is actually executed. In this paper, we present a binary analysis framework that implements a number of analysis techniques that have been proposed in the past. We present a systematized implementation of these techniques, which allows other researchers to compose them and develop new approaches. In addition, the implementation of these techniques in a unifying framework allows for the direct comparison of these approaches and the identification of their advantages and disadvantages. The evaluation included in this paper is performed using a recent dataset created by DARPA for evaluating the effectiveness of binary vulnerability analysis techniques. Our framework has been open-sourced and is available to the security community.

3.19 Hardware / Software Binding Using DRAM PUFs

Stefan Katzenbeisser (Universität Passau, DE)

License © Creative Commons BY 3.0 Unported license
© Stefan Katzenbeisser

Joint work of Stefan Katzenbeisser, Wenjie Xiong, Andre Schaller, Jakub Szefer

Low-end computing devices are becoming increasingly ubiquitous, especially due to the widespread deployment of Internet-of-Things products. There is, however, much concern about sensitive data being processed on these low-end devices which have limited protection mechanisms in place. This paper proposes a Hardware-Entangled Software Protection (HESP) scheme that leverages hardware features to protect software code from malicious modification before or during run-time. It also enables implicit hardware authentication. Thus, the software will execute correctly only on an authorized device and if the timing of the software, e.g., control flow, was not changed through malicious modifications. The proposed ideas are based on the new concept of Dynamic Physically Unclonable Functions (PUFs). Dynamic PUFs have time-varying responses and can be used to tie the software execution to the timing of software and the physical properties of a hardware device. It is further combined with existing approaches for code self-checksumming, software obfuscation, and call graph and register value scrambling to create the HESP scheme. HESP is demonstrated on commodity, off-the-shelf computing devices, where a DRAM PUF is used as an instance of a Dynamic PUF.

3.20 Decision processes @ Guardsquare


Eric Lafortune (Guardsquare – Leuven, BE)

License  Creative Commons BY 3.0 Unported license
© Eric Lafortune

Guardsquare develops software to protect mobile apps against reverse engineering and tampering. Its users are engineers who need to integrate and configure the software to process their apps, and then evaluate the results. The current approach is driven by the technology. Required configuration to make sure processed Android apps continue to work is facilitated by the widespread use of our open-source software ProGuard. Configuration to actually harden the apps follows the same conventions. Based on our experience, configuration to harden iOS apps works at a higher declarative level. External penetration testers typically provide feedback.

3.21 Binary Ninja Demonstration

Peter Lafosse (Vector 35 – Melbourne, US)

License  Creative Commons BY 3.0 Unported license
© Peter Lafosse

A hands-on demonstration was given of the disassembler tool Binary Ninja, a tool growing in popularity for reverse engineering of binaries.

3.22 Usable Security

Katharina Pfeffer (SBA Research – Wien, DE)

License  Creative Commons BY 3.0 Unported license
© Katharina Pfeffer

Usable security aims to investigate how IT systems can be designed so that end-users and software developers use them correctly and the attack surface is minimized. In this talk we present 2 research projects on usable security and discuss how the insights gained and the methodology used can help defending MATE attacks and develop appropriate metrics for prevention evaluation.

3.23 Case Study in Deobfuscation: Compile-Time Obfuscation

Rolf Rolles (Mobius Strip Reverse Engineering – San Francisco, US)

License  Creative Commons BY 3.0 Unported license
© Rolf Rolles

Software obfuscation has always been a controversially discussed research area. While theoretical results indicate that provably secure obfuscation in general is impossible, its widespread application in malware and commercial software shows that it is nevertheless popular in practice. Still, it remains largely unexplored to what extent today's software

obfuscations keep up with state-of-the-art code analysis, and where we stand in the arms race between software developers and code analysts. The main goal of this survey is to analyze the effectiveness of different classes of software obfuscation against the continuously improving de-obfuscation techniques and off-the-shelf code analysis tools. The answer very much depends on the goals of the analyst and the available resources. On the one hand, many forms of lightweight static analysis have difficulties with even basic obfuscation schemes, which explains the unbroken popularity of obfuscation among malware writers. On the other hand, more expensive analysis techniques, in particular when used interactively by a human analyst, can easily defeat many obfuscations. As a result, software obfuscation for the purpose of intellectual property protection remains highly challenging.

3.24 Protecting software through obfuscation: Can it keep pace with progress in code analysis?

Sebastian Schrittwieser (FH – St. Pölten, AT), Stefan Katzenbeisser (Universität Passau, DE), Johannes Kinder, Georg Merzdovnik, and Edgar Weippl

License © Creative Commons BY 3.0 Unported license
© Sebastian Schrittwieser, Stefan Katzenbeisser, Johannes Kinder, Georg Merzdovnik, and Edgar Weippl

Main reference Sebastian Schrittwieser, Stefan Katzenbeisser, Johannes Kinder, Georg Merzdovnik, Edgar R. Weippl: “Protecting Software through Obfuscation: Can It Keep Pace with Progress in Code Analysis?”, *ACM Comput. Surv.*, Vol. 49(1), pp. 4:1–4:37, 2016.

URL <https://doi.org/10.1145/2886012>

Software obfuscation has always been a controversially discussed research area. While theoretical results indicate that provably secure obfuscation in general is impossible, its widespread application in malware and commercial software shows that it is nevertheless popular in practice. Still, it remains largely unexplored to what extent today’s software obfuscations keep up with state-of-the-art code analysis, and where we stand in the arms race between software developers and code analysts. The main goal of this survey is to analyze the effectiveness of different classes of software obfuscation against the continuously improving de-obfuscation techniques and off-the-shelf code analysis tools. The answer very much depends on the goals of the analyst and the available resources. On the one hand, many forms of lightweight static analysis have difficulties with even basic obfuscation schemes, which explains the unbroken popularity of obfuscation among malware writers. On the other hand, more expensive analysis techniques, in particular when used interactively by a human analyst, can easily defeat many obfuscations. As a result, software obfuscation for the purpose of intellectual property protection remains highly challenging.

3.25 Software Protection, Cloakware Style

Bahman Sistani (Irdeto – Ottawa, CA)


License © Creative Commons BY 3.0 Unported license
© Bahman Sistani

In this talk, we introduce Irdeto’s Cloakware Software Protection and go into some detail about the Obfuscation engine and various obfuscations and transformations that it offers. We present a summary of how guidance is given to users on the use of Irdeto’s Software

Protection to protect their assets. We discuss why there is a need to rank code/data entities in terms of level/kind of protection needed and how performance and size considerations are taken into account. We also cover how heuristics may be used to identify and rank entities as low and high security and how we can draw further inferences about the whole code units and use it as a basis for creating training datasets for Machine Learning based security application. Finally some early research results in ML models and their datasets were presented.

3.26 Security levels for white-box crypto


Atis Straujums (whiteCryption – Riga, LV)

License  Creative Commons BY 3.0 Unported license
© Atis Straujums

We present a potential list of levels, assigned to our white-box algorithms based on their strength. The level criteria don't define a strict ordering but nevertheless help quickly assess the strength of protection and decide whether to spend any more resources on improving it for each algorithm.

3.27 Cheating in Online Games


Stijn Volckaert (KU Leuven – Ghent, BE)

License  Creative Commons BY 3.0 Unported license
© Stijn Volckaert

In this talk, I presented an overview of the most prevalent types of cheats used in competitive online games. I discussed the functionality of these cheats and explained which techniques cheat coders use to construct them. I then shifted to cheat protection tools (so-called anti-cheats). After reviewing the overall architecture of an anti-cheat, I zoomed in on my own anti-cheat tool, ACE, which I built to protect Unreal Engine 1 games. I briefly talked about ACE's most prominent features and then reflected on things that have not worked when rolling out new functionality.

3.28 Modern Static Analysis of Obfuscated Code


John Wagner (Vector 35 – Melbourne, US)

License  Creative Commons BY 3.0 Unported license
© John Wagner

Static analysis tools have improved significantly in recent years. This talk is an exploration of how modern static analysis tools analyze binary code and its impact on deobfuscation techniques. Various obfuscation techniques are discussed, including those that have been defeated by modern tools, those that are easier to defeat using the scripting features of these tools, and those that are still very difficult to analyze.

3.29 Kudelski Decision Support

Brecht Wyseur (Kudelski Group SA – Cheseaux, CH)

License  Creative Commons BY 3.0 Unported license
© Brecht Wyseur

Kudelski has been developing Software Protection Tools internally. First for use on its own products – to protect Digital TV applications on a wide variety of devices. Now this has become also a product offering where Kudelski is helping its customers to protect its applications. In this presentation, we elaborate on the constraints and requirements that have been taken into account and how this has been managed in the product design and development processes.

3.30 Seminar introduction

Brecht Wyseur (Kudelski Group SA – Cheseaux, CH)

License  Creative Commons BY 3.0 Unported license
© Brecht Wyseur

At the start of the seminar, we presented an opening introduction presentation, setting the scene and aligning on the seminar objectives. We also did a tour de table.

4 Working groups on Software Evaluation Methodology White Paper

As described in the executive summary, different groups brainstormed on evaluation methodology best practices for different use cases, to serve as the initial input to a white paper. The goal of this white paper is not to prescribe how exactly evaluations should be done, but rather what aspects are relevant to consider explicitly in evaluations, which assumptions might make sense and which might not. As a good way to convey, we consider the following potential structure of a software protection paper:

1. Abstract & Introduction
2. Attack Model – Background – Related Work
3. Technical Contribution
4. Evaluation
5. Discussion
6. Availability
7. Conclusions

Our guidelines are not bound to such a structure, nor do we put forward this structure. It simply allows us to put some structure in the many relevant aspects, as we can then formulate advice on what aspects to consider and discuss in each section.

For each of those supposed sections, we can then formulate advice that would be relevant for (almost all) papers in the domain of MATE software protection, or advice that would only be relevant for papers focused on either offensive or defensive techniques, or advice that would only be relevant for specific classes of techniques, such as trace-based techniques or static analyses, or for specific tools, such as obfuscators or disassemblers. So on top of the aforementioned structure, we structure the white paper itself into the following parts:

1. Introduction – to the white paper.
2. Methodology – explaining the methodology followed to get to the white paper.
3. General Principles
4. Defensive Techniques
5. Offensive Techniques
6. Benchmarks – specific advice on the use of benchmarks.
7. Appendices – one per separate class of techniques.

As this is an evolving field, the white paper would be a living document.

For the four case studies discussed during the seminar, the initial input for the corresponding appendices was collected, and potential plans were agreed upon to continue the necessary work towards an initial publishable white paper. Here, we list, as an example, a number of items for two of them: trace-based attack techniques on the one hand, and (anti-)disassembly techniques on the other. Principles or practices considered to be clearly more generally applicable are marked with an asterisk, even when some examples are given to clarify them that clearly only apply to the technique at hand. We do not repeat such generally applicable concepts in the second use case.

4.1 Class 1: (Anti-) Disassembly

4.1.1 Abstract & Introduction

- * Explicitly mention the specific goals of the defense or attack methods that your technique tries to overcome or to mitigate: disassembly, hiding code, identifying function starts, letting the disassembler produce incomplete information (such as incomplete CFGs) or letting it produce wrong information (i.e., incorrect graphs), identifying all basic blocks within a function, hindering dataflow analysis, call graph reconstruction, increase false-positive or false-negative rates of function starts, branches, function “ends”, etc
- * Explicitly mention the advantages of your technique.
- * Discuss upfront negative side-effects and impacts, as well as limitations
- * Discuss upfront the maturity of the proposed technique (e.g., tested on state-of-the-art combinations of protections or on single play version of one obfuscation, multiple platforms or not)

4.1.2 Attack Model – Background – Related Work

- * Describe the attack goals against which you defend, e.g., extracting instructions, function starts, modifying code, symbolic execution, dataflow analysis, ...
- * Discuss which attacks are out of scope, e.g., instruction tracing.
- * Make assumptions explicit, e.g., regarding dynamic analysis being difficult or impossible on some target, regarding the disassembler working on (the basis of) executable files on disks, memory dumps, traces, regarding defender and attacker capabilities, limitations on existing state-of-the-art techniques that you will use or build on.
- * Discuss relevant, concrete (real-world) scenarios in which the proposed technique will be demonstrated or is claimed to be useful, i.e., pushes the state of the art.
- Discuss related work that and the extent to which it targets cases that your technique can handle or is compatible with, such as overlapping instructions, polyglot code, dynamic jumps, self-modifying code, architectures designed against static disassembly (next instruction depends on current one – see Malbolge), opaque predicates, edit distance as a measure.

4.1.3 Technical Contribution

- * Explicitly discuss technical constraints, such as where in the build process is this applied (compile time, link time, post-link, runtime, etc)? What platform (incl. OS) / architecture / compilers / compiler options and other features are required or are the techniques limited to (e.g., data in code or not).
- * Explicitly discuss any diverges in the experimental setup / prototype implementation from the relevant scenarios that were discussed in the attack model section.

4.1.4 Evaluation

- * Explicitly mentioned the baseline you are comparing against.
- * Measure performance on standard benchmarks.
- * Advice in the form of “You can use metrics X, Y, and Z, for stealth, potency, performance, cost, resilience, ... but not A, B, and C because they are useless”.
- Consider scalability and sensitivity regarding program size, amount of indirection in it, amount of aliasing in code when data flow analysis is used,...
- Ideally use multiple disassemblers and not just the tools out-of-the-box, but, e.g., with additional scripts that make up for the fact that the existing tools might not include some simple heuristics they would have improved their performance on your binaries but that were irrelevant before.
- Ideally do not evaluate on binaries with only your new protection, but in combination with relevant protections taken from existing state of the art (commercial or academic).
- You should look ahead a bit in the cat-and-mouse-game: not all possible attacks on a new defense should be tested, but an analysis of some basic new attacks or small adaptations to existing attacks should be discussed, and ideally already be evaluated (e.g., by writing and running IDA Pro plugins that mimic simple heuristics that an attacker-improved version of IDA Pro would include once attackers get to know your new protection).
- As long as the benchmarks span the relevant ranges as needed to assess scalability and sensitivity, any benchmarks will typically do, such as SPEC (i.e., no assets required in software).
- Provide at least 1 set of ground truth experiments, account for false positives & false negatives.
- Moving towards a standard set of benchmarks would be good.

4.1.5 Discussion

- * Discuss threats to validity
- * In case there were divergences between real-world attack model and experimental setup, the potential impact thereof should be discussed.

4.1.6 Availability

- * Anything that matters for reproducibility and to let others build on your results should be discussed, including the following items
- * Our code is available under license XXX.
- * Our benchmarks are standard, and available at ZZZ.
- * Output data is available at ...
- * Intermediate results are available ...
- * Scripts for summarizing results from raw data are available ...

4.1.7 Conclusions

4.2 Class 2: Trace-based Attack Techniques

4.2.1 Attack Model – Background – Related Work

- Explicitly describe and define the attack goals (e.g., finding keys, finding data value, modifying data value, ...) in relation to legitimate software protection or malware. For example, define what it means to “deobfuscate” in the context of your paper, and why that is chosen as a goal (e.g., undo VM-based protection or analyse code at the bytecode level).
- Be explicit about the defenses and features thereof that you consider in scope or not, as illustrated in the next bullets.
- Depending on the type of tracing (including data flow analysis on the fly or not), different types of tracing tools might be necessary, so limitations should be discussed if any exist that relate to anti-emulation protections that can break the use of tools.
- The granularity of traces should be discussed.
- Stealth is very important to defend against these attacks (hide fragments or operations that are relevant), so consider what stealth measures you can handle and how.
- Taint-tracking is difficult to do because it is easy to thwart techniques, so if your method depends on it, discuss this explicitly.
- Hiding boundaries between relevant and irrelevant code is important and easy (e.g., by inlining sensitive code) for many techniques, e.g., because they handle short sequences in traces that need to be identified first. So benchmarks should span a wide range in terms of stealth and complexity, but of course that also means we need a good definition of stealth first.
- It is okay to assume that a trace can always be collected, given the capabilities of modern virtualization technologies.
- It is okay to assume that deterministic replay of an execution on one input is possible. Making the trace artificially different for different inputs can be a real problem, so don't assume delta-techniques are trivial.
- If relevant, take into account in practice, deployed obfuscations might not be limited to small parts (that are easily identifiable).
- * The paper should explicitly cover whether or not the attacker is assumed to know the obfuscator internals, and whether or not the paper assumes security through obscurity (preferably not).

4.2.2 Evaluation

- * Use microbenchmarks to measure performance (CPU and memory runtime overhead, code size overhead)
- * Also measure performance on standard benchmarks, for example taken from a competition.
- * When analyzing obfuscations, it is important to measure the variability they introduce in the execution and report averages and confidence intervals.
- * Check scalability by measuring performance on benchmarks that covers a range (includes program size & complexity, trace size & complexity), measure time, memory consumption and anything else needed. Identify bottlenecks if there is an issue with scalability.

- * Both evaluation for specific protection-attack combinations (i.e., single protection) and on more real-world relevant cases (with more protections combined) can be useful and are ideally included. If you only chose one, explain why it's enough.

4.2.3 Discussion

- * Threads to validity should include how easy or difficult it is to port the proposed techniques to different platforms or setups and to deploy them in other scenarios.
- Coverage requirements should be discussed. What are the expectations in terms of coverage and how does it affect the validity of the attack?

Participants

- Mohsen Ahmadvand
TU München, DE
- Sébastien Bardin
CEA LIST, FR
- Cataldo Basile
Polytechnic University of
Torino, IT
- Tim Blazytko
Ruhr-Universität Bochum, DE
- Richard Bonichon
CEA LIST – Nano-INNOV, FR
- Richard Clayton
University of Cambridge, GB
- Christian Collberg
University of Arizona –
Tucson, US
- Moritz Contag
Ruhr-Universität Bochum, DE
- Bart Coppens
Ghent University, BE
- Jorge R. Cuéllar
Siemens AG – München, DE
- Mila Dalla Preda
University of Verona, IT
- Bjorn De Sutter
Ghent University, BE
- Laurent Dore
EDSI – Cesson-Sevigne, FR
- Ninon Eyrolles
Paris, FR
- Roberto Giacobazzi
University of Verona, IT
- Yuan Xiang Gu
Irdeto – Ottawa, CA
- Christophe Hauser
USC – Marina del Rey, US
- Stefan Katzenbeisser
Universität Passau, DE
- Eric Lafortune
Guardsquare – Leuven, BE
- Peter Lafosse
Vector 35 – Melbourne, US
- Patrik Marcacci
Kudelski Security –
Cheseaux, CH
- J. Todd McDonald
University of South Alabama –
Mobile, US
- Christian Mönch
Conax – Oslo, NO
- Leon Moonen
Simula Research Laboratory –
Lysaker, NO
- Jan Newger
Google Switzerland – Zürich, CH
- Katharina Pfeffer
SBA Research – Wien, DE
- Yannik Potdevin
Universität Kiel, DE
- Uwe Resas
QuBalt GmbH, DE
- Rolf Rolles
Mobius Strip Reverse
Engineering – San Francisco, US
- Sebastian Schrittwieser
FH – St. Pölten, AT
- Bahman Sistany
Irdeto – Ottawa, CA
- Natalia Stakhanova
University of Saskatchewan –
Saskatoon, CA
- Atis Straujums
whiteCryption – Riga, LV
- Stijn Volckaert
KU Leuven – Ghent, BE
- John Wagner
Vector 35 – Melbourne, US
- Andreas Weber
Gemalto – München, DE
- Brecht Wyseur
Kudelski Group SA –
Cheseaux, CH
- Michael Zunke
SFNT Germany GmbH –
München, DE



Algorithms and Complexity for Continuous Problems

Edited by

Dmitriy Bilyk¹, Aicke Hinrichs², Frances Y. Kuo³, and
Klaus Ritter⁴

1 University of Minnesota – Minneapolis, US, dbilyk@math.umn.edu

2 Johannes Kepler Universität Linz, AT, aicke.hinrichs@jku.at

3 UNSW Sydney, AU, f.kuo@unsw.edu.au

4 TU Kaiserslautern, DE, ritter@mathematik.uni-kl.de

Abstract

From 18.08. to 23.08.2019, the Dagstuhl Seminar 19341 Algorithms and Complexity for Continuous Problems was held in the International Conference and Research Center (LZI), Schloss Dagstuhl. During the seminar, participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar can be found in this report. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Seminar August 18–23, 2019 – <http://www.dagstuhl.de/19341>

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis, Theory of computation → Lower bounds and information complexity, Mathematics of computing → Stochastic differential equations, Mathematics of computing → Approximation, Mathematics of computing → Quadrature

Keywords and phrases applied harmonic analysis, computational stochastics, computing and complexity in infinite dimensions, discrepancy theory, tractability analysis

Digital Object Identifier 10.4230/DagRep.9.8.26

Edited in cooperation with David Krieg, JKU Linz, AT, david.krieg@jku.at


1 Executive Summary

Dmitriy Bilyk

Aicke Hinrichs

Frances Y. Kuo

Klaus Ritter

License  Creative Commons BY 3.0 Unported license
© Dmitriy Bilyk, Aicke Hinrichs, Frances Y. Kuo, and Klaus Ritter

This was already the 13th Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems over a period of 28 years. It brought together researchers from different communities working on complexity of continuous problems. Such problems, which originate from numerous areas, including physics, chemistry, finance, and economics, can almost never be solved analytically, but rather only approximately to within some error threshold. The complexity analysis ideally includes the construction of (asymptotically) optimal algorithms. Although the seminar title has remained the same, many of the topics and participants change with each seminar and each seminar in this series is of a very interdisciplinary nature. The current seminar attracted 41 participants from nine different countries all over the world. About 30% of them were young researchers including PhD students. There were 34 presentations.



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

Algorithms and Complexity for Continuous Problems, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 26–48

Editors: Dmitriy Bilyk, Aicke Hinrichs, Frances Y. Kuo, and Klaus Ritter



DAGSTUHL
REPORTS

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

The following topics were covered:

Tractability analysis of high-dimensional problems: Tractability analysis is an area of applied mathematics and theoretical computer science that studies the minimal computational resources needed for the approximate solution of problems with a huge number of variables, and it can be seen as a unifying theme for the preceding seminars in this series. Many concrete problems from applications have been analyzed in this context, new algorithms were developed, approaches to break the curse of dimensionality were established, but there remain a number of important open problems. Tractability analysis will serve as a guideline and a tool for establishing complexity results and for constructing algorithms for infinite dimensional problems.

Computational stochasticity: The focus was on weak and strong approximation as well as on the quadrature problem for stochastic ordinary or partial differential equations, i.e., on models with a random dynamics in a finite- or infinite-dimensional state space. A major topic was the complexity analysis for stochastic differential equations under non-standard assumptions.

Computing and complexity in infinite dimensions: Computational problems with infinitely many variables naturally arise in rather different application areas. Results and techniques from tractability analysis are available and thus permit one to study infinite dimensional problems as the limit of finite dimensional ones. Moreover, the availability of generic types of algorithms, like the multivariate decomposition method or the multi-level approach, will contribute to the complexity analysis and practical application in integration and approximation problems of infinitely many variables.

Discrepancy theory: Classical discrepancy theory is concerned with the question how uniformly finite point sets can be distributed. The geometric notion of discrepancy is intimately connected to the complexity of integration for functions from certain function classes. For problems in both fixed low dimension and high dimension, there are intriguing open questions whose solution would impact both fields of discrepancy theory and tractability studies.

Computational/applied harmonic analysis: Harmonic analysis plays an increasingly important role both in discrepancy theory and tractability analysis. One highlight is the proof of the currently best known lower bound for the star discrepancy in fixed dimension, which showed close connections between different areas, so similar techniques could be used to establish better bounds for the celebrated small ball problem for Gaussian processes. Equally important for the workshop is that many of the interesting spaces of functions occurring in numerical problems are well suited to the application of harmonic analysis.

As we understand better and better, these subjects are highly interrelated, and they are probably the most active and promising ones in the fields for the next decade. Bringing together a mix of junior and senior researchers from these diverse but interrelated subjects in a Dagstuhl seminar resulted in considerable progress both for the theory and the applications in these areas.

Seminars in applied mathematics and theoretical computer science typically consist of presentations, followed by short discussions in the plenum, and numerous informal discussions in smaller groups. In this seminar, we added another new feature. A moderator was assigned to three preselected talks (based on their particular relevance and on the experience of the speaker) in order to inspire a longer, in-depth discussion in the plenum. The three speakers were Jan Vı́byral, Erich Novak, and Martin Hutzenthaler. The talks were scheduled as the

first talks on Tuesday, Wednesday and Thursday. It was indeed very inspiring to witness the long and deep discussions following these special talks. We feel that this format was successful and should be used also in other workshops and conferences of the community.

The work of the attendants was supported by a variety of funding agencies. This includes the Deutsche Forschungsgemeinschaft, the Austrian Science Fund, the National Science Foundation (USA), and the Australian Research Council.

As always, the excellent working conditions and friendly atmosphere provided by the Dagstuhl team have led to a rich exchange of ideas as well as a number of new collaborations. Selected papers related to this seminar will be published in a special issue of the *Journal of Complexity*.

2 Table of Contents

Executive Summary

Dmitriy Bilyk, Aicke Hinrichs, Frances Y. Kuo, and Klaus Ritter 26

Overview of Talks

Bound on the expected number of function evaluations required to approximate the minimum of a smooth Gaussian process <i>James M. Calvin</i>	31
Lattice Algorithms for Multivariate Approximation in Periodic Spaces with General Weight Parameters <i>Ronald Cools, Frances Y. Kuo, Dirk Nuyens, and Ian Sloan</i>	31
Convergence in Hölder and Sobolev norms for approximations of Gaussian fields <i>Sonja Cox</i>	32
CLTs for stochastic approximation schemes under non-standard assumptions <i>Steffen Dereich</i>	32
The spectral decomposition of discrepancy kernels on manifolds <i>Martin Ehler</i>	32
Nested multilevel Monte Carlo and use of approximate random variables <i>Michael Giles</i>	33
Mixed Randomized Sequences, Negative Dependence, and Probabilistic Discrepancy Bounds <i>Michael Gnewuch</i>	33
Multilevel Monte Carlo methods for estimating the expected value of sample information <i>Takashi Goda</i>	34
Adaptive Quantile Computation for Brownian Bridge in Change-Point Analysis <i>Mario Hefter</i>	35
Overcoming the curse of dimensionality for parabolic PDEs <i>Martin Hutzenthaler</i>	35
Quasi-Monte Carlo Methods and Artificial Neural Networks <i>Alexander Keller</i>	36
Fast simulation of non-stationary Gaussian random fields <i>Kristin Kirchner</i>	36
The power of random information <i>David Krieg, Aicke Hinrichs, Erich Novak, Joscha Prochno, Mario Ullrich</i>	37
Exponential tractability of linear tensor product problems <i>Peter Kritzer</i>	37
Optimal confidence for Monte Carlo integration of smooth functions <i>Robert J. Kunsch and Daniel Rudolf</i>	38
Uniform Recovery Guarantees for Least Squares Approximation <i>Lutz Kämmerer</i>	38

In the search for all zeros of smooth functions <i>Leszek Plaskota</i>	39
Convergence order of the Euler-Maruyama scheme in dependence of the Sobolev regularity of the drift <i>Michaela Szölgényi</i>	39
On strong approximation of SDEs with a discontinuous drift coefficient <i>Thomas Müller-Gronbach and Larisa Yaroslavtseva</i>	40
Algorithms and Complexity for Functions on General Domains <i>Erich Novak</i>	40
Lattice algorithms for approximation: new constructions <i>Dirk Nuyens, Ronald Cools, Ian H. Sloan, and Frances Y. Kuo</i>	40
Randomized Euler scheme for strong approximation of SDEs under Sobolev- Slobodeckij smoothness <i>Paweł Przybyłowicz</i>	41
Tractability properties of discrepancy <i>Friedrich Pillichshammer</i>	41
Large deviations in geometric functional analysis <i>Joscha Prochno</i>	42
Wasserstein contraction and spectral gap of simple slice sampling <i>Daniel Rudolf</i>	43
Complexity of stochastic integration <i>Stefan Heinrich</i>	43
Potential Theory, inverse Laplacians and new Low(?) -Discrepancy Sequences <i>Stefan Steinerberger</i>	43
Dimension-independent convergence of Gaussian process regression <i>Aretha Teckentrup</i>	44
Sampling discretization error of integral norms for function classes <i>Vladimir N. Temlyakov</i>	44
Discrepancy, Dispersion and Fixed Volume Discrepancy <i>Mario Ullrich</i>	44
Approximation of shallow neural networks <i>Jan Vybíral</i>	45
Randomized Smolyak Algorithm: Explicit Cost Bounds and an Application to Infinite-Dimensional Integration <i>Marcin Wnuk</i>	45
Tractability for Volterra problems with convolution kernels <i>Henryk Wozniakowski</i>	47
Approximation complexity for additive random fields <i>Marguerite Zani</i>	47
Participants	48

3 Overview of Talks

3.1 Bound on the expected number of function evaluations required to approximate the minimum of a smooth Gaussian process

James M. Calvin (NJIT – Newark, US)

License © Creative Commons BY 3.0 Unported license
© James M. Calvin

We consider the problem of approximating the minimum of a function using sequentially chosen points at which to evaluate the function. Given a random function, we want an algorithm that approximates the minimum to a prescribed accuracy with few function evaluations on average.

In this talk we consider the function to be a centered stationary Gaussian process on the unit interval with three-times continuously differentiable paths. We assume that the covariance function of the process has positive second and fourth spectral moments.

We describe an algorithm that takes as input an error tolerance ϵ and confidence level γ , and stops when the probability that the error exceeds ϵ is at most γ . For our probability model and algorithm, the expected number of function evaluations required, in terms of the error tolerance ϵ , is of order $\log(1 + 1/\epsilon) \log \log(1 + 1/\epsilon)$.

3.2 Lattice Algorithms for Multivariate Approximation in Periodic Spaces with General Weight Parameters

Ronald Cools (KU Leuven, BE), Frances Y. Kuo (UNSW Sydney, AU), Dirk Nuyens (KU Leuven, BE), and Ian Sloan

License © Creative Commons BY 3.0 Unported license
© Ronald Cools, Frances Y. Kuo, Dirk Nuyens, and Ian Sloan


Main reference Ronald Cools, Frances Y. Kuo, Dirk Nuyens, Ian H. Sloan: “Lattice algorithms for multivariate approximation in periodic spaces with general weight parameters”, CoRR, Vol. abs/1910.06604, 2019.

URL <http://arxiv.org/abs/1910.06604>

This talk summarizes a recent manuscript by the authors on the theoretical foundation for the construction of lattice algorithms for multivariate L_2 approximation in the worst case setting, for functions in a periodic space with general weight parameters. Our construction leads to an error bound that achieves the optimal rate of convergence for lattice algorithms.

3.3 Convergence in Hölder and Sobolev norms for approximations of Gaussian fields

Sonja Cox (University of Amsterdam, NL)

License  Creative Commons BY 3.0 Unported license
© Sonja Cox

Joint work of Sonja Cox, Kristin Kirchner

Main reference Sonja G. Cox, Kristin Kirchner: “Regularity and convergence analysis in Sobolev and Hölder spaces for generalized Whittle-Matérn fields”, CoRR, Vol. abs/1904.06569, 2019.

URL <https://arxiv.org/abs/1904.06569>

In models involving a Gaussian field one frequently assumes the covariance operator to be given by a negative fractional power of a second-order elliptic differential operator of the form $L := -\nabla \cdot (A\nabla) + \kappa^2$. Whittle-Matérn fields form an well-known example of such a model. Such covariance operators allow for a reasonable amount of model flexibility (adjustable correlation length and the smoothness of the field) whilst being relatively easy to simulate. In our work we established optimal strong convergence rates in Hölder and Sobolev norms for Galerkin approximations of such Gaussian random fields. More specifically, we considered both spectral Galerkin methods and finite element methods. The latter, although significantly more tedious to analyse, are more suitable for non-stationary fields on non-standard domains.

3.4 CLTs for stochastic approximation schemes under non-standard assumptions


Steffen Dereich (Universität Münster, DE)

License  Creative Commons BY 3.0 Unported license
© Steffen Dereich

We establish new CLTs for Ruppert-Polyak averaged stochastic gradient descent schemes. Instead of isolated attractors we consider attracting manifolds. On the event of convergence we prove a stable limit theorem which is of the optimal order $n^{-1/2}$.

3.5 The spectral decomposition of discrepancy kernels on manifolds

Martin Ehler (Universität Wien, AT)

License  Creative Commons BY 3.0 Unported license
© Martin Ehler

We study the spectral decomposition of discrepancy kernels when restricted to compact kernels of \mathbb{R}^d . For restrictions to the Euclidean ball in odd dimensions, to the rotation group $SO(3)$, and to the Grassmannian manifold, we compute the kernel’s Fourier coefficient and determine their asymptotics.

3.6 Nested multilevel Monte Carlo and use of approximate random variables

Michael Giles (University of Oxford, GB)

License © Creative Commons BY 3.0 Unported license
© Michael Giles

Joint work of Mike Giles, Oliver Sheridan-Methven

Main reference Michael B. Giles: “Multilevel Monte Carlo methods”, Acta Numer., Vol. 24, pp. 259–328, 2015.

URL <http://dx.doi.org/10.1017/S096249291500001X>

The multilevel Monte Carlo (MLMC) method has been used for a wide variety of stochastic applications. In this talk we consider its use in situations in which input random variables can be replaced by similar approximate random variables which can be computed much more cheaply. A nested MLMC approach is adopted in which a two-level treatment of the approximated random variables is embedded within a standard MLMC application. We analyse the resulting nested MLMC variance in the specific context of an SDE discretisation in which Normal random variables can be replaced by approximately Normal random variables, and provide numerical results to support the analysis.

3.7 Mixed Randomized Sequences, Negative Dependence, and Probabilistic Discrepancy Bounds

Michael Gnewuch (Universität Osnabrück, DE)

License © Creative Commons BY 3.0 Unported license
© Michael Gnewuch

Joint work of Michael Gnewuch, Benjamin Doerr, Nils Hebbinghaus, Marcin Wnuk

We consider sampling schemes in the d -dimensional unit cube $[0, 1]^d$. A simple example would be Monte Carlo (MC) points $\mathbf{X} := (X_i)_{i=1}^n$, which are independent and uniformly distributed in $[0, 1]^d$. It is known that MC points satisfy the probabilistic star discrepancy bound

$$\text{disc}^*(\mathbf{X}) \leq c\sqrt{d/n} \tag{1}$$

with positive probability (Heinrich et al. 2001, Aistleitner 2011), where the smallest value for the constant so far, $c = 2.5287$, was achieved in [2]. This bound is a pre-asymptotic bound, since it gives useful information for a moderate number of points n (only depending linearly on d) and the dependence of all constants on the number of points n and the dimension d is made explicit. So far there is no sampling scheme known that satisfies a better pre-asymptotic bound for the star discrepancy.

Our goal is to identify those sampling schemes $\mathbf{X} := (X_i)_{i=1}^n$, whose points are “well spreaded” in $[0, 1]^d$ in the sense that the probabilistic bound for the star discrepancy of X_1, \dots, X_n is (essentially) not worse than bound (1). One sufficient condition is that the sampling scheme satisfies certain negative dependence properties.

If \mathbf{X} satisfies, e.g., a certain negative dependence property with respect to arbitrary axis-parallel boxes anchored in 0, then a discrepancy bound of the form

$$\text{disc}^*(\mathbf{X}) \leq c\sqrt{d/n}\sqrt{\ln(1+n/d)},$$

c small, holds with positive probability, see [3].

If \mathbf{X} satisfies even the corresponding negative dependence property with respect to arbitrary differences of axis-parallel boxes anchored in 0, then a discrepancy bound of the form (1) for small c holds with positive probability, see [2].

Examples of sampling schemes that satisfy these negative dependence properties include, apart from MC points, Latin hypercube sampling, see [2], generalized stratified sampling or certain mixed randomized sequences, see [3].

The notion of negative dependence used is a relaxation of the notion of (upper and lower) negative orthant dependence. The relaxation allows for a parameter $\gamma \in [1, \infty)$ that in the case of negative orthant dependence is fixed to be one. In a project started at the Dagstuhl seminar 19341, we recently have been able to show that the negative dependence property for Latin hypercube samples proved in [2] for parameters $\gamma = \gamma(d) = e^d$ can actually only be proved for a $\gamma = \gamma(d)$ that grows at least of the order $\Omega(\sqrt{d})$ as d tends to infinity, see [1]. That is, although Latin hypercube sampling definitely does not satisfy negative orthant dependence with respect to arbitrary differences of axis-parallel boxes anchored in 0, it satisfies the corresponding new relaxed negative dependence property with (dimension-dependent) parameter γ .

References

- 1 B. Doerr, M. Gnewuch. *On negative dependence properties of Latin hypercube samples and scrambled nets*. In preparation.
- 2 M. Gnewuch, N. Hebbinghaus. *Discrepancy bounds for a class of negatively dependent random points including Latin hypercube samples*. Preprint 2018 (submitted).
- 3 M. Wnuk, M. Gnewuch, N. Hebbinghaus. *On negatively dependent sampling schemes, variance reduction, and probabilistic upper discrepancy bounds*. Preprint, arXiv:1904.10796. (To appear in: D. Bylik, J. Dick, F. Pillichshammer (Eds.), Proceedings of the RICAM Special Semester on Multivariate Algorithms and their Foundations in Number Theory, Linz 2018, DeGruyter.)

3.8 Multilevel Monte Carlo methods for estimating the expected value of sample information

Takashi Goda (University of Tokyo, JP)

License  Creative Commons BY 3.0 Unported license
© Takashi Goda

Joint work of Michael B. Giles, Takashi Goda, Tomohiko Hironaka, Howard Thom

Main reference Michael B. Giles, Takashi Goda: “Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI”, *Statistics and Computing*, Vol. 29(4), pp. 739–751, 2019.

URL <http://dx.doi.org/10.1007/s11222-018-9835-1>

Motivated by applications to medical decision making, we study Monte Carlo estimation of the expected value of partial perfect information (EVPPI) and the expected value of sample information (EVSU). Both EVPPI and EVSU are defined as nested expectations, for which the standard (nested) Monte Carlo methods requires $O(\varepsilon^{-3})$ or $O(\varepsilon^{-4})$ computational costs to achieve the root-mean-square accuracy ε . To reduce these costs to $O(\varepsilon^{-2})$, we introduce antithetic multilevel Monte Carlo (MLMC) estimators for these quantities in this study. Under some assumptions on decision models, the antithetic property of the MLMC estimator enables to prove such a computational complexity for estimating EVPPI (Giles and Goda, 2019). The result can be extended to EVSU, by directly using the Bayes’ formula and showing auxiliary results on the MLMC estimation of nested ratio expectations (Hironaka, Giles, Goda and Thom, in preparation). Numerical experiments support our theoretical analysis.

3.9 Adaptive Quantile Computation for Brownian Bridge in Change-Point Analysis

Mario Hefter (TU Kaiserslautern, DE)

License © Creative Commons BY 3.0 Unported license
© Mario Hefter

Joint work of Mario Hefter, Jürgen Franke, André Herzwurm, Klaus Ritter, Stefanie Schwaar

In change-point analysis, weighted partial sum processes are used to detect changes. A well-known test statistic for change-points is their maximum. Asymptotically, its distribution is specified by the supremum of a weighted Brownian bridge, for which the distribution function is not known in general such that critical values have to be calculated numerically by simulation. We construct an adaptive Monte Carlo algorithm for generating weighted Brownian bridges with the goal of approximating the distribution of their suprema. We compare the new method with the classical algorithm based on evaluating the stochastic process on an equidistant grid. For prescribed approximation quality, the new algorithm provides a much faster calculation of, e.g., critical values.

3.10 Overcoming the curse of dimensionality for parabolic PDEs

Martin Hutzenthaler (Universität Duisburg-Essen, DE)

License © Creative Commons BY 3.0 Unported license
© Martin Hutzenthaler

Joint work of Weinan E, Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, Tuan Anh Nguyen, Philippe von Wurstemberger

Main reference Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, Tuan Anh Nguyen, Philippe von Wurstemberger: “Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations”, CoRR, Vol abs/1807.01212, 2018.

URL <https://arxiv.org/abs/1807.01212>

For a long time it is well-known that high-dimensional linear parabolic partial differential equations (PDEs) can be approximated by Monte Carlo methods with a computational effort which grows polynomially both in the dimension and in the reciprocal of the prescribed accuracy. In other words, linear PDEs do not suffer from the curse of dimensionality. For general semilinear PDEs with Lipschitz coefficients, however, it remained an open question whether these suffer from the curse of dimensionality. This talk explains a new numerical approximation algorithm introduced in [1] and [2] which overcomes the curse of dimensionality in the numerical approximation of general semilinear heat equations with gradient-independent nonlinearities.

References

- 1 E, W., HUTZENTHALER, M., JENTZEN, A., AND KRUSE, T. Linear scaling algorithms for solving high-dimensional nonlinear parabolic differential equations. *arXiv:1605.00856* (2016).
- 2 HUTZENTHALER, M., JENTZEN, A., KRUSE, T., NGUYEN, T. A., AND VON WURSTEMBERGER, P. Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations. *arXiv preprint arXiv:1807.01212* (2018).

3.11 Quasi-Monte Carlo Methods and Artificial Neural Networks

Alexander Keller (NVIDIA, DE)

License © Creative Commons BY 3.0 Unported license
© Alexander Keller

Joint work of Gonçalo Mordido, Matthijs Van Keirsbilck, Alexander Keller
URL <https://developer.nvidia.com/gtc/2019/video/S9389>

The average human brain has about 10^{11} nerve cells, where each of them may be connected to up to 10^4 others. We therefore investigate the question whether there are algorithms for artificial neural networks that are linear in the number of neurons, while the number of connections incident to a neuron is bounded by a constant.

Representing artificial neural networks by paths, we offer two approaches to answer this question: First, we derive an algorithm that quantizes a trained artificial neural network such that the resulting complexity is linear [1]. Second, we demonstrate that training networks, whose connections are determined by uniform sampling can achieve a similar precision as using fully connected layers. Due to sparsity upfront, these networks can be trained much faster. Finally, we explain how generating the paths using quasi-Monte Carlo methods, especially the Sobol' low discrepancy sequence, leads to a new parallel hardware architecture for artificial neural networks.

References

- 1 Gonçalo Mordido, Matthijs Van keirsbilck, Alexander Keller. *Instant Quantization of Neural Networks using Monte Carlo Methods*. <https://arxiv.org/abs/1905.12253>, 2019

3.12 Fast simulation of non-stationary Gaussian random fields

Kristin Kirchner (ETH Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Kristin Kirchner

Joint work of Lukas Herrmann, Kristin Kirchner, Christoph Schwab
Main reference Lukas Herrmann, Kristin Kirchner, Christoph Schwab: "Multilevel Approximation of Gaussian Random Fields: Fast Simulation", *Mathematical Models and Methods in Applied Sciences*(ja), 2019.
URL <http://dx.doi.org/10.1142/S0218202520500050>

We propose and analyze multilevel algorithms for the fast simulation of possibly non-stationary Gaussian random fields (GRFs for short) indexed, e.g., by a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ or by a compact d -manifold \mathcal{M} . A *colored* GRF \mathcal{Z} , admissible for our algorithms, solves the stochastic fractional-order equation $\mathcal{A}^\beta \mathcal{Z} = \mathcal{W}$ for some $\beta > d/4$, where \mathcal{A} is a linear, local, second-order elliptic differential operator in divergence form and \mathcal{W} is white noise. We thus consider GRFs with covariance operators of the form $\mathcal{C} = \mathcal{A}^{-2\beta}$.

The proposed algorithms numerically approximate samples of \mathcal{Z} on nested sequences $\{\mathcal{T}_\ell\}_{\ell \geq 0}$ of regular, simplicial partitions \mathcal{T}_ℓ of \mathcal{D} and \mathcal{M} , respectively. Work and memory to compute one approximate realization of the GRF \mathcal{Z} on the triangulation \mathcal{T}_ℓ with consistency $\mathcal{O}(N_\ell^{-\rho})$, for some consistency order $\rho > 0$, scale essentially linear in $N_\ell = \#(\mathcal{T}_\ell)$, independent of the possibly low regularity of the GRF. The algorithms are based on a sinc quadrature for an integral representation of (the application of) the negative fractional-order elliptic operator $\mathcal{A}^{-\beta}$. For the proposed numerical approximation, we prove bounds of the computational cost and the consistency error.

3.13 The power of random information

David Krieg (Johannes Kepler Universität Linz, AT), Aicke Hinrichs (Johannes Kepler Universität Linz, AT), Erich Novak (Universität Jena, DE), Joscha Prochno (Universität Graz, AT), and Mario Ullrich (Johannes Kepler Universität Linz, AT)

License © Creative Commons BY 3.0 Unported license

© David Krieg, Aicke Hinrichs, Erich Novak, Joscha Prochno, Mario Ullrich

Main reference Aicke Hinrichs, David Krieg, Erich Novak, Joscha Prochno, Mario Ullrich: “On the power of random information”, CoRR, Vol. abs/1903/006081, 2019.

URL <https://arxiv.org/abs/1903.00681>

We study problems like recovering a function from a finite number of function values. Usually, it is assumed that these function values can be computed at arbitrary points. In this talk, we assume that we do not get to choose the points. We compare the quality of random sampling points with the quality of optimal sampling points. How much do we loose?

References

- 1 Aicke Hinrichs, David Krieg, Erich Novak, Joscha Prochno, Mario Ullrich. *Random sections of ellipsoids and the power of random information*. arXiv:1901.06639 [math.FA]
- 2 Aicke Hinrichs, David Krieg, Erich Novak, Joscha Prochno, Mario Ullrich. *On the power of random information*. arXiv:1903.00681 [math.NA]
- 3 David Krieg, Mario Ullrich. *Function values are enough for L_2 -approximation*. arXiv:1905.02516 [math.NA]

3.14 Exponential tractability of linear tensor product problems

Peter Kritzer (Österreichische Akademie der Wissenschaften – Linz, AT)

License © Creative Commons BY 3.0 Unported license

© Peter Kritzer

Joint work of Fred J. Hickernell, Peter Kritzer, Henryk Wozniakowski

Main reference Fred J. Hickernell, Peter Kritzer, Henryk Wozniakowski: “Exponential tractability of linear tensor product problems”, CoRR, Vol. abs/1811.05856, 2018.

Main reference <https://arxiv.org/abs/1811.05856>

We consider the approximation of compact linear operators defined over tensor product Hilbert spaces. Necessary and sufficient conditions on the singular values of the problem under which we can or cannot achieve different notions of exponential tractability were given by Papageorgiou, Petras, and Wozniakowski in 2017. Here we present an alternative proof method based on a more recent result to obtain these conditions. As opposed to the algebraic setting, several tractability notions cannot be achieved for non-trivial cases in the exponential setting.

3.15 Optimal confidence for Monte Carlo integration of smooth functions

Robert J. Kunsch (RWTH Aachen, DE) and Daniel Rudolf (Universität Göttingen, DE)

License © Creative Commons BY 3.0 Unported license
© Robert J. Kunsch and Daniel Rudolf

Main reference Robert J. Kunsch, Daniel Rudolf: “Optimal confidence for Monte Carlo integration of smooth functions”, CoRR, Vol. abs/1809/09890, 2018.

URL <https://arxiv.org/abs/1809.09890>

We study the information-based complexity of approximating integrals of smooth functions at absolute precision $\varepsilon > 0$ with confidence level $1 - \delta \in (0, 1)$ using function evaluations within randomized algorithms. The probabilistic error criterion is new in the context of integrating smooth functions. In previous research, Monte Carlo integration was studied in terms of the expected error (or the root mean squared error), for which linear methods achieve optimal rates of the error $e(n)$ in terms of the number n of function evaluations. In our context, usually methods that provide optimal confidence properties exhibit non-linear features. The optimal probabilistic error rate $e(n, \delta)$ for multivariate functions from classical isotropic Sobolev spaces $W_p^r(G)$ with sufficient smoothness on bounded Lipschitz domains $G \subset \mathbb{R}^d$ is determined. It turns out that the integrability index p has an effect on the influence of the uncertainty δ in the complexity. In the limiting case $p = 1$ we see that deterministic methods cannot be improved by randomization. In general, higher smoothness reduces the additional effort for diminishing the uncertainty. Finally, we add a discussion about this problem for function spaces with mixed smoothness.

3.16 Uniform Recovery Guarantees for Least Squares Approximation

Lutz Kämmerer (TU Chemnitz, DE)

License © Creative Commons BY 3.0 Unported license
© Lutz Kämmerer

Joint work of Lutz Kämmerer, Tino Ullrich, Toni Volkmer


Main reference Lutz Kämmerer: “Multiple Lattice Rules for Multivariate L_∞ Approximation in the Worst-Case Setting”, CoRR, Vol. abs/1909.02290, 2019.

URL <http://arxiv.org/abs/1909.02290>

We recapitulate recent results for least squares approximation using random point sets. In particular, for the $L_2(\mathbb{T}^d)$ approximation of functions from periodic Sobolev spaces $H_{\text{mix}}^s(\mathbb{T}^d)$ of dominating mixed smoothness s , the uniform recovery guarantees $\sup_{\|f\|_{H_{\text{mix}}^s}} \|f - \tilde{f}\|_{L_2} \lesssim n^{-s} \log^{ds} n$ hold, which is an improvement compared to best known so far sparse grid algorithms for small smoothness. Furthermore, the $L_\infty(\mathbb{T}^d)$ approximation using a set of rank-1 lattices as sampling nodes provides an efficient approximation algorithm that uses several least squares solutions in order to build up an approximation. This approach achieves the best possible main rate $s - 1/2$ in $1/n$ of the sampling error.

3.17 In the search for all zeros of smooth functions

Leszek Plaskota (University of Warsaw, PL)

License  Creative Commons BY 3.0 Unported license
© Leszek Plaskota


We report results obtained in an on going research on the problem of finding the set of all zeros of functions $f \in C^r([0, 1])$, $r \in \{0, 1, 2, \dots\}$, such that $f^{(r)}$ is Hölder continuous with exponent $\varrho \in (0, 1]$. We also allow $r = +\infty$, in which case f is infinitely many times continuously differentiable. Possible algorithms use information about values of f and/or its derivatives at n points. The error between the true solution $Z(f)$ and approximate solution $Z_n(f)$ is measured via the Hausdorff distance $d_H(Z(f), Z_n(f))$ between sets. We construct a nonadaptive algorithm using function evaluations at equally spaced points whose error converges to zero as $n \rightarrow +\infty$, for all functions f from our class. On the other hand, the convergence is arbitrarily slow. Specifically, for any sequence $\{Z_n\}_{n \geq 1}$ of approximations and for any positive sequence $\{\tau_n\}_{n \geq 1}$ converging to zero there are functions f^* having exactly one zero for which the errors $d_H(Z(f^*), Z_n(f^*))$ do not converge to zero or converge slower than τ_n .

We also note that the same results hold for finding zeros of functions from the corresponding class of multivariate functions, and for other problems, such as finding all fixed points or finding all global minima.

These results confirm a common belief that smoothness itself is not enough to have faster convergence of algorithms for those problems.

3.18 Convergence order of the Euler-Maruyama scheme in dependence of the Sobolev regularity of the drift

Michaela Szölgvényi (Alpen-Adria-Universität Klagenfurt, AT)

License  Creative Commons BY 3.0 Unported license
© Michaela Szölgvényi
Joint work of Michaela Szölgvényi, Andreas Neuenkirch

We study the strong convergence rate of the Euler-Maruyama scheme for scalar SDEs with additive noise and irregular drift. We provide a framework for the error analysis by reducing it to a weighted quadrature problem for irregular functions of Brownian motion. By analysing the quadrature problem we obtain for arbitrarily small $\epsilon > 0$ a strong convergence order of $(1 + \kappa)/2 - \epsilon$ for a non-equidistant Euler-Maruyama scheme, if the drift has Sobolev-Slobodeckij-type regularity of order $\kappa \in (0, 1)$.

3.19 On strong approximation of SDEs with a discontinuous drift coefficient


Thomas Müller-Gronbach (Universität Passau, DE) and Larisa Yaroslavtseva (Universität Passau, DE)

License  Creative Commons BY 3.0 Unported license
© Thomas Müller-Gronbach and Larisa Yaroslavtseva

Recently a lot of effort has been invested in the literature to analyze the L_p -error of the Euler-Maruyama scheme in the case of stochastic differential equations (SDEs) with a drift coefficient that may have discontinuities in space. For scalar SDEs with a piecewise Lipschitz drift coefficient and a Lipschitz diffusion coefficient that is non-zero at the discontinuity points of the drift coefficient so far only an L_p -error rate of at least $1/(2p)$ has been proven in the literature. In this talk we show that under the latter assumptions on the coefficients of the SDE the Euler-Maruyama scheme in fact achieves an L_p -error rate of at least $1/2$ for all $p \in [1, \infty)$ as in the case of SDEs with Lipschitz coefficients. We furthermore present a numerical method, which achieves an L_p -error rate of at least $3/4$ for all $p \in [1, \infty)$ if, additionally to the assumptions stated above, both the drift and the diffusion coefficients are piecewise differentiable with Lipschitz derivatives.

3.20 Algorithms and Complexity for Functions on General Domains

Erich Novak (Universität Jena, DE)


License  Creative Commons BY 3.0 Unported license
© Erich Novak

Error bounds and complexity bounds in numerical analysis and information-based complexity are often proved for functions that are defined on very simple domains, such as a cube, a torus, or a sphere. We study optimal error bounds for the approximation and integration and only assume that the domain is a bounded Lipschitz domain in \mathbb{R}^d . It is known that for many problems the order of convergence does not depend on the domain. We present examples for which the following is true:

- 1) Also the asymptotic constant does not depend on the shape of the domain, only of its volume.
- 2) There are explicit and uniform lower (or upper, respectively) bounds for the error that are only slightly smaller (or larger, respectively) than the asymptotic error bound.

3.21 Lattice algorithms for approximation: new constructions

Dirk Nuyens (KU Leuven, BE), Ronald Cools (KU Leuven, BE), Ian H. Sloan, and Frances Y. Kuo (UNSW Sydney, AU)

License  Creative Commons BY 3.0 Unported license
© Dirk Nuyens, Ronald Cools, Ian H. Sloan, and Frances Y. Kuo

We derive a new CBC algorithm for the construction of good generating vectors for rank-1 lattice point sets which can be used for approximation in the Korobov space with general weights. The good news is that this construction is independent of the index set on which we represent our approximated function which makes for nice and fast construction algorithms in the case of product, POD and SPOD weights.

3.22 Randomized Euler scheme for strong approximation of SDEs under Sobolev-Slobodeckij smoothness

Paweł Przybyłowicz (AGH Univ. of Science & Technology-Krakow, PL)

License © Creative Commons BY 3.0 Unported license
© Paweł Przybyłowicz

Joint work of Paweł Przybyłowicz, Raphael Kruse

We investigate the problem of strong approximation of solution of the following scalar SDE

$$\begin{cases} dX(t) = a(t, X(t))dt + b(t)dW(t), & t \in [0, T], \\ X(0) = \eta, \end{cases} \quad (2)$$

driven by a standard one-dimensional Wiener process $W = (W(t))_{t \in [0, T]}$. We assume that $a = a(t, y)$ and $b = b(t)$ are only measurable with respect to the time variable t , and a is globally Lipschitz with respect to the space variable y .

We investigate behavior of the randomized Euler scheme X_n^{RE} , which evaluates a and b at randomly chosen points. By using Information-Based Complexity framework we show that randomized Euler scheme converges to the solution X of the underlying SDE but the convergence of X_n^{RE} to X may be arbitrarily slow ([5]). In order to get positive results we assume that b belongs to the Sobolev-Slobodeckij space $W^{\sigma, p}$, $\sigma \in (0, 1)$, $p > 2$. In this case we show that the $L^2(\Omega)$ -error of the algorithm X_n^{RE} is $O(n^{-\min\{\frac{1}{2} - \frac{1}{p}, \sigma\}})$. Moreover, we investigate corresponding lower bounds ([3]). In particular, this extends the results from [1], [2], [4], and [6], obtained for the randomized Euler scheme.

References

- 1 S. Heinrich, and B. Milla, The randomized complexity of initial value problems, *J. Complexity*, **24**: 77–88 (2008).
- 2 A. Jentzen, and A. Neuenkirch, A random Euler scheme for Carathéodory differential equations, *J. Comput. Appl. Math.*, **224**: 346–359 (2009).
- 3 R. Kruse, and P. Przybyłowicz, Approximation of solutions of SDEs with fractional Sobolev regularity, in preparation
- 4 R. Kruse, and Y. Wu, Error analysis of randomized Runge–Kutta methods for differential equations with time-irregular coefficients, *Comput. Methods Appl. Math.*, **17**: 479–498 (2017).
- 5 P. Przybyłowicz, On arbitrary slow rate of convergence for randomized Euler scheme, in preparation
- 6 P. Przybyłowicz, and P. Morkisz, Strong approximation of solutions of stochastic differential equations with time-irregular coefficients via randomized Euler algorithm, *Appl. Numer. Math.*, **78**: 80–94 (2014).

3.23 Tractability properties of discrepancy

Friedrich Pillichshammer (Johannes Kepler Universität Linz, AT)

License © Creative Commons BY 3.0 Unported license
© Friedrich Pillichshammer

Joint work of Josef Dick, Aicke Hinrichs, Friedrich Pillichshammer

Discrepancies are quantitative measures for the irregularity of distribution of point sets in $[0, 1]^d$ which are closely related to the error of quasi-Monte Carlo (QMC) integration rules. Classical results consider discrepancy with respect to its asymptotic dependence when the

size N of a point set tends to infinity. In this sense optimal results are known, but often these results give no information on the pre-asymptotic scale, especially when the dimension d is large.

In 2001 Heinrich, Novak, Wasilkowski and Woźniakowski [1] initiated the study of the dependence of discrepancy on the dimension d with a remarkable result for the star discrepancy. They showed that for every N and d there exists a N -point set in $[0, 1]^d$ with classical star discrepancy of at most $C\sqrt{d/N}$, where C is a positive constant independent of N and d . Since then a lot of papers on this topic with exciting results have appeared. Nevertheless, a lot of problems are still open.


In this talk we give a review of this topic and present some new results concerning the periodic L_2 discrepancy and the discrepancy with respect to the exponential Orlicz norm.

References

- 1 S. Heinrich, E. Novak, G.W. Wasilkowski, and H. Woźniakowski: The inverse of the star-discrepancy depends linearly on the dimension. *Acta Arith.* 96: 279-302, 2001.

3.24 Large deviations in geometric functional analysis

Joscha Prochno (Universität Graz, AT)

License  Creative Commons BY 3.0 Unported license

© Joscha Prochno

Joint work of Joscha Prochno, Zakhar Kabluchko, Christoph Thäle

Large deviations are a classical topic in probability theory, but have only recently entered the scene of asymptotic geometric analysis. After giving a short introduction to the theory of large deviations, we present a large deviations principle for the q -norm length of a random vector chosen uniformly at random from the unit ball of ℓ_p^n . More precisely, we show that for $1 \leq p < \infty$ and $q > p$, the sequence $(n^{1/p-1/q}\|X\|_q)_{n \in \mathbb{N}}$ satisfies a large deviations principle with speed $n^{p/q}$ and rate

$$\mathbb{I}(z) = \begin{cases} \frac{1}{p}(z^q - M_p(q))^{p/q} & \text{for } z^q \geq M_p(q), \\ \infty & \text{else.} \end{cases}$$

We shall also mention large deviations results that can be proved in the noncommutative setting of Schatten classes. In this case, the rate function is essentially the logarithmic energy plus some perturbation by a constant strongly connected to the famous Ullman distribution. As a consequence of the Sanov-type large deviations, one obtains a strong law of large numbers showing that the empirical spectral measure converges weakly almost surely to the Ullman distribution.

3.25 Wasserstein contraction and spectral gap of simple slice sampling

Daniel Rudolf (Universität Göttingen, DE)

License © Creative Commons BY 3.0 Unported license
© Daniel Rudolf

Joint work of Daniel Rudolf, Viacheslav Natarovskii, Björn Sprungk

Main reference Viacheslav Natarovskii, Daniel Rudolf, Björn Sprungk: “Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling”, CoRR, Vol abs/1903/03824, 2019.

URL <https://arxiv.org/abs/1903.03824>

We provide results on Wasserstein contraction of simple slice sampling for approximate sampling w.r.t. distributions with log-concave and rotational invariant Lebesgue densities. This leads to an explicit quantitative lower bound of the spectral gap of simple slice sampling. In addition to that this lower bound carries over to more general target distributions depending only on the volume of the (super-)level sets of their unnormalized density.

3.26 Complexity of stochastic integration

Stefan Heinrich (TU Kaiserslautern, DE)

License © Creative Commons BY 3.0 Unported license
© Stefan Heinrich

We study the complexity of stochastic integration with respect to the Wiener sheet measure $\int_{[0,1]^d} f(t) dW_t$ of stochastic functions $f = f(t, \omega)$ with Besov $B_{pp}^r([0,1]^d)$ and Bessel potential $H_p^r([0,1]^d)$ regularity in t . We determine the complexity in the deterministic and randomized setting, which includes finding and analyzing algorithms of optimal order and proving matching lower bounds.

3.27 Potential Theory, inverse Laplacians and new Low(?) -Discrepancy Sequences

Stefan Steinerberger (Yale University – New Haven, US)

License © Creative Commons BY 3.0 Unported license
© Stefan Steinerberger

Main reference Stefan Steinerberger: “A Nonlocal Functional Promoting Low-Discrepancy Point Sets”, CoRR, Vol. abs/1902.00441, 2019.

URL <https://arxiv.org/abs/1902.00441>

Main reference Stefan Steinerberger: “Dynamically Defined Sequences with Small Discrepancy”, CoRR, Vol. abs/1902.03269, 2019.

URL <https://arxiv.org/abs/1902.03269>


Main reference Florian Pausinger: “Greedy energy minimization can count in binary: point charges and the van der Corput sequence”, CoRR, Vol. abs/1905.09641, 2019.

URL <https://arxiv.org/abs/1905.09641>

We discuss a new way to construct sequences in the unit interval with very favorable distribution properties. Our construction is based on a greedy algorithm that uses the Green function of the fractional Laplacian as a kernel; we can prove that the discrepancy of this set is at least $N^{-1/2} \log(N)$ but presumably much stronger results hold true (and this is also backed up by numerical investigations). This seems to open several different lines of research.

3.28 Dimension-independent convergence of Gaussian process regression


Aretha Teckentrup (University of Edinburgh, GB)

License  Creative Commons BY 3.0 Unported license
 © Aretha Teckentrup

We consider the problem of interpolating a function $f : [0, 1]^s \rightarrow \mathbb{R}$, where the input dimension s is potentially large. In particular, we study kernel based meshless methods such as kernel based interpolants and Gaussian process emulators. Using results from high-dimensional quadrature, we prove error estimates that are independent of s . The errors are measured in the L^2 -norm or the supremum norm.

3.29 Sampling discretization error of integral norms for function classes


Vladimir N. Temlyakov (University of South Carolina – Columbia, US)

License  Creative Commons BY 3.0 Unported license
 © Vladimir N. Temlyakov

The new ingredient of this paper is that we consider infinitely dimensional classes of functions and instead of the relative error setting, which was used in previous papers on norm discretization, we consider the absolute error setting. We demonstrate how known results from two areas of research – supervised learning theory and numerical integration – can be used in sampling discretization of the square norm on different function classes.

3.30 Discrepancy, Dispersion and Fixed Volume Discrepancy

Mario Ullrich (Johannes Kepler Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
 © Mario Ullrich

Joint work of Vladimir N. Temlyakov, Mario Ullrich

Main reference Vladimir N. Temlyakov, Mario Ullrich: “On the fixed volume discrepancy of the Fibonacci sets in the integral norms”, CoRR, Vol. abs/1908.04658, 2019.

URL <http://arxiv.org/abs/1908.04658>

We present a bunch of recent results on the discrepancy and dispersion, especially in high dimensions, and give an introduction to a new geometric quantity – the fixed volume discrepancy. One of the implications that can be obtained from this new quantity can be stated like this: “Bad boxes” for the discrepancy cannot be “too small”.

3.31 Approximation of shallow neural networks

Jan Vybíral (Czech Technical University – Prague, CZ)

License © Creative Commons BY 3.0 Unported license
© Jan Vybíral

Joint work of Jan Vybíral, Massimo Fornasier, Ingrid Daubechies, Karin Schnass, Tino Ullrich, Sebastian Mayer
Main reference Massimo Fornasier, Jan Vybíral, Ingrid Daubechies: “Identification of Shallow Neural Networks by Fewest Samples”, CoRR, Vol. abs/1804.01592, 2018.
URL <http://arxiv.org/abs/1804.01592>

We address the structure identification and the uniform approximation of sums of ridge functions $f(x) = \sum_{i=1}^m g_i(a_i \cdot x)$ on \mathbb{R}^d , representing a general form of a shallow feed-forward neural network, from a small number of query samples. Higher order differentiation, as used in our constructive approximations, of sums of ridge functions or of their compositions, as in deeper neural network, yields a natural connection between neural network weight identification and tensor product decomposition identification. We prove that in the case of the shallowest feed-forward neural network, second order differentiation and tensors of order two (i.e., matrices) suffice. Based on multiple gathered approximated first and second order differentials, our general approximation strategy is developed as a sequence of algorithms to perform individual sub-tasks. We first perform an active subspace search by approximating the span of the weight vectors a_1, \dots, a_m . Then we use a straightforward substitution, which reduces the dimensionality of the problem from d to m . The core of the construction is then the stable and efficient approximation of weights expressed in terms of rank-1 matrices $a_i \otimes a_i$, realized by formulating their individual identification as a suitable nonlinear program. We prove the successful identification by this program of weight vectors being close to orthonormal and we also show how we can constructively reduce to this case by a whitening procedure, without loss of any generality.

3.32 Randomized Smolyak Algorithm: Explicit Cost Bounds and an Application to Infinite-Dimensional Integration

Marcin Wnuk (Universität Kiel, DE)

License © Creative Commons BY 3.0 Unported license
© Marcin Wnuk

Joint work of Michael Gnewuch, Marcin Wnuk
Main reference Michael Gnewuch, Marcin Wnuk: “Explicit error bounds for randomized Smolyak algorithms and an application to infinite-dimensional integration”, CoRR, Vol. abs/1903/02276, 2019.
URL <https://arxiv.org/abs/1903.02276>

The Smolyak method is a generic tool to tackle tensor product problems. Let $d \in \mathbb{N}$. Generally, for $n = 1, \dots, d$, separable Hilbert spaces of functions $F^{(n)}$, separable Hilbert spaces $G^{(n)}$, and bounded linear operators $S^{(n)} : F^{(n)} \rightarrow G^{(n)}$ are given. The tensor product problem is defined by the solution operator $S_d = \bigotimes_{n=1}^d S^{(n)}$, so for $F_d = \bigotimes_{n=1}^d F^{(n)}$ and $G_d = \bigotimes_{n=1}^d G^{(n)}$ we have

$$S_d : F_d \rightarrow G_d.$$

Suppose that for every $n = 1, \dots, d$, one has a sequence of algorithms $(U_j^{(n)})_{j \in \mathbb{N}}$ meant to approximate $S^{(n)}$. Usually, with growing j the algorithms $U_j^{(n)}$ give better approximation, but are at the same time more expensive. The algorithms $(U_j^{(n)})$ are often referred to as building

blocks. Define the algorithm differences $\Delta_j^{(n)} = U_j^{(n)} - U_{j-1}^{(n)}, j \geq 2$ and $\Delta_1^{(n)} = U_1^{(n)}$. The d -variate Smolyak method of level $L \geq d$ is now given by

$$A(L, d) = \sum_{j \in Q(L, d)} \bigotimes_{n=1}^d \Delta_{j_n}^{(n)},$$

where $Q(L, d) = \{j = (j_1, \dots, j_d) \in \mathbb{N}^d \mid \sum_n j_n \leq L\}$. One speaks of a randomized Smolyak method if the $(U_j^{(n)})_{n,j}$ are randomized algorithms.

In this talk, as an error criterion we consider the randomized error given by

$$e^{ran}(S_d, A(L, d)) = \left[\sup_{f \in F_d, \|f\|_{F_d}=1} \mathbb{E} \| (S_d - A(L, d))f \|_{G_d}^2 \right]^{\frac{1}{2}},$$

and show under some regularity conditions on the building blocks that if for every $n = 1, \dots, d$, the convergence rate of $(U_j^{(n)})_j$ is of the order $\mathcal{O}(\frac{1}{N^\alpha})$ then one also has for some positive constants C_0, C_1 , not depending on N nor d

$$e^{ran}(A(L, d), S_d) \leq C_0 C_1^d \left(1 + \frac{\log(N)}{d-1}\right)^{(d-1)(\alpha+1)} N^{-\alpha}, \quad d \geq 2, \quad (3)$$

Here N denotes the cardinality of information used by the respective algorithms.

Our interest in the Smolyak method is twofold. Firstly, the upper bound (3) shows that the Smolyak method is quite an efficient generic tool to tackle tensor product problems in moderate dimension d . Secondly, even if d is very large, Smolyak method may be used as a building block of more complicated algorithms.

We illustrate the second statement with an example of infinite-dimensional integration on weighted function spaces. There one considers input from some Hilbert space

$$H = \bigotimes_{n=1}^{\infty} H_n,$$

where for each $n \in \mathbb{N}$, H_n is a reproducing kernel Hilbert space. Moreover, with growing n the spaces H_n are assigned decreasing weights, meaning basically that even though one is considering as input functions of infinitely many variables, the impact of variables from higher coordinates gets smaller and smaller. Under some technical assumptions in this setting one may define in a sensible way the integral of functions from H . The problem is now to approximate the integral with the help of randomized algorithms. It turns out that our bound (3) in combination with the results of Plaskota and Wasilkowski on multivariate decomposition methods (MDMs) [3] and the embedding results of Gnewuch, Hefter, Hinrichs and Ritter [1] allows us to show that multivariate decomposition methods using our randomized Smolyak algorithms as building blocks achieve optimal convergence rate.


Details and further results may be found in [2].

References

- 1 Gnewuch Michael, Hefter Mario, Hinrichs Aicke, Ritter Klaus, 'Embeddings of weighted Hilbert spaces and applications to multivariate and infinite-dimensional integration', Journal of Approximation Theory, Volume 222, 2017
- 2 Gnewuch Michael, Wnuk Marcin 'Explicit error bounds for randomized Smolyak algorithms and application to infinite-dimensional integration', Preprint, ArXiv 1903.02276, 2019
- 3 Plaskota Leszek, Wasilkowski Grzegorz, 'Tractability of infinite-dimensional integration in the worst case and randomized settings', Journal of Complexity, Volume 27, 2011

3.33 Tractability for Volterra problems with convolution kernels

Henryk Wozniakowski (Columbia University – New York, US)

License  Creative Commons BY 3.0 Unported license
© Henryk Wozniakowski

We show that the information complexity of the Volterra problems considered in this talk is the same (essentially) as the information complexity of multivariate approximation. Therefore the Volterra problems enjoy the same notions of tractability.

We also analyze the combinatory cost of Picard's algorithm for the Volterra problems. The bounds we obtain are not necessarily optimal.

3.34 Approximation complexity for additive random fields

Marguerite Zani (Université d'Orléans, FR)

License  Creative Commons BY 3.0 Unported license
© Marguerite Zani
Joint work of Marguerite Zani, Alexey Khartov, Mikhail A. Lifshits

We study the approximation complexity of additive random fields. For example for $Y_d(t) = \sum_{j=1}^d X_j(t)$ where the X_j are uncorrelated, square-integrable, centered random processes of dimension 1.

The complexity $n^{Y_d}(\varepsilon)$ in the average case setting is considered here. We give asymptotics for $\log n^{Y_d}(\varepsilon)$ when ε is fixed and d goes to infinity. We give results in case constant 1 is an eigenfunction of the covariance function associated to Y_d , and if not we can boil down to this situation.

Participants

- Dmitriy Bilyk
University of Minnesota –
Minneapolis, US
- James M. Calvin
NJIT – Newark, US
- Ronald Cools
KU Leuven, BE
- Sonja Cox
University of Amsterdam, NL
- Steffen Dereich
Universität Münster, DE
- Benjamin Doerr
Ecole Polytechnique –
Palaiseau, FR
- Martin Ehler
Universität Wien, AT
- Michael Giles
University of Oxford, GB
- Michael Gnewuch
Universität Osnabrück, DE
- Takashi Goda
University of Tokyo, JP
- Mario Hefter
TU Kaiserslautern, DE
- Stefan Heinrich
TU Kaiserslautern, DE
- Aicke Hinrichs
Johannes Kepler Universität
Linz, AT
- Martin Hutzenhaler
Universität Duisburg-Essen, DE
- Lutz Kämmerer
TU Chemnitz, DE
- Alexander Keller
NVIDIA, DE
- Kristin Kirchner
ETH Zürich, CH
- David Krieg
Johannes Kepler Universität
Linz, AT
- Peter Kritzer
österreichische Akademie der
Wissenschaften – Linz, AT
- Robert J. Kunsch
RWTH Aachen, DE
- Frances Y. Kuo
UNSW Sydney, AU
- Gunther Leobacher
Universität Graz, AT
- Thomas Müller-Gronbach
Universität Passau, DE
- Erich Novak
Universität Jena, DE
- Dirk Nuyens
KU Leuven, BE
- Friedrich Pillichshammer
Johannes Kepler Universität
Linz, AT
- Leszek Plaskota
University of Warsaw, PL
- Joscha Prochno
Universität Graz, AT
- Pawel Przybylowicz
AGH Univ. of Science &
Technology-Krakow, PL
- Klaus Ritter
TU Kaiserslautern, DE
- Daniel Rudolf
Universität Göttingen, DE
- Stefan Steinerberger
Yale University – New Haven, US
- Michaela Szölgényi
Alpen-Adria-Universität
Klagenfurt, AT
- Aretha Teckentrup
University of Edinburgh, GB
- Vladimir N. Temlyakov
University of South Carolina –
Columbia, US
- Mario Ullrich
Johannes Kepler Universität
Linz, AT
- Jan Vybíral
Czech Technical University –
Prague, CZ
- Marcin Wnuk
Universität Kiel, DE
- Henryk Wozniakowski
Columbia University –
New York, US
- Larisa Yaroslavtseva
Universität Passau, DE
- Marguerite Zani
Université d’Orléans, FR



Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction

Edited by

Rolf Backofen¹, Yael Mandel-Gutfreund², Uwe Ohler³, and Gabriele Varani⁴

1 Universität Freiburg, DE, backofen@informatik.uni-freiburg.de

2 Technion – Haifa, IL, yaelm@technion.ac.il

3 Max-Delbrück-Centrum – Berlin, DE, uwe.ohler@mdc-berlin.de

4 University of Washington – Seattle, US, varani@chem.washington.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19342 “Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction”.

Seminar August 18–23, 2019 – <http://www.dagstuhl.de/19342>

2012 ACM Subject Classification Mathematics of computing → Probability and statistics, Theory of computation → Design and analysis of algorithms, Theory of computation → Theory and algorithms for application domains, Applied computing → Life and medical sciences, Applied computing → Chemistry, Applied computing → Mathematics and statistics

Keywords and phrases Machine learning, algorithms, genomics analysis, gene expression networks, big data analysis, quantitative prediction, proteins, RNA, CLIP-Seq

Digital Object Identifier 10.4230/DagRep.9.8.49

Edited in cooperation with Florian Heyl, Michael Uhl

1 Executive Summary

Rolf Backofen

Yael Mandel-Gutfreund

Uwe Ohler

Gabriele Varani

License © Creative Commons BY 3.0 Unported license
© Rolf Backofen, Yael Mandel-Gutfreund, Uwe Ohler, and Gabriele Varani

DNA is often described as the blueprint of life, since it encodes all the information necessary for an organism to develop and maintain its biological functions. Single blueprints for specific functions are stored inside DNA regions called genes. The primary product produced (also termed expressed) from genes is RNA, which can either become biologically active itself (non-coding RNA or ncRNA) or is further translated into proteins (messenger RNA or mRNA), which then executes the gene functions. Given the astonishing complexity of biological functions, it is not surprising that the regulation of gene expression itself is a highly complex matter. Proteins, RNA, and DNA all can interact with each other, forming regulatory networks in order to control the expression of genes. To elucidate these networks, experimental scientists rely more and more on high-throughput methods, producing vast amounts of raw data. Computational methods to analyze these huge datasets are therefore of highest demand. The main focus of this seminar lies on RNA-protein and RNA-RNA



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

Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 49–69

Editors: Rolf Backofen, Yael Mandel-Gutfreund, Uwe Ohler, and Gabriele Varani



DAGSTUHL Dagstuhl Reports

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

interactions. In particular, transcriptome-wide binding patterns of RNA-binding proteins (RBPs), their computational predictability, and the biological effects of binding are discussed. Moreover, the seminar dealt with topics like combinatorial RBP binding prediction, RBP binding kinetics, RNA-RNA interaction prediction, subcellular RNA imaging, and RBP binding site classification. Regarding the computational methodology, several newly developed deep learning methods are presented, e.g. for RBP binding site prediction. Taken together, the aim of the seminar is to bring experimental and computational scientists together for the aforementioned topics and to engage them in fruitful discussions in order to:

- present the current experimental and computational methodologies,
- understand their implications, strengths, and limitations from first-hand experience,
- and spark ideas for developing new computational and experimental methods and improving on existing ones.

2 Table of Contents

Executive Summary

Rolf Backofen, Yael Mandel-Gutfreund, Uwe Ohler, and Gabriele Varani 49

Introduction

Seminar Format 53

Studying protein-RNA interactions 53

Functional analysis of RBPs 54

Study of non-coding RNAs 54

Improvement of CLIP-seq Data 54

Provision of Tools and Data for and of protein-RNA experiments 54

Conclusions 55

Overview of Talks

How to make Sense out of CLIP-seq data
Rolf Backofen 55

The kinetic landscape of Dazl-RNA binding in cells
Eckhard Jankowsky 55

RNA structure as mediator of cooperative/antagonistic RBP interaction
Jörg Fallmann 56

Associating non-coding RNAs to proteins: from RNA structure to literature mining
Jan Gorodkin 56

Evaluation and Classification of Peak Profiles for Protein-RNA Binding Predictions
Florian Heyl and Rolf Backofen 57

Computational Approaches to Posttranscriptional Gene Regulation in Human
Biology and Disease
Katharina Zarnack 57

A Translational Repression Complex in Developing Mammalian Neural Stem Cells
that Regulates Neuronal Specification
Kazan Hilal 58

in vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on
regulation by co-factors
Julian König 58

Posttranscriptional regulation in cellular and time
Markus Landthaler 59

Deconstructing an essential RNA regulatory program
Donny Licatalosi 59

RNA-mediated transcriptional regulation: a systematic search for new players
Yael Mandel-Gutfreund 60

A new method to predict novel trans RNA-RNA interactions
Irmtraud Meyer 60

Characterizing the snoRNome through transcriptomics profiling and RNA-RNA interactomics <i>Michelle Scott</i>	61
RNA regulatory dynamics controlling human steroidogenesis <i>Neelanjan Mukherjee</i>	62
Deep learning for protein-RNA interactions <i>Yaron Orenstein</i>	63
Dynamic post-transcriptional RNA regulation in early zebrafish development <i>Michal Rabani</i>	63
Exploring inter-domain cooperation in RNA binding proteins <i>Andres Ramos</i>	64
Coding regions regulate mRNA stabilities in human cells <i>Olivia Rissland</i>	64
Eukaryotic-wide reconstruction of RNA-binding protein specificity by joint matrix factorization <i>Alexander Sasse</i>	65
Decoding regulatory protein-RNA interactions by combining integrative structural biology and large-scale approaches <i>Michael Sattler</i>	66
GraphProt2: deep learning for graphs meets RBP binding site prediction <i>Michael Uhl</i>	66
Breaking apart 3'-UTRs to model in vivo post-transcriptional regulation <i>Charles E. Vejnar</i>	67
Deep Learning for Modeling Translation events <i>Jianyang Zeng</i>	68
Participants	69

3 Introduction

3.1 Seminar Format

The seminar “Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction” emerged from the organizer’s combined experience that new experimental and computational methods to understand RNA-based gene regulation are positively influenced and more prolific if both sides exchange their ideas, findings, and hypotheses in order to answer critical biological, biomedical, and bioinformatical questions. Experimental and computational biology are two big fields that require different expertise, which are even further divided into sub-fields such as structural and chemical biology. To foster the research of highly interdisciplinary fields, such as the study of protein-RNA interactions, it is imperative that all sides talk and discuss to understand in-depth the underlying problems and goals to find concrete paths. The biggest opportunities are collaborative analyses combining newly and different computational and experimental methods, but events and venues that facilitate an open platform to form such collaborations and opportunities are missing.

The Dagstuhl seminar brought experimental and computation biologists together, allowing them to present and discuss their findings and newly developed powerful computational and experimental methods to investigate RNA-protein interactions across the genome and transcriptome of different organisms, cell tissues and cells. Each day and each session consisted of a different higher-order topic that combined several presentations of computational and experimental groups to intertwine both sides and flourish vivid discussions. Each session was intermittent by a discussion round to talk about open questions, ideas, and problems for the current topic. These discussions also catalyzed the birth of new and enhanced technologies to improve the analysis of protein-RNA interactions. The seminar was attended by many leading scientists in their field of expertise from all around the world to find solutions and ideas of ongoing problems.

3.2 Studying protein-RNA interactions

In order to understand the complexity of post-transcriptional regulation by RBPs, it is essential to have experimental methods for detecting RBP binding sites on RNAs with high resolution. In this context, CLIP-seq (cross-linking and immunoprecipitation followed by next generation sequencing) together with its popular variants PAR-CLIP, iCLIP, and eCLIP has become the state-of-the-art procedure for determining transcriptome-wide binding sites of RBPs with single-nucleotide resolution.

The seminar saw both the presentation of a new iCLIP variant as well several prediction methods that utilize CLIP-seq data to train models for predicting new binding sites for RBPs of interest. The protocol called *in vitro* iCLIP makes it possible to identify RBP binding sites *in vitro* for selected RNAs, which when also taking into account *in vivo* iCLIP data allows for estimating the effects of trans-acting RBPs on the binding patterns of the studied RBP. The shown prediction methods offer various deep learning approaches to RBP binding site prediction, using, for example, convolutional neural networks (CNNs) or graph convolutional neural networks (GCNs) to learn features from RBP binding sites determined by RNAcompete or CLIP-seq methods.

3.3 Functional analysis of RBPs

Several studies on newly described functions of RBPs or the discovery of RBPs as contributors to certain cellular functions have been presented. Among these were, for example, reports on the ZFP36 family of RBPs with potential functions in steroidogenesis and hypertension, roles of the DAZL RBP in gametogenesis, or a translational repression complex made up of PUM2 and 4E-T active in neuronal specification. Computational approaches have also been shown, for example, the prediction of binding preferences of an RBP based on its protein sequence, which could be used to roughly classify the RBP based on RBPs with known functional roles that share similar binding motifs.

3.4 Study of non-coding RNAs

It is currently estimated that the number of non-coding RNA genes is higher than the amount of protein-coding genes in the human genome. On the other hand, the vast majority of these non-coding RNAs have no assigned functions yet, urging the need for functional studies. In particular, long non-coding RNAs (lncRNAs) have drawn much attention in recent years due to their diverse cellular roles, such as RNA-DNA triple helix formation, or in general DNA interactions to control gene expression. In this context, a work on lncRNAs that bind to transcription factors (TFs) has been presented, linking lncRNA NORAD with TF STAT3 in human stem cells. A second work presented a computational prediction method based on RNA structure alignment and structure-based clustering to identify novel ncRNA-RNA or ncRNA-protein interactions in the human genome. In a third presentation, new functional roles for small nucleolar RNAs (snoRNAs) have been discussed, and a machine learning approach was shown to predict known and new RNAs targeted by snoRNAs.

3.5 Improvement of CLIP-seq Data

The genome-wide approaches reported in Session 1 (CLIP-seq) determine *in vivo* interactions. An interesting discussion evolved about the noise of CLIP-seq experiments. New approaches were discussed and presented to reduce the noise of CLIP-seq experiments, including a better normalization, improved CLIP-seq protocols, and new methods to improve the peak calling quality, such as, improved peak calling algorithms and new methods to reduce the number of false positives and false negatives.

3.6 Provision of Tools and Data for and of protein-RNA experiments

The last vivid discussion session picked up at aforementioned topics about a better contribution and provision of bioinformatical tools for the analysis of protein-RNA experiments, and obtained results of completed analysis of protein-RNA data. A big part of this discussion was Galaxy as an interactive platform for bioinformatical tools. The participants uttered their wishes and ideas about new tools and training materials for Galaxy, such as, IntaRNA, or omniCLIP. Furthermore, the participants discussed about an integrative browser for CLIP-seq results for a quick visualization of the binding sites of different RBPs to improve the investigation of common sequence and structural motifs.

3.7 Conclusions

The third Dagstuhl seminar achieved its goal to bring computational and experimental people together to exchange and discuss the advances and challenges in protein-RNA recognition, regulation and prediction, which is a very interdisciplinary field and thus challenging from both the experimental and computational standpoint. There are rare occasions and venues that facilitate a platform for open discussions and presentations between experimental and computational people to foster the research of protein-RNA interactions. Consequently, the third Dagstuhl seminar was another successful bridge, which was well appreciated by all participants and gave birth to new and exciting collaborations, ideas, and relationships that would not have been formed without this meeting.

4 Overview of Talks

4.1 How to make Sense out of CLIP-seq data

Rolf Backofen (Universität Freiburg, DE)

License © Creative Commons BY 3.0 Unported license
© Rolf Backofen

It is becoming increasingly clear that RNA-binding proteins are key elements in regulating the cell's transcriptome. CLIP-seq is one of the major tools to determine binding sites but suffers from high false negative rate due its expression dependency. This critical hinders the use of public CLIP-data. We will show in several examples how use of raw public CLIP-seq data can lead to false biological reasoning and how advanced machine learning approach can overcome this problem. I will also introduce some recent application of approach that allows us to determine mechanism underlying post-transcriptional regulation. I will further discuss our results from our new Nature paper, showing that the human RNA helicase DHX9 predominantly binds to IRAlu elements and such suppresses the negative effect of Alu inflation in transcripts.

4.2 The kinetic landscape of Dazl-RNA binding in cells

Eckhard Jankowsky

License © Creative Commons BY 3.0 Unported license
© Eckhard Jankowsky

The kinetics by which RNA binding proteins (RBPs) interact with their cellular RNA sites are thought to be critical for the biological function of RBPs, but it has not been possible to measure these kinetic parameters in cells. Here, we describe a new approach to determine kinetic parameters of protein binding to individual RNA sites in cells and show how kinetic data quantitatively link RNA binding patterns to biological RBP function.

We combine time-resolved, multi-photon RNA-protein crosslinking with Immunoprecipitation, Next Generation Sequencing, and large scale kinetic modeling to determine rate constants for association, dissociation, crosslinking and fractional occupancy for thousands of individual binding sites of the RBP Dazl in mouse GC1 cells. Association and dissociation rate constants for Dazl vary by several orders of magnitude among different binding sites.

Dazl resides at individual binding sites at most for few seconds or less, indicating exceptionally high dynamics of Dazl-RNA binding. We further find that the presence of only few Dazl proteins on a given RNA per time interval ultimately determines the impact of Dazl on translation and decay of a given RNA. Dazl presence on an RNA is controlled in a complex fashion through the binding kinetics at individual binding sites, the collective kinetics of Dazl clusters and the combination of these clusters on a given RNA. The data explain how similar Dazl effects on translation can be accomplished by distinct Dazl-RNA binding patterns. Collectively, our results show that and how previously inaccessible, kinetic parameters for RNA-protein interactions in cells allow the development of detailed mechanistic models for cellular RNA-protein interactions.

4.3 RNA structure as mediator of cooperative/antagonistic RBP interaction

Jörg Fallmann (Universität Leipzig, DE)

License  Creative Commons BY 3.0 Unported license
© Jörg Fallmann

Joint work of Jörg Fallmann, Julian König, Katharina Zarnack

RNA is a molecule known for its ability to form stable secondary structures on an inter- and intramolecular level. Such structures can influence the binding site of an RBP. The probability for the formation of basepairs on intramolecular level can be expressed as the accessibility of a stretch of RNA. This accessibility directly influences the availability of binding sites for an interacting molecule. If one RBP binds its target site, this can also lead to changes in the conformation of the RNA molecule, a feature we can model *in silico* via constraint folding. This approach is used to predict positive and negative effects of RBP interactions on the accessibility of the binding sites for other RBPs or RNAs. From changes in accessibility we can via the calculation of pseudo energy contributions derive a measure for the affinity of the corresponding interaction partner. With this we want to model and infer potential antagonistic or cooperative behavior of RBP pairs or other pairs of interaction partners.

4.4 Associating non-coding RNAs to proteins: from RNA structure to literature mining

Jan Gorodkin (University of Copenhagen, DK)

License  Creative Commons BY 3.0 Unported license
© Jan Gorodkin

In the first strategy, we performed a genome-wide prediction of Conserved RNA Structures (CRSs) using the syntenic regions that have been subjected to RNA structural alignment in the mammalian genomes. The downstream integrative analysis of the 774 K predicted CRSs showed that the resulting CRSs are enriched to overlap protein binding sites from CLiP data. Using RNA structure based clustering approach we cluster the CRSs from UTRs and identify thousands of putative CRS clusters that contained at least two CRSs located at different genomic positions. Upon filtering these clusters for consistent overlap to the same protein, we identify a few hundred clusters associating with proteins. In the second strategy, we employ

search for non-coding RNA (ncRNA) – RNA interactions and ncRNA –protein interaction in a similar fashion as done in the STRING database for protein-protein interactions. We use an integrative scoring scheme to obtain confidence scores from four channels: curated examples, experimental data, interaction predictions and automatic literature mining. To evaluate the method we show for the largest class of ncRNAs, microRNAs, that the combined scoring scheme outperform that of individual microRNA target predictors. The obtained interactions link directly into STRINGs payload mechanisms and hence allowing uses to fuse the ncRNA interactions with protein networks.

4.5 Evaluation and Classification of Peak Profiles for Protein-RNA Binding Predictions

Florian Heyl (Universität Freiburg, DE) and Rolf Backofen (Universität Freiburg, DE)

License © Creative Commons BY 3.0 Unported license
© Florian Heyl and Rolf Backofen

Main reference Florian Heyl, Rolf Backofen: “StoatyDive: Evaluation and Classification of Peak Profiles for Sequencing Data”, bioRxiv, Cold Spring Harbor Laboratory, 2019.

URL <https://doi.org/10.1101/799114>

The prediction of binding sites (peak calling) is a common task in the data analysis of methods such as crosslinking immunoprecipitation in combination with high-throughput sequencing (CLIP-Seq). The peaks are often further analyzed to predict sequence motifs or structure patterns. However, the obtained peak set can vary in their profile shapes because of the used peakcaller method, different binding domains of the protein, protocol biases, or other factors. Thus, a prior step is missing to evaluate and classifies the predicted peaks based on their shapes. We investigated different shapes in CLIP data and pronounce a filter step to distinguish different peak shapes and thus improve subsequent analysis tasks.

4.6 Computational Approaches to Posttranscriptional Gene Regulation in Human Biology and Disease

Katharina Zarnack (Goethe-Universität – Frankfurt am Main, DE)

License © Creative Commons BY 3.0 Unported license
© Katharina Zarnack

Main reference Simon Braun, Mihaela Enculescu, Samarth T. Setty, Mariela Cortés-López, Bernardo P. de Almeida, F. X. Reymond Sutandy, Laura Schulz, Anke Busch, Markus Seiler, Stefanie Ebersberger, Nuno L. Barbosa-Morais, Stefan Legewie, Julian König, Kathi Zarnack: “Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis”. *Nat Commun* 9, 3315, 2018.

URL <https://doi.org/10.1038/s41467-018-05748-7>

We employ a systems approach to better understand how multiple protein complexes dynamically interact on pre-mRNA sequence to control splicing (splice code). As a prototypical example, we study the alternative splicing of the MSTR1 gene which is frequently altered in cancer. Starting with a high-throughput mutagenesis screen, the complex splicing patterns are interpreted using mathematical models to infer changes in the splicing kinetics and to identify causative mutations. Importantly, the measured effects correlate with RON alternative splicing in cancer patients bearing the same mutations. Moreover, they highlight the RNA-binding protein HNRNPH as a key regulator of RON splicing in healthy tissues and cancer. Our results thereby offer insights into splicing regulation and the impact of mutations on alternative splicing in cancer.

4.7 A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification

Kazan Hilal (Antalya International University, TR)

License © Creative Commons BY 3.0 Unported license
© Kazan Hilal

Joint work of Siraj K. Zahr, Guang Yang, Hilal Kazan, Michael J. Borrett, Scott A. Yuzwa, Anastassia Voronova, David R. Kaplan, Freda D. Miller

Main reference Siraj K. Zahr, Guang Yang, Hilal Kazan, Michael J. Borrett, Scott A. Yuzwa, Anastassia Voronova, David R. Kaplan, Freda D. Miller: “A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification”, *Neuron*, Vol. 97(3), pp. 520–537.e6, 2018.

URL <http://dx.doi.org/10.1016/j.neuron.2017.12.045>

The mechanisms instructing genesis of neuronal sub- types from mammalian neural precursors are not well understood. To address this issue, we have characterized the transcriptional landscape of radial glial precursors (RPs) in the embryonic murine cortex. We show that individual RPs express mRNA, but not protein, for transcriptional specifiers of both deep and superficial layer cortical neurons. Some of these mRNAs, including the superficial versus deep layer neuron transcriptional regulators *Brn1* and *Tle4*, are translationally repressed by their association with the RNA- binding protein Pumilio2 (*Pum2*) and the 4E-T protein. Disruption of these repressive complexes in RPs mid-neurogenesis by knocking down 4E-T or *Pum2* causes aberrant co-expression of deep layer neuron specification proteins in newborn superficial layer neurons. Thus, cortical RPs are transcriptionally primed to generate diverse types of neurons, and a *Pum2*/4E-T complex represses translation of some of these neuronal identity mRNAs to ensure appropriate temporal specification of daughter neurons.

4.8 in vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by co-factors

Julian König (Institut für Molekulare Biologie – Mainz, DE)

License © Creative Commons BY 3.0 Unported license
© Julian König

Main reference F.X. Reymond Sutandy, Stefanie Ebersberger, Lu Huang, Anke Busch, Maximilian Bach, Hyun-Seo Kang, Jörg Fallmann, Daniel Maticzka, Rolf Backofen, Peter F. Stadler, Kathi Zarnack, Michael Sattler, Stefan Legewie, Julian König: “In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res* 28(5), 699–713, 2018.

URL <https://doi.org/10.1101/gr.229757.117>

Alternative splicing generates distinct mRNA isoforms and is crucial for proteome diversity in eukaryotes. The RNA-binding protein (RBP) U2AF2 is central to splicing decisions, as it recognizes 3' splice sites and recruits the spliceosome. We establish 'in vitro iCLIP' experiments, in which recombinant RBPs are incubated with long transcripts, to study how U2AF2 recognizes RNA sequences and how this is modulated by trans-acting RBPs. We measure U2AF2 affinities at hundreds of binding sites, and compare in vitro and in vivo binding landscapes by mathematical modeling. We find that trans-acting RBPs extensively regulate U2AF2 binding in vivo, including enhanced recruitment to 3' splice sites and clearance of introns. Using machine learning, we identify and experimentally validate novel trans-acting RBPs (including FUBP1, BRUNOL6 and PCBP1) that modulate U2AF2 binding and affect splicing outcomes. Our study offers a blueprint for the high-throughput characterization of in vitro mRNP assembly and in vivo splicing regulation.

4.9 Posttranscriptional regulation in cellular and time

Markus Landthaler (*Max-Delbrück-Centrum – Berlin, DE*)

License © Creative Commons BY 3.0 Unported license
© Markus Landthaler

Spatial compartmentalization of RNA is central to many biological processes and enables diverse regulatory schemes that exploit both coding as well as noncoding functions of the transcriptome. Spatiotemporal RNA dynamics are typically examined by single molecule imaging techniques, but can simultaneously only be applied to a small number of transcripts. The combination of metabolic RNA labeling with biochemical nucleoside transitions adds a broadly applicable temporal dimension to RNA sequencing. To obtain insights in the spatiotemporal mRNA distribution in mouse embryonic stem cells we are using SLAM-seq (in collaboration with the labs of Stefan Ameres and Nils Blüthgen), a method for time-resolved measurement of newly synthesized and pre-existing RNA in cultured cells, in combination with cellular fraction and biochemical isolations. Current efforts are aiming at measuring transcriptome-wide kinetics of mRNA export from the nucleus, association with membranes and ribosomes. The goal is to identify sequence and/or structure features that modulate the spatiotemporal distribution of mRNAs.

4.10 Deconstructing an essential RNA regulatory program

Donny Licatalosi (*Case Western Reserve University – Cleveland, US*)

License © Creative Commons BY 3.0 Unported license
© Donny Licatalosi
Joint work of Leah L. Zagore, Molly M. Hannigan, Sebastian M. Weyn-Vanhentenryck, Raul Jobava, Maria Hatzoglou, Chaolin Zhang, Donny D. Licatalosi

The RNA binding protein DAZL is essential for gametogenesis, but its direct *in vivo* functions, RNA targets, and the molecular basis for germ cell loss in *Dazl*-null mice are unknown. Here, we mapped transcriptome-wide DAZL-RNA interactions *in vivo*, revealing DAZL binding to thousands of mRNAs via polyA-proximal 3' UTR interactions. In parallel, fluorescence-activated cell sorting and RNA-seq identified mRNAs sensitive to DAZL deletion in male germ cells. Despite binding a broad set of mRNAs, integrative analyses indicate that DAZL post-transcriptionally controls only a subset of its mRNA targets, namely those corresponding to a network of genes that are critical for germ cell proliferation and survival. In addition, we provide evidence that polyA sequences have key roles in specifying DAZL-RNA interactions across the transcriptome. Our results reveal a mechanism for DAZL-RNA binding and illustrate that DAZL functions as a master regulator of a post-transcriptional mRNA program essential for germ cell survival.

4.11 RNA-mediated transcriptional regulation: a systematic search for new players

Yael Mandel-Gutfreund (Technion – Haifa, IL)

License  Creative Commons BY 3.0 Unported license
© Yael Mandel-Gutfreund

Joint work of Amir Argoetti, Rina Ben-El, Shlomi Dvir, Dor Shalev, Nathan Salomonis, Yael Mandel-Gutfreund

Transcription factors (TFs) play a pivot role in embryonic stem cells as key pluripotent markers. In recent years, it has been shown that long non-coding RNAs (lncRNAs) are involved in activation and repression of pluripotency-related genes via epigenetic and transcriptional regulation. To predict novel interactions between TFs and lncRNAs with regulatory functions in pluripotency and differentiation, we sampled RNA from eleven time points during directed differentiation of human Induced Pluripotent Stem Cells (iPSs) to Cardiomyocytes. Analyzing the differential expression patterns of coding and non-coding RNAs across time revealed pairs of TFs and lncRNAs that are significantly co-expressed, suggesting co-regulatory relationships. To confirm direct interactions between TFs and lncRNAs we performed eCLIP (enhanced crosslinking and immunoprecipitation) followed by sequencing on selected TFs. Computational analysis of the CLIP data revealed a small subset of non-coding RNAs with significantly enriched protein binding peaks. Specifically, the eCLIP results signify a direct association between the STAT3 (signal transducer and activator of transcription 3) TF and the lncRNA NORAD (non-coding RNA activated by DNA damage) in human pluripotent cells. Strikingly, knockdown of NORAD in hESCs significantly impaired STAT3 localization to the nucleus. Based on our findings, we propose that lncRNAs may contribute to stemness by directly interacting with TFs, possibly acting as co-regulators to modulate and fine-tune the transcriptional program of their target genes.

4.12 A new method to predict novel trans RNA-RNA interactions

Irmtraud Meyer (Max-Delbrück-Centrum – Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Irmtraud Meyer

Joint work of Sabine Reißer, Irmtraud Meyer

Many key mechanisms of gene regulation happen on transcriptome level. Key examples include miRNA-mRNA interactions, RNA editing and RNA splicing. These are already known to crucially determine the functional products of any given cell. At the core of these interactions are trans RNA-RNA interactions, i.e. direct interactions between two or more transcripts that may happen at different time points of the transcript's life in the cell. Compared to networks of protein-protein interactions, the universe of trans RNA-RNA interactions remains vastly underexplored. Even the most recent duplex-based experimental methods for probing the RNA interactome and RNA structure in vivo have biases and inefficiencies. On the computational side, there already exist numerous prediction methods. These, however, either cater for specific classes of biological interactions (e.g. miRNA-mRNA) where the key features of the interaction site are already well-known or aim to predict novel classes of trans RNA-RNA interactions while having a range of significant limitations.

To overcome these challenges, we have developed a new computational method that can detect trans RNA-RNA interactions in an unbiased manner provided the corresponding

functional features have been conserved in evolution. Our method takes a given multiple-sequence alignment and a corresponding phylogenetic tree as input and predicts helices (i.e. a helix being defined as a stretch of consecutive base-pairs) that have been well conserved in evolution. Our method employs a fully probabilistic framework that compares for each candidate helix the likelihood of having evolved as base-paired entity to the likelihood of having evolved as unpaired entity. The corresponding log-likelihood scores are derived from two probabilistic models of evolution that model how base-pairs evolve over time and how un-paired nucleotides evolve over time, respectively. Our method is capable of also estimate p-values for all predicted helices. Compared to two existing state-of-the-art programs, the prediction accuracy of our method is significantly higher for a test set of known snoRNA-rRNA interactions from Lai *et al.* [1]. Due to its time- and memory-efficiency, our method readily extends to long biological transcripts which has been one major limitation of existing methods. We thus hope to apply our method on transcriptome-wide data sets in order to identify novel biological classes of trans RNA-RNA interactions.

References

- 1 D. Lai, I. M. Meyer, A comprehensive comparison of general RNA-RNA interaction prediction methods, *Nucleic Acids Research* 44(7):e61 (2016)

4.13 Characterizing the snoRNome through transcriptomics profiling and RNA-RNA interactomics


Michelle Scott (University of Sherbrooke, CA)

License  Creative Commons BY 3.0 Unported license
© Michelle Scott

SnoRNAs have long been characterized for their role as guides for the site-specific modification of rRNA. In recent years however, increasing numbers of studies report diverse novel functions for snoRNAs including the regulation of alternative splicing, the control of the stability of mRNAs and pre-mRNAs, the regulation of chromatin architecture and as essential intermediates in cell stress responses. The characterization of snoRNAs has lagged behind other main RNA families, likely in part due to the difficulty in quantifying them by RNA-seq, because of their strong structure. We have recently established a methodology to accurately measure the abundance of snoRNAs using a reverse transcriptase with low structure bias. Despite the assumed housekeeping role of snoRNAs, expression profiles across various normal human tissues show snoRNAs covering a wide abundance range and a subset displaying tissue specificity or tissue variability, which relates to their host gene and their targets. Our data show that approximately 65% of snoRNAs are uniformly expressed across all tissues considered. Uniformly expressed snoRNAs are typically highly expressed, are often encoded in translation-related protein-coding host genes, target ribosomal RNA and are strongly conserved across evolution. On the other hand, snoRNAs displaying variable expression across tissues are less expressed, less conserved, display characteristics of feedback relationships with their host genes and are more likely to be involved in non-canonical functions in post-transcriptional regulation such as the regulation of alternative splicing. We are also using machine learning approaches to predict canonical and non-canonical targets of snoRNAs and integrating expression profiling and target prediction to characterize the extent of snoRNA functionality.

4.14 RNA regulatory dynamics controlling human steroidogenesis

Neelanjan Mukherjee (University of Colorado – Aurora, US)

License  Creative Commons BY 3.0 Unported license
© Neelanjan Mukherjee

Joint work of Kimberly Wellman, Kent Riemondy, Austin Gillen, Amber Baldwin, Neelanjan Mukherjee

Human steroid hormones produced by the adrenal cortex control important physiology including metabolism, inflammation, blood pressure/volume, and sexual characteristics. Many human disorders are caused by the lack or excess of adrenal hormones. For example, 1 in 20 Americans suffer from high blood pressure caused by excessive adrenal aldosterone production. While the signaling components, transcriptional regulators, and steroidogenic enzymes necessary for production of hormones have been identified, little is known about post-transcriptional regulation of steroidogenesis by RNA-binding proteins (RBPs). Recently technological advances have revolutionized our ability to investigate RBP-driven RNA regulation, making it possible for the first time to investigate how these mechanisms contribute to steroidogenesis. We have recently carried out a screen for RBPs regulating human aldosterone production that revealed numerous RBPs. Many of these RBPs are regulators of cytoplasmic RNA stability and translation.

Prominent among the hits in our screen were members of the ZFP36 family of RNA-binding protein. These RBPs binds to AU-rich elements (ARE) in 3'UTRs and consequently destabilizes and/or translationally represses these ARE-containing mRNAs. Through time-course experiments we found that mRNA stability controls the temporal pattern of RNA expression during steroidogenesis. Indeed, mRNAs with AU-rich elements (AREs) in their 3' UTR were rapidly induced and cleared out in response to steroidogenic stimulation. Furthermore, depletion of either ZFP36L2 or ZFP36L1 significantly increased aldosterone levels. These represents one of the first RBPs implicated in control of human aldosterone synthesis. Notably, over-production of aldosterone is a major cause of hypertension, suggesting that failure of this negative feedback loop could have important implications for human health. Additionally, genome-wide association studies have reported variants in both ZFP36L1 and ZFP36L2 associated with changes in systolic blood pressure.

We propose that the ZFP36 family of proteins operate a negative feedback loop that prevent overproduction of aldosterone by destabilizing and/or translationally repressing ARE-containing mRNAs encoding steroidogenic proteins. Our ongoing research will elucidate the mechanism underlying this negative feedback loop that controls aldosterone biosynthesis post-transcriptionally through the action of ZFP36L1/2 RNA binding. Post-transcriptional regulation of mRNA stability and translation by AREs is a critical gene regulatory pathway important in many different tissues and conditions. Finally, the adrenal cortex is amenable to the delivery of modified oligonucleotides; thus, our discoveries can facilitate the design of oligonucleotide therapeutics that can be used to precisely and specifically modulate steroidogenesis through RBP-RNA disruption.

4.15 Deep learning for protein-RNA interactions

Yaron Orenstein (Ben Gurion University – Beer Sheva, IL)

License © Creative Commons BY 3.0 Unported license
© Yaron Orenstein

Joint work of Ilan Ben-Bassat, Benny Chor, Yaron Orenstein

Main reference Ilan Ben-Bassat, Benny Chor, Yaron Orenstein: “A deep neural network approach for learning intrinsic protein-RNA binding preferences”, *Bioinformatics*, Vol. 34(17), pp. i638–i646, 2018.

URL <http://dx.doi.org/10.1093/bioinformatics/bty600>

Protein-RNA binding, mediated through both RNA sequence and structure, plays vital role in many cellular processes, including neurodegenerative-diseases. Modelling the sequence and structure binding preferences of an RNA-binding protein is a key computational challenge. Accurate models will enable prediction of new interactions and better understanding of the binding mechanism.

DLPRB is a new deep learning based approach to learn RNA sequence and structure binding preferences from large biological datasets. DLPRB outperforms the state of the art, both in vitro and in vivo. Moreover, biological insights can be gained by applying neural networks to large datasets of protein-RNA interactions.

4.16 Dynamic post-transcriptional RNA regulation in early zebrafish development

Michal Rabani (The Hebrew University of Jerusalem, IL)

License © Creative Commons BY 3.0 Unported license
© Michal Rabani

Joint work of Michal Rabani, Lindsey Pieper, Guo-Liang Chew, Alexander F. Schier

Main reference Michal Rabani, Lindsey Pieper, Guo-Liang Chew, Alexander F. Schier: “Massively parallel reporter assay of 3’UTR sequences identifies in vivo rules for mRNA degradation”, *Molecular Cell*, Vol. 68(6), pp. 1083–1094, 2017.

URL <https://doi.org/10.1016/j.molcel.2017.11.014>

The stability of mRNAs is regulated by signals within their sequences, but a systematic and predictive understanding of the underlying sequence rules remains elusive. In this talk, I will introduce UTR-Seq, a combination of massively parallel reporter assays and regression models, to survey the dynamics of tens-of-thousands of 3’UTR sequences during early zebrafish embryogenesis. I will focus on the massive degradation of maternally provided mRNAs, a key developmental transition in early embryos that is shared in all animals, as a powerful system to study mRNA dynamics in the absence of de-novo transcription. Applying UTR-Seq in this system, we revealed two temporal degradation programs: a maternally encoded early-onset program and a late-onset program that accelerated degradation after zygotic genome activation. Our analysis identifies regulatory sequences with specific roles: stabilizing poly-U and UUAG signals and destabilizing GC-rich signals act via early-onset pathways; and miR-430 seeds, ARE signals and PUM sites promote late-onset degradation. These elements identified through UTR-Seq also influence the stability of endogenous maternal mRNAs. Finally, Sequence based regression models translated 3’UTR sequences into their unique decay patterns, and predicted the in vivo impact of sequence signals on mRNA stability. Their application led to the successful design of artificial 3’UTRs that conferred specific mRNA dynamics. By using UTR-Seq as a general strategy to uncover the rules of RNA cis-regulation, we aim to learn the code of genomic information within native maternal mRNAs that defines their unique decay profiles, and its physiological roles during early developmental transitions.

4.17 Exploring inter-domain cooperation in RNA binding proteins

Andres Ramos (University College London, GB)

License © Creative Commons BY 3.0 Unported license
© Andres Ramos

Joint work of Giuseppe Nicastro, Robert Dagil, V. Castilla-Llorente, C. Gallagher, Ian A. Taylor, J. Ule, Andres Ramos

Main reference Robert Dagil, Neil J. Ball, Roksana W. Ogradowicz, Fruzsina Hobor, Andrew G. Purkiss, Geoff Kelly, Stephen R. Martin, Ian A. Taylor, Andres Ramos: “IMP1 KH1 and KH2 domains create a structural platform with unique RNA recognition and re-modelling properties”, *Nucleic Acids Research*, Vol. 47(8), pp. 4334–4348, 2019.

URL <http://dx.doi.org/10.1093/nar/gkz136>

Main reference Giuseppe Nicastro, Adela M. Candel, Michael Uhl, Alain Oregioni, David Hollingworth, Rolf Backofen, Stephen R. Martin, Andres Ramos: “Mechanism of β -actin mRNA Recognition by ZBP1”, *Cell Reports*, Vol. 18(5), pp. 1187–1199, 2017.

URL <http://dx.doi.org/10.1016/j.celrep.2016.12.091>

Most RNA-binding proteins recognise their RNA targets with the combinatorial action of multiple RNA-binding domains. This complexity is tunable, and allows the target-dependent recognition of a diverse range of features and sequences, underlying the capability of individual proteins to regulate multiple steps of RNA metabolism. A key question in RNA regulation is how the domains cooperate in target recognition, both for individual targets and at the transcriptome level.

We answer this question on the IGF2-mRNA binding protein 1 (IMP1), a key regulator of RNA metabolism, transport and translation. IMP1 plays an important role in defining synaptic morphology in human neurons and has a general function in regulating cell motility and differentiation. Further, IMP1 is expressed at very low levels in most cells in the adult, but high level of IMP1 expression in cancer are connected to cancer cell invasion and to the final outcome of the pathology. At the molecular level, IMP1 regulates the localisation, translation and stability of different sets of mRNA targets. The protein contains six putative RNA-binding domains – two RNA recognition motifs (RRMs) and four K-homology (KH) domains – that are organized in two-domain units.

We are using an ensemble of in vitro (e.g. NMR, X-ray crystallography, BLI, CD etc) to characterise the interaction of the individual domains and the larger multi-domain units in RNA binding. We have then built computational models of the kinetic pathways followed by the individual binding events, and that we can relate to microscopy data. Our results indicate that the two domains within one unit are strongly coupled and that each KH di-domain unit can bind RNA with high affinity and re-model it. However, the sequence selectivity and binding mechanisms are different in the two di-domains, which act quasi-independently and with very different kinetic properties. Importantly, we find these concepts are overall valid at the transcriptome level by performing in a novel computational analysis to compare sets of in vivo transcriptome-wide binding data recorded on protein mutants where RNA binding of individual domains has been knocked off using structure driven mutations.

4.18 Coding regions regulate mRNA stabilities in human cells

Olivia Rissland (University of Colorado – Denver, US)

License © Creative Commons BY 3.0 Unported license
© Olivia Rissland

A new paradigm has emerged that coding regions can regulate mRNA stability. Here, due to differences in cognate tRNA abundance, synonymous codons are translated at different speeds, and slow codons then stimulate mRNA decay. To ask if this phenomenon also

occurs in humans, we isolated RNA stability effects due to coding regions with the human ORFeome collection. We find that coding regions change mRNA stability primarily through translation. Instability-associated codons are translated more slowly, providing the first connection in humans between elongation speed and mRNA decay. Surprisingly, and in contrast to the existing model, the encoded amino acid also plays a key role. Analysis of ribosome profiling datasets indicates that decoding rates generally determine elongation speeds and are themselves likely controlled by both tRNA abundance and charging. Thus, we propose that both codon and amino acid usage regulate human mRNA stability, which may allow for coordinated regulation of related genes.

4.19 Eukaryotic-wide reconstruction of RNA-binding protein specificity by joint matrix factorization

Alexander Sasse (University of Toronto, CA)

License © Creative Commons BY 3.0 Unported license

© Alexander Sasse

Joint work of Alexander Sasse, Debashish Ray, Kaitlin U. Laverty, Hong Zheng, Kate Nie, Mihai Albu, Matthew H. Weirauch, Timothy R. Hughes, Quaid Morris

Main reference Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecnas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid Morris, Timothy R. Hughes: “A compendium of RNA-binding motifs for decoding gene regulation”, *Nature*, Vol. 499(7457), pp. 172–177, 2013.

URL <http://dx.doi.org/10.1038/nature12311>

Messenger RNA (mRNA) maturation is defined by co- and post-transcriptional interactions with RNA binding proteins (RBPs). Most RBPs possess unique binding specificities towards sequence, or sequence-structure patterns, called motifs. The largest set of these sequence motifs has been measured by RNAcompete, an in vitro assay which measures binding strength of the protein to a designed pool of 250,000 RNA sequences (Ray et al. 2013). Previously, these measurements were used to infer motifs of uncharacterized eukaryotic RBPs that shared at least 70% sequence identity to a measured protein sequence. However, for sequences containing RNA recognition motif domains (RRM), the most abundant RNA binding domain, predictions based on sequence identity between 40 to 70% were ambiguous. To address this issue, we developed a new computational method, called joint matrix factorization (jMF), which infers binding preferences from peptide profiles (k-mers). jMF circumvents the need for sequence alignments, which can be error prone, and enables high confidence predictions for about 15% more RBPs, formerly ambiguous cases. Moreover, jMF predicts the importance of individual peptides for different binding specificities. Tested on co-complex structures of RRMs, these peptide scores showed significant improvements in predicting RNA binding sites from protein sequence compared to classical conservation scores. Combining RNAcompete measurements with computational predictions from jMF we increased the total number of eukaryotic RBPs with known specificities from 12,000 to 32,500 (RRM/KH 36%) across more than 700 species, leading to 121 motifs for *Homo sapiens*, 51 for *C. elegans*, and 48 for *Arabidopsis thaliana*. We used the inferred set of binding specificities for *Arabidopsis thaliana* and combined it with gene expression data from 67 tissue types to determine crucial post-transcriptional regulators. The extended compendium of measured and inferred RNA binding specificities will be available on the CisBP-RNA database (<http://cisbp-rna.cabr.utoronto.ca/>)

4.20 Decoding regulatory protein-RNA interactions by combining integrative structural biology and large-scale approaches

Michael Sattler (*Helmholtz Zentrum – München, DE*)

License © Creative Commons BY 3.0 Unported license

© Michael Sattler

Main reference Tim Schneider, Lee-Hsueh Hung, Masood Aziz, Anna Wilmen, Stephanie Thaum, Jacqueline Wagner, Robert Janowski, Simon Müller, Silke Schreiner, Peter Friedhoff, Stefan Hüttelmaier, Dierk Niessing, Michael Sattler, Andreas Schlundt, Albrecht Bindereif: “Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3”. *Nature Communications* 10:1–18, 2019

URL <https://doi.org/10.1038/s41467-019-09769-8>

Main reference Hamed Kooshapur, Nila Roy Choudhury, Bernd Simon, Max Mühlbauer, Alexander Jussupow, Noemi Fernandez, Alisha N. Jones, Andre Dallmann, Frank Gabel, Carlo Camilloni, Gracjan Michlewski, Javier F. Caceres, Michael Sattler: “Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1”. *Nat Commun* 9, 2479, 2018.

URL <https://doi.org/10.1038/s41467-018-04871-9>

RNA plays essential roles in virtually all aspects of gene regulation, where single-stranded or folded regulatory RNA motifs are recognized by RNA binding proteins (RBPs). Most eukaryotic RBPs are multi-domain proteins that comprise various structural domains to mediate protein-RNA or protein-protein interactions. Linked to this molecular mechanisms of the formation and function of regulatory protein-RNA complexes often involve dynamic structural ensembles and can be controlled by population shifts between inactive and inactive conformations. The domains in these proteins are often connected or flanked by intrinsically disordered regions, where posttranslational modifications can further modulate the protein-RNA interactions to regulate the biological activity. We employ integrative structural biology combining solution techniques such as NMR, small angle scattering (SAXS/SANS) and FRET with X-ray crystallography and biophysical techniques to unravel the molecular recognition and dynamics for the assembly and molecular function of regulatory RNP (ribonucleoprotein) complexes. Three examples are discussed that highlight the role of conformational changes and dynamics in the function of RNPs. 1) We found that the U2AF2 RNA binding specificity for Py-tract RNA depends on an intrinsically disordered linker region flanking the canonical RNA binding domains. 2) Recognition of multipartite cis-regulatory motifs by the multi-domain RBP IMP3 involves cooperative protein-RNA interactions. 3) Recognition of the terminal loop of the pri-miR-18a hairpin by hnRNP A1 involves partial melting of the upper stem region to enhance its processing and function. Our data provide unique insight into conformational dynamics underlying the regulation of essential biological processes. The combination of experimental biophysical and structural biology techniques with large-scale genome-wide mapping and efficient computational tools is essential to unravel the protein-RNA code.

4.21 GraphProt2: deep learning for graphs meets RBP binding site prediction

Michael Uhl (*Universität Freiburg, DE*)

License © Creative Commons BY 3.0 Unported license

© Michael Uhl

CLIP-seq is the current state-of-the-art technique to experimentally determine transcriptome-wide binding sites of RNA-binding proteins (RBPs). However, since it relies on gene expression which is highly variable between conditions, it cannot provide a complete picture of the RBP

binding landscape. This necessitates the use of computational methods to predict missing binding sites, which is usually done by learning relevant features from identified sites and then use the learned model for prediction on unseen sequences. Here we present GraphProt2, a computational RBP binding site prediction method based on graph convolutional neural networks (GCNs). GraphProt2 converts the input binding sites into graphs and uses these for model training and prediction. In contrast to popular convolutional neural network (CNN) methods, this allows for variable length input as well as the possibility to add base pair information. Furthermore, additional features such as accessibility, conservation or region type information can be added as feature vectors to each node to improve predictive performance. Preliminary results show superior performance when compared to GraphProt as well as iDeepS, a CNN-based method that also utilizes a long short-term memory (LSTM) extension. For single RBP models, average accuracy in 10-fold cross validation over 33 eCLIP datasets was 86.13% (SD: ± 0.84) for GraphProt2, 81.87% (SD: ± 1.20) for iDeepS, and 77.54% (no SD information given) for GraphProt.

4.22 Breaking apart 3'-UTRs to model in vivo post-transcriptional regulation

Charles E. Vejnar (Yale University – New Haven, US)

License © Creative Commons BY 3.0 Unported license
© Charles E. Vejnar

Main reference Charles E. Vejnar, Mario Abdel Messih, Carter M. Takacs, Valeria Yartseva, Panos Oikonomou, Romain Christiano, Marlon Stoeckius, Stephanie Lau, Miler T. Lee, Jean-Denis Beaudoin, Damir Musaev, Hiba Darwich-Codore, Tobias C. Walthers, Saeed Tavazoie, Daniel Cifuentes, Antonio J. Giraldez: “Genome wide analysis of 3' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish”, *Genome Res.*, Vol. 29(7), pp. 1100–1114, 2019.

URL <https://doi.org/10.1101/gr.245159.118>

Post-transcriptional regulation plays a crucial role in shaping gene expression. During the Maternal-to-Zygotic Transition (MZT), thousands of maternal transcripts are regulated. However, how different cis-elements and trans-factors are integrated to determine mRNA stability remains poorly understood. Here, we show that most transcripts are under combinatorial regulation by multiple decay pathways during zebrafish MZT. Using a massively parallel reporter assay, we identified cis-regulatory sequences in the 3'-UTR, including U-rich motifs that are associated with increased mRNA stability. In contrast, miR-430 target sequences, UAUUUUU AU-rich elements (ARE), CCUC and CUGC elements emerged as destabilizing motifs, with miR-430 and AREs causing mRNA deadenylation upon genome activation. We identified trans-factors by profiling RNA-protein interactions and found that poly(U) binding proteins are preferentially associated with 3'-UTR sequences and stabilizing motifs. We demonstrate that this activity is antagonized by C-rich motifs and correlated with protein binding. Finally, we integrated these regulatory motifs into a machine learning model that predicts reporter mRNA stability in vivo.

4.23 Deep Learning for Modeling Translation events

Jiayang Zeng (Tsinghua University – Beijing, CN)

License  Creative Commons BY 3.0 Unported license
© Jiayang Zeng

Conventionally mRNAs are thought to only transfer the genetic information into protein sequences during translation. Recently more and more evidence has shown that mRNA sequences also encode the regulatory code that modulates translation initiation, elongation and termination. Now the high-throughput technique, Ribosome profiling, provides a large amount of data to measure the translational activities. In addition, deep learning has become a powerful machine learning tool for addressing the large-scale learning tasks. Then it remains unknown whether we can apply deep learning techniques to fully exploit the available large-ribosome profiling data to decode the sequence determinants of translation regulation. Here, we develop three deep learning models to achieve this goal. In particular, we first apply a CNN model to predict the ribosome stalling events from the normalized ribosome footprints. Then we develop a deep reinforcement learning framework to select the most important codon features and make accurate prediction on ribosome density. Finally, we propose a hybrid deep learning model to predict translation initiation sites. Tests on real ribosome profiling data show that our models can achieve accurate predictions, outperform conventional learning models, and provide useful biological insights into understanding the translation mechanisms.

Participants

- Amir Argoetti
Technion – Haifa, IL
- Rolf Backofen
Universität Freiburg, DE
- Marina Chekulaeva
Max-Delbrück-Centrum –
Berlin, DE
- Jörg Fallmann
Universität Leipzig, DE
- Jan Gorodkin
University of Copenhagen, DK
- Florian Heyl
Universität Freiburg, DE
- Eckhard Jankowsky
Case Western Reserve University
– Cleveland, US
- Hilal Kazan
Antalya International
University, TR
- Julian König
Institut für Molekulare Biologie –
Mainz, DE
- Markus Landthaler
Max-Delbrück-Centrum –
Berlin, DE
- Donny Licatalosi
Case Western Reserve University
– Cleveland, US
- Yael Mandel-Gutfreund
Technion – Haifa, IL
- Irmtraud Meyer
Max-Delbrück-Centrum –
Berlin, DE
- Neelanjan Mukherjee
University of Colorado –
Aurora, US
- Uwe Ohler
Max-Delbrück-Centrum –
Berlin, DE
- Yaron Orenstein
Ben Gurion University –
Beer Sheva, IL
- Teresa Przytycka
National Center for
Biotechnology – Bethesda, US
- Michal Rabani
The Hebrew University of
Jerusalem, IL
- Andres Ramos
University College London, GB
- Olivia Rissland
University of Colorado –
Denver, US
- Alexander Sasse
University of Toronto, CA
- Michael Sattler
Helmholtz Zentrum –
München, DE
- Michelle Scott
University of Sherbrooke, CA
- Michael Uhl
Universität Freiburg, DE
- Charles E. Vejnár
Yale University – New Haven, US
- Katharina Zarnack
Goethe-Universität –
Frankfurt am Main, DE
- Jianyang Zeng
Tsinghua University –
Beijing, CN



Computational Proteomics

Edited by

Nuno Bandeira¹ and Lennart Martens²

1 University of California – San Diego, US, bandeira@ucsd.edu

2 Ghent University, BE, lennart.martens@ugent.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19351 “Computational Proteomics”. The Seminar was originally built around four topics, identification and quantification of DIA data; algorithms for the analysis of protein cross-linking data; creating an online view on complete, browsable proteomes from public data; and detecting interesting biology from proteomics findings. These four topics were led to four corresponding breakout sessions, which in turn led to five offshoot breakout sessions.

The abstracts presented here first describe the four topic introduction talks, as well as a fifth, cross-cutting topic talk on bringing proteomics data into clinical trials. These talk abstracts are followed by one abstract each per breakout session, documenting that breakout’s discussion and outcomes.

An Executive Summary is also provided, which details the overall seminar structure, the relationship between the breakout sessions and topics, and the most important conclusions for the four topic-derived breakouts.

Seminar August 25–30, 2019 – <http://www.dagstuhl.de/19351>

2012 ACM Subject Classification Applied computing → Bioinformatics


Keywords and phrases computational biology, computational mass spectrometry, proteomics

Digital Object Identifier 10.4230/DagRep.9.8.70

1 Executive Summary

Lennart Martens (Ghent University, BE)

Nuno Bandeira (University of California – San Diego, US)

License  Creative Commons BY 3.0 Unported license
© Lennart Martens and Nuno Bandeira

The Dagstuhl Seminar 19351 ‘Computational Proteomics’ discussed several key challenges of facing the field of computational proteomics. The topics discussed were varied and wide-ranging, and radiated out from the four topics set out at the start.

These four topics were (i) personally identifiable proteomics data; (ii) unique computational challenges in data-independent analysis (DIA) approaches; (iii) computational approaches for cross-linking proteomics; and (iv) the visual design of proteomics data and results, to communicate more clearly to the broad life sciences community. A cross-cutting topic was introduced as well, which focused on proteotyping in clinical trials as it brings many of the previous challenges together, by asking the logical but complex question of how proteomics approaches, data, and associated computational methods and tools can become part of routine clinical trial data acquisition, monitoring and processing.



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

Computational Proteomics, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 70–83

Editors: Nuno Bandeira and Lennart Martens



Dagstuhl Reports

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

Based on these initial topics, breakout sessions were organized around proteomics data privacy, dealing with data from DIA approaches, how to best utilize computational approaches to use cross-linking for structural elucidation, and the importance of visualisation of proteomics data and results to engender excitement for the field's capabilities in the life sciences in general. However, these breakout sessions in turn inspired additional breakout sessions on associated topics.

The DIA and cross-linking breakouts both yielded the issue of ambiguity in identification as a cross-cutting topic that merited its own dedicated breakout session. A closely related breakout session, derived from the proteomics privacy and DIA sessions, centered on open modification searches, which are now becoming feasible in proteomics for the first time, but which are also prone to potentially crippling ambiguity issues while raising even more complex privacy issues. The visual design breakout explicitly identified multi-omics data integration as a direct offshoot of its discussions, which led to a dedicated breakout session on this topic as well. Another emerging breakout session concerned public data, which was triggered by both the DIA and cross-linking topics because of their shared need to disseminate their respective specialised data and results in a standardised, uniform, and well-structured manner. Finally, the cross-linking and DIA topics also led to a breakout session on ion mobility, as this technological advance was seen as a key aspect in the future of these technologies.

Each of these breakout sessions had exciting outcomes, and gave rise to future research ideas and collaborations. The proteomics privacy breakout concluded that the field is now ready to delve in more detail into the issues surrounding proteomics data privacy concerns, and that a white paper will be written that can be used to propose policy and to inform the community. The DIA breakout identified three such future tasks: (i) to develop a perspective manuscript that will discuss peptide-centric and spectrum-centric FDR, as well as the effects of shared evidence; (ii) to conduct an experiment for testing DDA versus DIA on the same sample to discover the sampling space for precursors and fragments; and (iii) to conduct a second experiment for understanding target/decoy scoring for different decoy generation models using both synthetic and predicted target/decoy peptides. The cross-linking breakout concluded that a cross-linked ribosomal protein complex should be used as a standardized dataset publicly available to the community, while a 'Minimum Information Requirements About a Cross Linking Experiment (MIRACLE)' was proposed to unify results from many crosslinking tools. The results will also be presented at the Symposium on Structural Proteomics in Göttingen in November 2019. The visual design breakout came up with many fine-grained conclusions, but also with an overall design philosophy which centered on three levels of technical detail, depending on the audience: i) interfaces for detailed data exploration for experienced consumers; ii) interfaces with minimal technical information, focusing on high-level data for the specific scientific question for novice consumers; and iii) interfaces with only relevant information for clinical decision making (e.g. short list of proteins significantly affected by the disease) for clinicians.

The five offshoot breakouts described above also came to conclusions, and the interested reader is referred to the corresponding abstracts for details.

Overall, the 2019 Dagstuhl Seminar on Computational Proteomics was extremely successful as a catalyst for careful yet original thinking about key challenges in the field, and as a means to make progress by setting important, high impact goals to work on in close collaboration. Moreover, during the Seminar, several highly interesting topics for a future Dagstuhl Seminar on Computational Proteomics were proposed, showing that this active and inspired community has not yet run out of challenges, nor out of ideas and opportunities!

2 Table of Contents

Executive Summary

<i>Lennart Martens and Nuno Bandeira</i>	70
--	----

Overview of Talks

Topic Introduction: Protein Cross-linking <i>Michael Götze, Robert Chalkley, Michael Hoopmann, and Lennart Martens</i>	74
Topic Introduction: Public Proteomics Data: Visual Design and Extraction of Biological Data <i>Lennart Martens and Nuno Bandeira</i>	74
Topic Introduction: DIA Challenges and Opportunities <i>Brian Searle and Maarten Dhaenens</i>	75
Topic Introduction: Proteomics Data and Personal Identification <i>Juan Antonio Vizcaino</i>	75
Topic Introduction: Proteotyping in Clinical Trials <i>Bernd Wollscheid</i>	75

Working groups

Working Group Report: Public Proteomics Data <i>Nuno Bandeira, Harald Barsnes, Frank Conlon, Eric Deutsch, Joshua Elias, Rebekah Gundry, Sicheng Hao, Nils Hoffmann, Michelle Kennedy, Benoît Kunath, Lennart Martens, Renee Salz, Natalia Sizochenko, Yves Vandenbrouck, Olga Vitek, Juan Antonio Vizcaino, and Bernd Wollscheid</i>	76
Working Group Report: Excitement and Visualization <i>Harald Barsnes, Michael Götze, Rebekah Gundry, Sicheng Hao, Michael Hoopmann, Michelle Kennedy, Benoît Kunath, Lennart Martens, Magnus Palmblad, Renee Salz, Natalia Sizochenko, Yves Vandenbrouck, Olga Vitek, and Bernd Wollscheid</i>	77
Working Group Report: Multi-Omics Data Integration (role of proteomics; how to interface) <i>Pedro Beltrao, Frank Conlon, Lukas Käll, Renee Salz, Brian Searle, Natalia Sizochenko, Yves Vandenbrouck, Olga Vitek, Mathias Wilhelm, Bernd Wollscheid, and Roman Zubarev</i>	78
Working Group Report: Open Modification Searches <i>Robert Chalkley, Nuno Bandeira, Lieven Clement, David Creasy, Bernard Delanghe, Joshua Elias, Michael Götze, Lukas Käll, and Juan Antonio Vizcaino</i>	78
Working Group Report: Ambiguity in Identification (at multiple levels, including FDR) <i>Lieven Clement, Robert Chalkley, Bernard Delanghe, Joshua Elias, and Michael Hoopmann</i>	79
Working Group Report: Cross-linking <i>Michael Hoopmann, Pedro Beltrao, Robert Chalkley, David Creasy, Bernard Delanghe, Michael Götze, Lennart Martens, and Magnus Palmblad</i>	80

Working Group Report: Ion Mobility
Hannes Röst, Sebastian Böcker, David Creasy, Eric Deutsch, Maarten Dhaenens, Birgit Schilling, Brian Searle, Stefan Tenzer, Hans Vissers, and Mathias Wilhelm 81

Working Group Report: Data Independent Acquisition
Brian Searle, Sebastian Böcker, Lieven Clement, Maarten Dhaenens, Lukas Käll, Hannes Röst, Birgit Schilling, Stefan Tenzer, Hans Vissers, and Mathias Wilhelm 81

Working Group Report: Proteomics Data Privacy
Juan Antonio Vizcaino, Nuno Bandeira, Eric Deutsch, and Benoît Kunath 82

Participants 83

3 Overview of Talks

3.1 Topic Introduction: Protein Cross-linking

Michael Götze (ETH Zürich, CH), Robert Chalkley (University of California – San Francisco, US), Michael Hoopmann (Institute for Systems Biology – Seattle, US), and Lennart Martens (Ghent University, BE)

License © Creative Commons BY 3.0 Unported license
© Michael Götze, Robert Chalkley, Michael Hoopmann, and Lennart Martens

The data acquired from cross-linking mass spectrometry (MS) poses specific challenges. These can be split into data processing and analysis concerns on the one hand, and meta-context issues on the other hand. The former revolve around combinatorial problems, due to the large number of possible cross-links that need to be explored.

This in turn leads to ambiguity problems, which are similar to, but exaggerated compared to, classical shotgun proteomics. A further consequence is the apparent limited overlap between different identification algorithms. A last data processing and analysis issue is protein inference, as not only do we need to infer proteins for each linked peptide, we also need to take into account that one of the linked peptides can be very short, which in turn exacerbates the problem.

When it comes to meta-context issues, the first concerns the wide range of the scale: from within-individual protein crosslinking to whole proteome crosslinking. Standard formats are not optimally accommodating right now, and this hampers data dissemination. While initial progress is being made, standard (reference) data sets are not yet sufficiently developed.

Finally, there are opportunities to bring crosslinking results to structural biologists.

3.2 Topic Introduction: Public Proteomics Data: Visual Design and Extraction of Biological Data

Lennart Martens (Ghent University, BE) and Nuno Bandeira (University of California – San Diego, US)

License © Creative Commons BY 3.0 Unported license
© Lennart Martens and Nuno Bandeira

Public proteomics data is currently focused internally primarily. This means that our visualisations are not readily understood outside of our field.

We therefore need a (new) visual design language that can communicate the pertinent information in a readily understood context to specific (outside) users.

We also need to consider the value of public data for novel biological discovery. There is undoubtedly low-hanging fruit there, but we should also look forward at what kind of data (and metadata!) we need to go beyond the low-hanging fruit. In that context, is there something ‘special’ about proteomics data that makes it more interesting or more relevant to reprocess?

Finally, can we leverage the novel biology that can be found in proteomics data to cement the unique contributions of proteomics data in the context of multi-omics data?

3.3 Topic Introduction: DIA Challenges and Opportunities

Brian Searle (Institute for Systems Biology – Seattle, US) and Maarten Dhaenens (Ghent University, BE)

License © Creative Commons BY 3.0 Unported license
© Brian Searle and Maarten Dhaenens

Data independent acquisition (DIA) mass spectrometry is emerging as a powerful alternative to data dependent acquisition (DDA) and parallel reaction monitoring (PRM). We posit the following questions:

- Are we quantifying peptides at the cost of making detections? How can we convince people to move beyond summing fragments and peptides for protein quant?
- What does FDR mean for DIA? Does target/decoy work the same way as for DDA?
- How to best incorporate ion mobility for DIA? Is establishing peptide overlap for selection potentially more useful than using ion mobility for separation?
- Spectrum-centric versus peptide-centric; and what can we learn from combining these, especially for shared evidence between peptides and PTM positional isomers?
- Can we build and query DIA-based repositories at a raw data level? What types of questions can we answer with “unanticipated” peptide queries across experiments and labs? Is DIA-data (considering it as a digital copy of a sample) even transferable or re-usable between individual labs?
- What can we learn from DIA for proteomics, and how can we apply it to measure metabolites?
- Is it possible to re-use libraries for DIA?

3.4 Topic Introduction: Proteomics Data and Personal Identification

Juan Antonio Vizcaino (EBI – Hinxton, GB)

License © Creative Commons BY 3.0 Unported license
© Juan Antonio Vizcaino

The detection of genomic variants on a proteome level implies that clinically sensitive proteomics data could be patient-identifiable, and then it should be protected appropriately (for instance, in the context of GDPR guidelines in the European Union).

It is now the right time to assess the state-of-the-art and develop guidelines that are applicable to the community as a whole. Future data management policies for access to human proteomics data in the public domain are part of these efforts.

3.5 Topic Introduction: Proteotyping in Clinical Trials

Bernd Wollscheid (ETH Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Bernd Wollscheid

Thinking about and discussing “Clinical Proteotype Analysis” is helpful in order to focus, connect, compare & to make strategic decisions.

- Tumor Profiler project as an example for making such strategic decisions
- Which molecular data is 2020 useful in order to support clinical decision-making beyond the current state-of-the-art?
- What is/could be the role of proteotype analysis in the clinical decision-making process?
- In order to participate in observational & interventional clinical trials we (the proteotype analysis community) needs to make sensible decisions at all levels (sample, sample processing (ID, quant, crosslinking, interactomics, surfaceome analysis etc), data acquisition, data analysis, data visualization, data privacy, data sharing etc)
- Matched data generation from the same clinical specimen

4 Working groups

4.1 Working Group Report: Public Proteomics Data

Nuno Bandeira (University of California – San Diego, US), Harald Barsnes (University of Bergen, NO), Frank Conlon (University of North Carolina – Chapel Hill, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), Joshua Elias (Chan Zuckerberg Biohub, US), Rebekah Gundry (University of Nebraska – Omaha, US), Sicheng Hao (Northeastern University – Boston, US), Nils Hoffmann (ISAS – Dortmund, DE), Michelle Kennedy (Princeton University, US), Benoît Kunath (University of Luxembourg, LU), Lennart Martens (Ghent University, BE), Renee Salz (Radboud University Nijmegen, NL), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandenbrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), Juan Antonio Vizcaino (EBI – Hinxton, GB), and Bernd Wollscheid (ETH Zürich, CH)

License © Creative Commons BY 3.0 Unported license

© Nuno Bandeira, Harald Barsnes, Frank Conlon, Eric Deutsch, Joshua Elias, Rebekah Gundry, Sicheng Hao, Nils Hoffmann, Michelle Kennedy, Benoît Kunath, Lennart Martens, Renee Salz, Natalia Sizochenko, Yves Vandenbrouck, Olga Vitek, Juan Antonio Vizcaino, and Bernd Wollscheid

Public availability of proteomics mass spectrometry data has continued to increase to hundreds of terabytes in thousands of datasets from very diverse studies and organisms. However, the lack of metadata describing the samples, experimental design and details of data acquisition and analysis continue to complicate data reutilization and make it difficult for most community members to benefit from the large volume of available data.

This breakout group aimed to compose a vision for the future of public proteomics mass spectrometry data, with a special emphasis on how to make the data most useful to enable clinical and biological discovery.

Two major use cases were proposed to guide the discussion: a) controlled-access clinical proteomics data, typically acquiring larger sample sizes using uniform protocols and featuring extensive sample metadata (often in electronic medical records) and b) open-data research proteomics data, typically acquiring small sample sizes using one or more lab-specific protocols and providing little-to-no metadata describing the study and experiments.

The discussion then focused on incentives that could be implemented to increase the level of annotation of public datasets: i) global associations of expression patterns (e.g., protein expression across tissues, samples-like-mine, etc), ii) offering research tools on the repositories that eventually store the public version of the data (e.g., differential expression, visualization, etc), iii) principal investigator tools (e.g., lab-wide statistics, reports and query features to

support grant writing, etc) and iv) publication guidelines and requirements (e.g., minimal metadata to describe the statistical tests, file and reporting formats, etc).

Finally the group acknowledged the need for example reference datasets illustrating the levels of data and metadata annotation that would be ideal for several classes of technical and biological datasets.

4.2 Working Group Report: Excitement and Visualization

Harald Barsnes (University of Bergen, NO), Michael Götze (ETH Zürich, CH), Rebekah Gundry (University of Nebraska – Omaha, US), Sicheng Hao (Northeastern University – Boston, US), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Michelle Kennedy (Princeton University, US), Benoît Kunath (University of Luxembourg, LU), Lennart Martens (Ghent University, BE), Magnus Palmblad (Leiden University Medical Center, NL), Renee Salz (Radboud University Nijmegen, NL), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandembrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), and Bernd Wollscheid (ETH Zürich, CH)

License © Creative Commons BY 3.0 Unported license

© Harald Barsnes, Michael Götze, Rebekah Gundry, Sicheng Hao, Michael Hoopmann, Michelle Kennedy, Benoît Kunath, Lennart Martens, Magnus Palmblad, Renee Salz, Natalia Sizochenko, Yves Vandembrouck, Olga Vitek, and Bernd Wollscheid

Creating excitement for proteomics in the community at large, and especially among fellow scientists, is an important goal for the field of proteomics. This starts by figuring out what (mass spectrometry-based) proteomics provides that related technologies do not, and then come up with useful visualizations showing these unique aspects.

Some of the highlighted topics were: i) biological context (e.g. proteins carry out the function, and the majority of drug targets are proteins); ii) the measurement of aggregate events such as post-transcriptional regulation; iii) antibody-independent detection and quantitation of proteins; and iv) the location of post-translational modifications can only be determined by proteomics.

The reasons why proteomics is not well appreciated by other fields was discussed next. This included: i) limited number of known success stories; ii) perceived as inconsistent; iii) higher complexity of the data, i.e. making it harder to interpret; and iv) the high variability in available technologies making it difficult to select the right one.

A couple of solutions were suggested: i) better management of expectations and mindful reporting; ii) create a central source of information on what “proteomics can do for you”; iii) resources promoting proteomics for the non-expert user community; and iv) editorial board members of biology-focused journals should invite contributions focusing on proteomic technologies for non-experts.

Finally, it was suggested that proteomics users can roughly be split into three general categories, all requiring different types of data and visualizations: i) experienced consumers wanting interactive, visual, interfaces for exploring the data in detail; ii) novice consumers requiring minimum amounts of technical information, focusing on what matters to their specific scientific question; iii) clinicians requiring only the information needed to make a clinical decision, e.g. the short list of proteins significantly affected by the disease.

4.3 Working Group Report: Multi-Omics Data Integration (role of proteomics; how to interface)

Pedro Beltrao (EBI – Hinxton, GB), Frank Conlon (University of North Carolina – Chapel Hill, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Renee Salz (Radboud University Nijmegen, NL), Brian Searle (Institute for Systems Biology – Seattle, US), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandenbrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), Mathias Wilhelm (TU München, DE), Bernd Wollscheid (ETH Zürich, CH), and Roman Zubarev (Karolinska Institute – Stockholm, SE)

License © Creative Commons BY 3.0 Unported license

© Pedro Beltrao, Frank Conlon, Lukas Käll, Renee Salz, Brian Searle, Natalia Sizochenko, Yves Vandenbrouck, Olga Vitek, Mathias Wilhelm, Bernd Wollscheid, and Roman Zubarev

Multi-omics data integration can be defined as deriving knowledge from the combination of different Omics measurements that is not possible to obtain from individual data types.

In our discussion, we identified as a major challenge in pursuing such multi-omics studies the increased complexity and skill sets required for the generation and analysis of data of multiple different types. Training is therefore a major issue for developing and carrying out multi-omics studies, and there is a need for combined expertise before a multi-omics project is started and before funding is requested. Most often researchers do not understand what are the opportunities and specific benefits from each Omics technology and how they can be combined in useful ways.

As a concrete step forward, we sought to answer the question, “What can we learn at the interface of omics?” We started to generate a document with pairwise -omics intersections and listed what we thought were a subset of open avenues of research made available by combining different -omics techniques. For some intersections, we added citations to relevant literature that showcases the power of these integrative approaches. This could be developed further into a perspective piece that would help new researchers develop questions to ask about their biological problems and determine which specific methods to focus on learning. This perspective could also serve as an opportunity to generate enthusiasm towards the use of proteomics methods and to highlight where new computational methods are most needed.

4.4 Working Group Report: Open Modification Searches

Robert Chalkley (University of California – San Francisco, US), Nuno Bandeira (University of California – San Diego, US), Lieven Clement (Ghent University, BE), David Creasy (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Joshua Elias (Chan Zuckerberg Biohub, US), Michael Götze (ETH Zürich, CH), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), and Juan Antonio Vizcaino (EBI – Hinxton, GB)

License © Creative Commons BY 3.0 Unported license

© Robert Chalkley, Nuno Bandeira, Lieven Clement, David Creasy, Bernard Delanghe, Joshua Elias, Michael Götze, Lukas Käll, and Juan Antonio Vizcaino

There is a long list of tools that have been developed for identifying unanticipated modifications through open mass modification searching. The group discussion mostly focused on three topics: 1. How best to convert an observed modification mass into a named structure; 2. How to assign biological significance to modifications to decide which are worthy of

follow-up; 3. How to create a knowledgebase such that other researchers can learn from previous identifications.

The major outcome from the discussion was a list of recommendations as to how discovered modifications should be reported and stored for community knowledge. This included submitting discovered modifications to Unimod and linking to example spectra in data submitted to a public repository through a universal spectral identifier. A spectral library of these modifications should also be created, although it was acknowledged that there may be challenges in controlling the FDR and FLR in this resource.

4.5 Working Group Report: Ambiguity in Identification (at multiple levels, including FDR)

Lieven Clement (Ghent University, BE), Robert Chalkley (University of California – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Joshua Elias (Chan Zuckerberg Biohub, US), and Michael Hoopmann (Institute for Systems Biology – Seattle, US)

License © Creative Commons BY 3.0 Unported license
© Lieven Clement, Robert Chalkley, Bernard Delanghe, Joshua Elias, and Michael Hoopmann

Ambiguity is introduced at different levels of the proteomics data analysis workflow. At the level of the identification, protein identification, quantification and differential analysis. Current reporting is driven towards unified results, but ambiguity requires hierarchical classification which is not generally supported by visualization and table schema. Without acknowledging incorrect, though “high-confidence” ambiguous results we risk to draw biological conclusions that may be false.

This breakout group aimed at discussing on a) important types of ambiguity at different levels in the data analysis workflow, b) how these types of ambiguity could be quantified and c) reported. The discussion then focused on challenges in possible strategies and solutions to report more efficiently on ambiguity.

Finally a number of actionable outcomes were selected to be realised on the short term:

1. FDR estimation in identification is currently monopolized by variations on the target decoy approach and it is difficult to publish on alternative ways to estimate the null distribution of false PSMs. With this respect we plan a perspective paper where we will review strategies based on decoys and parametric distributions. We will elaborate on the underlying assumptions of each approach and we will highlight the importance to assess the quality of the approximation of the null distribution within the identification step of the proteomics data analysis workflow.
2. One type of ambiguity that arises in the quantification involves peptides for which the ratios deviate from the proteome ratio. We will develop statistics to prioritise proteins where this type of and we will provide plots to assess the degree of ambiguity.
3. We will assess different types of ambiguity in existing datasets and present the resulting statistics.

4.6 Working Group Report: Cross-linking

Michael Hoopmann (Institute for Systems Biology – Seattle, US), Pedro Beltrao (EBI – Hinxton, GB), Robert Chalkley (University of California – San Francisco, US), David Creasy (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Michael Götze (ETH Zürich, CH), Lennart Martens (Ghent University, BE), and Magnus Palmblad (Leiden University Medical Center, NL)

License © Creative Commons BY 3.0 Unported license
© Michael Hoopmann, Pedro Beltrao, Robert Chalkley, David Creasy, Bernard Delanghe, Michael Götze, Lennart Martens, and Magnus Palmblad

Crosslinking presents many diverse challenges for computational proteomics due to its ever evolving nature of methods and tools. This diversity has hindered development of standardized datasets and workflows.

This breakout session discussed and presented major challenges current to crosslinking data analysis and solutions that will improve upon the field. We focused on three specific tiers for improvement guidelines: the developer, user, and reporting/publication levels.

Within these tiers, primary areas to focus on include improving upon data standardization. Current open standards are poorly implemented, yet there are existing tools such as mzTab that are immediately extensible and will provide greater utility. Additionally, the current paradigm in computational solutions include using tailored datasets. Instead a robust, curated and open dataset utilizing common paradigms and with input from the community was determined to be a better benchmark for algorithm development, and a suggestion was provided.

Additionally, the field suffers greatly from poor validation techniques. For example, current methods applied in standard shotgun proteomics perform poorly when challenged with the sparse datasets in crosslinking. More efforts must be made to explore viable alternatives while expanding the discussion into orthogonal disciplines, such as machine learning.

The discussion concluded by detailing actionable items for which these issues can be addressed. Specifically, a cross-linked ribosomal protein complex could be used as a standardized dataset publicly available to the community, A Minimum Information Requirements About a Cross Linking Experiment (MIRACLE) was proposed extending mzTab that could immediately unify the results from many crosslinking tools.

Furthermore, we expanded the discussion to better define the role of ambiguity in crosslinking, which is essential to understand to improve upon validation.

This topic will be further expanded and discussed in the community at the upcoming Symposium on Structural Proteomics in Göttingen November 3–6, 2019.

4.7 Working Group Report: Ion Mobility

Hannes Röst (University of Toronto, CA), Sebastian Böcker (Universität Jena, DE), David Creasy (Matrix Science Ltd. – London, GB), Eric Deutsch (Institute for Systems Biology – Seattle, US), Maarten Dhaenens (Ghent University, BE), Birgit Schilling (Buck Institute – Novato, US), Brian Searle (Institute for Systems Biology – Seattle, US), Stefan Tenzer (Universität Mainz, DE), Hans Vissers (Waters Corporation – Wilmslow, GB), and Mathias Wilhelm (TU München, DE)

License © Creative Commons BY 3.0 Unported license
 © Hannes Röst, Sebastian Böcker, David Creasy, Eric Deutsch, Maarten Dhaenens, Birgit Schilling, Brian Searle, Stefan Tenzer, Hans Vissers, and Mathias Wilhelm

Ion mobility separation (IMS) is an emerging analytical separation technique in various proteomics application areas. It is typically combined with liquid chromatography and mass spectrometry and can provide information on structure, adds an additional dimension of separation, and can have sensitivity benefits.

Discussion topics included IMS principles, hardware configurations, and the computational tools to analyse the multi-dimensional data, which comprises of the following coordinates: retention time, drift time, precursor and product ion m/z , and intensity.

It was concluded that IMS is not widely adopted yet for qualitative and quantitative high-throughput studies. One of the key aspects revolved around the question if IMS provides as solution to the problem of resolving chemic spectra. It appeared that this is an unresolved question in the field and that further investigation is required. An experiment has been designed to assess the magnitude of this phenomena on two platforms currently available.

A short discussion on data formats showed that current open source data formats are adequate to describe raw data. A document will be distributed describing best practices.

Prediction of CCS is being explored by a number of research groups. However, the benefits of CCS predictions are yet to be determined. Lastly, the impact of IMS on the quantitative performance of label-free quantitation workflows were discussed and experimental designs to evaluate this impact proposed.

IMS has significant potential for multiple applications in MS based proteomics, including structural biology, PTM analysis, and discovery experiments.

4.8 Working Group Report: Data Independent Acquisition

Brian Searle (Institute for Systems Biology – Seattle, US), Sebastian Böcker (Universität Jena, DE), Lieven Clement (Ghent University, BE), Maarten Dhaenens (Ghent University, BE), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Hannes Röst (University of Toronto, CA), Birgit Schilling (Buck Institute – Novato, US), Stefan Tenzer (Universität Mainz, DE), Hans Vissers (Waters Corporation – Wilmslow, GB), and Mathias Wilhelm (TU München, DE)

License © Creative Commons BY 3.0 Unported license
 © Brian Searle, Sebastian Böcker, Lieven Clement, Maarten Dhaenens, Lukas Käll, Hannes Röst, Birgit Schilling, Stefan Tenzer, Hans Vissers, and Mathias Wilhelm

We discussed several open questions for data independent acquisition (DIA). We first focused on what it means to produce a digital copy of a proteome. DIA produces digital copies of the precursor and fragment ion space, while DDA produces digital copies of only the precursor space. We feel that it might be better to talk about DIA as creating a “deterministic copy” or

“consistent copy” rather than a “digital copy” to emphasize that it is not a “comprehensive copy” and it doesn’t contain every possible peptide.

We have an interest in determining some statistics about DIA to determine the parameters in which it works better than DDA. In particular, we are interested in asking:

- How many of the precursors do we see in MS1, how many of them trigger MS2 spectra?
- Are we still under-sampling the precursor space?
- Are the MS2 spectra of poor quality due to triggering early and getting poor quality spectra?

We plan to conduct an experiment to probe these questions.

We then discussed the effect of peptide-centric versus spectrum-centric searching of DIA and DDA data. With spectrum-centric searching, the “currency” of detection are peptide-spectrum matches (PSMs), which are FDR corrected using target/decoy competition. With peptide-centric searching, the currency is a p-value for each peptide, where the FDR is estimated without competition. While it is possible to use spectrum-centric analysis for peptide detection, peptide-centric analysis is used for both DDA (MS1-level) and DIA (MS2-level). The re-use of ions in peptide-centric analysis has consequences over-reporting homologous or modified peptides, and we feel that the development of a hybrid analysis method by accounting for assigned ions in a peptide-centric search will be a necessary tool to ensure that FDRs are accurately assessed.

We planned three future tasks: to develop a perspective manuscript and two experiments. The perspective manuscript will discuss peptide-centric and spectrum-centric FDR, as well as the effects of shared evidence. We planned experiment for testing DDA versus DIA on the same sample to discover the sampling space for precursors and fragments. We also planned a second experiment for understanding target/decoy scoring for different decoy generation models using both synthetic and predicted target/decoy peptides.

4.9 Working Group Report: Proteomics Data Privacy

Juan Antonio Vizcaino (EBI – Hinxton, GB), Nuno Bandeira (University of California – San Diego, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), and Benoît Kunath (University of Luxembourg, LU)

License © Creative Commons BY 3.0 Unported license

© Juan Antonio Vizcaino, Nuno Bandeira, Eric Deutsch, and Benoît Kunath

“Proteomics data privacy issues: Is proteomic data Personally Identifiable Information (PII)?” The detection of genomic variants at a proteome level implies that clinical sensitive proteomics data can be patient-identifiable, and then it should be protected appropriately (for instance, in the context of the GDPR (General Data Protection Regulation) guidelines in the European Union).

It is now the right time to assess the current state of the art and develop guidelines that are applicable to the community as a whole. Future data management policies for access to human proteomics data in the public domain are part of these efforts.

The main objective is to write a white paper that can be used to propose policy and to inform the community.

Participants

- Nuno Bandeira
University of California –
San Diego, US
- Harald Barsnes
University of Bergen, NO
- Pedro Beltrao
EBI – Hinxton, GB
- Sebastian Böcker
Universität Jena, DE
- Robert Chalkley
University of California –
San Francisco, US
- Lieven Clement
Ghent University, BE
- Frank Conlon
University of North Carolina –
Chapel Hill, US
- David Creasy
Matrix Science Ltd. –
London, GB
- Bernard Delanghe
Thermo Fisher GmbH –
Bremen, DE
- Eric Deutsch
Institute for Systems Biology –
Seattle, US
- Maarten Dhaenens
Ghent University, BE
- Joshua Elias
Chan Zuckerberg Biohub, US
- Michael Götze
ETH Zürich, CH
- Rebekah Gundry
University of Nebraska –
Omaha, US
- Sicheng Hao
Northeastern University –
Boston, US
- Nils Hoffmann
ISAS – Dortmund, DE
- Michael Hoopmann
Institute for Systems Biology –
Seattle, US
- Lukas Käll
KTH Royal Institute of
Technology – Solna, SE
- Michelle Kennedy
Princeton University, US
- Benoît Kunath
University of Luxembourg, LU
- Lennart Martens
Ghent University, BE
- Magnus Palmblad
Leiden University Medical
Center, NL
- Hannes Röst
University of Toronto, CA
- Renee Salz
Radboud University
Nijmegen, NL
- Birgit Schilling
Buck Institute – Novato, US
- Brian Searle
Institute for Systems Biology –
Seattle, US
- Natalia Sizochenko
Dartmouth College –
Hanover, US
- Stefan Tenzer
Universität Mainz, DE
- Yves Vandenbrouck
CEA – Grenoble, FR
- Hans Vissers
Waters Corporation –
Wilmslow, GB
- Olga Vitek
Northeastern University –
Boston, US
- Juan Antonio Vizcaino
EBI – Hinxton, GB
- Mathias Wilhelm
TU München, DE
- Bernd Wollscheid
ETH Zürich, CH
- Roman Zubarev
Karolinska Institute –
Stockholm, SE



Computation in Low-Dimensional Geometry and Topology

Edited by

Maarten Löffler¹, Anna Lubiw², Saul Schleimer³, and
Erin Moriarty Wolf Chambers⁴

1 Utrecht University, NL, m.loffler@uu.nl

2 University of Waterloo, CA, alubiw@uwaterloo.ca

3 University of Warwick – Coventry, GB, s.schleimer@warwick.ac.uk

4 St. Louis University, US, echambe5@slu.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19352 “Computation in Low-Dimensional Geometry and Topology”. The seminar participants investigated problems in: knot theory, trajectory analysis, algorithmic topology, computational geometry of curves, and graph drawing, with an emphasis on how low-dimensional structures change over time.

Seminar August 25–30, 2019 – <http://www.dagstuhl.de/19352>

2012 ACM Subject Classification Mathematics of computing → Geometric topology, Human-centered computing → Graph drawings, Theory of computation → Computational geometry

Keywords and phrases Geometric topology, Graph Drawing, Computational Geometry

Digital Object Identifier 10.4230/DagRep.9.8.84

Edited in cooperation with Boris Klemz


1 Executive Summary

Erin Moriarty Wolf Chambers (St. Louis University, US)

Maarten Löffler (Utrecht University, NL)

Anna Lubiw (University of Waterloo, CA)

Saul Schleimer (University of Warwick – Coventry, GB)

License  Creative Commons BY 3.0 Unported license

© Erin Moriarty Wolf Chambers, Maarten Löffler, Anna Lubiw, and Saul Schleimer

One-dimensional structures embedded in higher-dimensional spaces are ubiquitous in both the natural and artificial worlds: examples include DNA strands, migration paths, planetary orbits, rocket trajectories, robot motion planning, chip design, and many more. These are studied in different areas of mathematics and computer science, under many names: knots, curves, paths, traces, trajectories, graphs, and others. However, researchers in many areas are just beginning to apply algorithmic techniques to find efficient algorithms for these structures, especially when more fundamental mathematical results are required. Broad examples of such problems include:

- classical algorithms on trajectories like the Fréchet distance as a way to formalize a distance measure as a curve changes;
- morphing between two versions of a common graph, which again tracks a higher dimensional space that corresponds to movement of a one-dimensional object;
- drawing and manipulating objects in three-manifolds, such as graphs, curves, or surfaces; and
- perhaps the simplest problem posed (in different ways) in all these areas, “how does one draw and morph a nice curve on a nice surface?”



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

Computation in Low-Dimensional Geometry and Topology, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 84–112
Editors: Maarten Löffler, Anna Lubiw, Saul Schleimer, and Erin Moriarty Wolf Chambers



DAGSTUHL
REPORTS

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

This seminar was the second in a series. In the first seminar, the goal was to identify connections and seed new research collaborations along the spectrum from knot theory and topology, through to computational topology and computational geometry, and all the way to graph drawing. After the success of the first seminar, the goal for this second round was to continue and extend prior work, in particular by focussing on how objects change over time.

The seminar began with three overview talks from researchers in different areas (trajectory analysis, algorithmic topology, and graph drawing) to motivate and introduce problems which would fit the theme of changes over time in the representations of low-dimensional objects in higher dimensional spaces. We then invited all participants to describe open problems (most of which were circulated in advance of the meeting) that fit with the topic of the workshop and could benefit from broad expertise. For the remainder of the workshop we split into small working groups each focussed on a particular open problem.

Throughout the workshop we used Coauthor, a tool for collaboration designed by Erik Demaine (MIT), to share progress and updates among all the working groups. This, together with twice-daily progress reports, allowed us to share ideas and expertise among all participants, which was very effective. Another advantage was that we had a record of the work accomplished when the workshop ended.

Below, we (the organizers) describe the main working group topics and how they connected to the overarching theme. The abstracts of talks in the seminar and preliminary results from the working groups are also outlined later in this report.

One group worked on open questions that were motivated by 3-manifolds. In particular, they considered lower bounds for deciding the complexity of a knot or link equivalence, with a goal of finding specific knots that require many simplification moves. Their work involved both designing smaller examples, as well as doing larger scale exhaustive search using the software tool Regina.

Another group considered representations of graphs and hypergraphs by touching polygons in 3-d. They were able to leverage the dual graph of the polyhedral complex in 3d, and make progress on classifying which types of graphs could (or could not) be realized. Their problem was primarily combinatorial, but the techniques used included several interesting topological arguments about embeddings of manifolds into 3d or into 3-manifolds.

Several groups considered problems about flows or morphs of curves in various settings. One question centered on visualizing actual embedded homotopies on a given surface; there is considerable prior work on how to compute such homotopies between curves quickly, but it generally focuses on computing the complexity of the homotopy as opposed to the actual sequence of simplifications needed. The group looked more closely at this algorithm, and was able to outline a proof that in fact an extension of that algorithm would generate the actual homotopy, for a slightly higher time cost. A second group considered curves in the plane, and investigated options for computing a “nice” morph between them. As the question was more vague, the group did quite a bit of background investigation on prior work, and then discussed a new technique based on 3-manifolds and normal surface theory which might lead to a new family of morphs. A third group looked at the problem of preprocessing a given curve so that the Fréchet distance to any other query curve could be efficiently computed, and were able to obtain improved time bounds for several variants of the problem.

In summary, the workshop fostered a highly collaborative environment where combining the expertise and knowledge of researchers from different communities allowed us to solve problems of common interest across those communities. A major theme was how connected the various problems could be; often, a proof technique or piece of literature suggested by a member of a different community proved useful or insightful to a group working in a different domain.

2 Table of Contents

Executive Summary

Erin Moriarty Wolf Chambers, Maarten Löffler, Anna Lubiw, and Saul Schleimer . 84

Overview of Talks

One-Dimensional Structures in Low-Dimensional Algorithmic Topology
Arnaud de Mesmay 87

Curves, Distance Measures, and Homotopies
Carola Wenk 87

Survey on Graph and Hypergraph Drawing
André Schulz and Alexander Wolff 87

Working groups

Frechet Distance Data Structures
Maïke Buchin, Tim Ophelders, Lena Schlipf, Rodrigo I. Silveira, Frank Staals, and Ivor van der Hoog 90

Lower Bounds for the Complexity of Knot and Link Equivalence
Benjamin Burton, Hsien-Chih Chang, Arnaud de Mesmay, Francis Lazarus, Maarten Löffler, Clément Maria, Saul Schleimer, Eric Sedgwick, and Jonathan Spreer . . . 91

Combinatorial Homotopies
Hsien-Chih Chang, Vincent Despré, Linda Kleist, Francis Lazarus, Anna Lubiw, Tim Ophelders, Hugo Parlier, Saul Schleimer, Stephan Tillmann, Birgit Vogtenhuber, Carola Wenk, and Erin Moriarty Wolf Chambers 94

Simple Graph Cycle in Homotopy Class
Hsien-Chih Chang, Arnaud de Mesmay, Vincent Despré, Francis Lazarus, and Erin Moriarty Wolf Chambers 96

Beautiful curves on beautiful surfaces
Saul Schleimer, Vincent Despré, Francis Lazarus, Hugo Parlier, Stephan Tillmann, and Erin Moriarty Wolf Chambers 99

Minimum Area Homotopies
Carola Wenk, Hsien-Chih Chang, Vincent Despré, Francis Lazarus, Anna Lubiw, Tim Ophelders, Hugo Parlier, and Erin Moriarty Wolf Chambers 101

Nice Morphs and Isotopic Frechet Distance
Erin Moriarty Wolf Chambers, Vincent Despré, Linda Kleist, Maarten Löffler, Anna Lubiw, Tim Ophelders, Hugo Parlier, Stephan Tillmann, Birgit Vogtenhuber, and Carola Wenk 103


Representing Graphs by Polygons with Edge Contacts in 3D
Alexander Wolff, Elena Arseneva, Arnaud de Mesmay, Linda Kleist, Boris Klemz, Maarten Löffler, André Schulz, and Birgit Vogtenhuber 106

Participants 112

3 Overview of Talks

3.1 One-Dimensional Structures in Low-Dimensional Algorithmic Topology


Arnaud de Mesmay (University of Grenoble, FR)

License  Creative Commons BY 3.0 Unported license
© Arnaud de Mesmay

In this talk, I survey algorithms to test the topological equivalence of one-dimensional objects in dimensions two and three. More precisely, I focus on algorithms to test homotopy of curves on surfaces (and connected topics), and isotopy of knots in \mathbb{R}^3 . I explain in particular why hyperbolic geometry helps for these problems, how researchers have come up with approaches to combinatorialize it into efficient algorithms in two dimensions, and how this is lacking in 3d. We also foray into implementations.

3.2 Curves, Distance Measures, and Homotopies


Carola Wenk (Tulane University – New Orleans, US)

License  Creative Commons BY 3.0 Unported license
© Carola Wenk

We study distance measures for curves, their relation to homotopies, and curve simplification. In particular we consider Hausdorff and Fréchet distances, and discuss algorithmic and hardness results for curve simplification under these distances. As a distance measure for metric spaces we consider the Gromov Hausdorff distance in Euclidean space. Finally we study the minimum homotopy area for a closed curve in the plane and as a distance measure for two curves.

3.3 Survey on Graph and Hypergraph Drawing

André Schulz (FernUniversität in Hagen, DE) and Alexander Wolff (Universität Würzburg, DE)

License  Creative Commons BY 3.0 Unported license
© André Schulz and Alexander Wolff

In this talk we review the basic concepts and the methodology of Graph Drawing, which is an active research area in the intersection of Graph Theory, Computational Geometry, and Information Visualization. Graph Drawing is about finding algorithms that map abstract, combinatorial objects (graphs or hypergraphs) to drawings, that is, “tangible” geometric objects. The goal is to find algorithms that guarantee a provable geometric quality measure in the worst case. For example, there are algorithms that draw any planar graph on a grid whose size is quadratic in the number of vertices of the graph. Other than in areas such as Information Visualization, the evaluation is usually *not* task-driven.

The Graph Drawing problem has numerous incarnations: supported graph classes (e.g., trees, outerplanar, planar, or bipartite graphs), drawing styles (e.g., orthogonal, straight-line, Bézier), quality measures (e.g., number of bends, number of crossings, crossing resolution),

the embedding space (2D or 3D), and the type of representation (node–link diagrams, contact or intersection representations). Often, optimizing one measure leads to drawings that are bad in other measures. There is a lack of algorithms that are “pretty good” or at least “not too bad” in many aspects.

There is a (surprisingly small) set of standard techniques for drawing graphs. For example, if the graph class for which we want to design a drawing algorithm has a recursive definition, an obvious approach is to construct drawings recursively. A prominent example are orthogonal straight-line drawings of binary trees. It turns out that they can be drawn in a compact way; on a grid of size $O(n \log n)$, where n is the number of vertices of the given tree [4]. Similarly, if a graph class has an inductive definition, we may try to draw the given graph of that class inductively. Such an approach is used to show that every n -vertex planar 3-tree can be drawn using $2n - 4$ segments [6]. Finally, there are two at first glance very different approaches for drawing planar graphs on a grid of quadratic size. One approach, the shift algorithm [5], constructs the drawing incrementally; the other approach [11] counts some combinatorial objects (using a Schnyder wood) and then turns the resulting numbers into coordinates. A more careful analysis, however, reveals structural similarities between the two approaches.

Topics that have received considerable attention over the last few years are simultaneous embedding, morphing of graphs, drawings with large crossing angles, drawings of beyond-planar graphs, and visual complexity. The visual complexity of a drawing is measured by the number of geometric objects needed to compose or cover the drawing. For example, the *segment number* of a planar graph is the smallest number of straight-line segments whose union represents a straight-line drawing of the given graph [6]. The *arc number* [12] is defined accordingly with respect to circular-arc drawings, which often allow for less complex or more compact representations. Another recent generalization [10] are variants of the segment number for nonplanar graphs, either admitting crossings or embedding in 3D. The *plane cover number* [2, 3] asks for the smallest number of planes needed to cover a straight-line drawing of a given graph in 3D. Accordingly, one can define the *line cover number* in 2D (for planar graphs) or in 3D (for arbitrary graphs), which is obviously upperbounded by the corresponding segment number. Also weak versions of these numbers have been studied where only the vertices of a crossing-free straight-line drawing of the graph need to be covered. While it is not hard to see that any outerplanar graph has weak line cover number 2 [2], even some cubic, 3-connected, bipartite planar graphs have unbounded weak line cover number (exceeding $\sqrt[3]{n}$, where n is the number of vertices) [7]. On the other hand, every 4-connected plane triangulation on n vertices has weak line cover number at most $\sqrt{2n}$ [9].

This seminar’s topic – Computation in Low-Dimensional Geometry and Topology – is in line with a recent effort to better understand the drawing of graphs and hypergraphs in 3D. For example, we now know that every graph has a contact representation in 3D where vertices are represented by pairwise interior-disjoint convex polygons and edges by vertex–vertex contacts between the corresponding polygons [8]. On the other hand, the 3-uniform complete hypergraph with six (or more) vertices does not admit a representation where vertices are represented by points and hyperedges by (pairwise interior-disjoint) triangles connecting the corresponding points [1].

References

- 1 Johannes Carmesin. Embedding simply connected 2-complexes in 3-space – I. A Kuratowski-type characterisation. ArXiv report, 2019. URL: <http://arxiv.org/abs/1709.04642>.
- 2 Steven Chaplick, Krzysztof Fleszar, Fabian Lipp, Alexander Ravsky, Oleg Verbitsky, and Alexander Wolff. Drawing graphs on few lines and few planes. In Yifan Hu and Martin

- Nöllenburg, editors, *Proc. 24th Int. Symp. Graph Drawing & Network Vis. (GD'16)*, volume 9801 of *LNCS*, pages 166–180. Springer, 2016. doi:10.1007/978-3-319-50106-2_14.
- 3 Steven Chaplick, Krzysztof Fleszar, Fabian Lipp, Alexander Ravsky, Oleg Verbitsky, and Alexander Wolff. The complexity of drawing graphs on few lines and few planes. In Faith Ellen, Antonina Kolokolova, and Jörg-Rüdiger Sack, editors, *WADS 2017*, volume 10389 of *LNCS*, pages 265–276. Springer, 2017. URL: <http://arxiv.org/abs/1607.06444>, doi:10.1007/978-3-319-62127-2_23.
 - 4 Pierluigi Crescenzi, Giuseppe Di Battista, and Adolfo Piperno. A note on optimal area algorithms for upward drawings of binary trees. *Comput. Geom. Theory Appl.*, 2(4):187–200, 1992. doi:10.1016/0925-7721(92)90021-J.
 - 5 Hubert de Fraysseix, János Pach, and Richard Pollack. How to draw a planar graph on a grid. *Combinatorica*, 10(1):41–51, 1990. doi:10.1007/BF02122694.
 - 6 Vida Dujmović, David Eppstein, Matthew Suderman, and David R. Wood. Drawings of planar graphs with few slopes and segments. *Comput. Geom. Theory Appl.*, 38(3):194–212, 2007. doi:10.1016/j.comgeo.2006.09.002.
 - 7 David Eppstein. Cubic planar graphs that cannot be drawn on few lines. In *Proc. 35th Int. Symp. Comp. Geom. (SoCG'19)*, volume 129 of *LIPICs*, pages 32:1–32:15, 2019. URL: <https://arxiv.org/abs/1903.05256>, doi:10.4230/LIPICs.SocG.2019.32.
 - 8 William Evans, Paweł Rzażewski, Noushin Saeedi, Chan-Su Shin, and Alexander Wolff. Representing graphs and hypergraphs by touching polygons in 3d. In Daniel Archambault and Csaba D. Tóth, editors, *Proc. 27th Int. Symp. Graph Drawing & Network Vis. (GD'19)*, volume 11904 of *LNCS*. Springer, 2019. To appear. URL: <https://arxiv.org/abs/1908.08273>, doi:10.1007/978-3-030-35802-0_2.
 - 9 Stefan Felsner. 4-connected triangulations on few lines. In Daniel Archambault and Csaba D. Tóth, editors, *Proc. 24th Int. Symp. Graph Drawing & Network Vis. (GD'16)*, volume 11904 of *LNCS*. Springer, 2019. To appear. URL: <https://arxiv.org/abs/1908.04524>, doi:10.1007/978-3-030-35802-0_30.
 - 10 Yoshio Okamoto, Alexander Ravsky, and Alexander Wolff. Variants of the segment number of a graph. In Daniel Archambault and Csaba D. Tóth, editors, *Proc. 27th Int. Symp. Graph Drawing & Network Vis. (GD'19)*, volume 11904 of *LNCS*. Springer, 2019. To appear. URL: <https://arxiv.org/abs/1908.08871>, doi:10.1007/978-3-030-35802-0_33.
 - 11 Walter Schnyder. Embedding planar graphs on the grid. In David S. Johnson, editor, *Proc. 1st ACM-SIAM Symp. Discrete Algorithms (SODA'90)*, pages 138–148, 1990. URL: <https://dl.acm.org/citation.cfm?id=320191>.
 - 12 André Schulz. Drawing graphs with few arcs. *J. Graph Alg. Appl.*, 19(1):393–412, 2015. doi:10.7155/jgaa.00366.

4 Working groups

4.1 Fréchet Distance Data Structures

Maïke Buchin (Ruhr-Universität Bochum, DE), Tim Ophelders (Michigan State University, US), Lena Schlipf (FernUniversität in Hagen, DE), Rodrigo I. Silveira (UPC – Barcelona, ES), Frank Staals (Utrecht University, NL), and Ivor van der Hoog (Utrecht University, NL)

License © Creative Commons BY 3.0 Unported license
 © Maïke Buchin, Tim Ophelders, Lena Schlipf, Rodrigo I. Silveira, Frank Staals, and Ivor van der Hoog

The Fréchet distance is a well known distance measure between two polygonal curves P and Q . It is also well known that computing the Fréchet distance can be done in $O(nm \log(nm))$ time, where n and m are the number of vertices in the curves P and Q respectively [1]. Furthermore, it is unlikely that a significantly faster algorithm is possible [2]. We consider a version of the problem in which we are given one of the curves, say P , in advance, and we can preprocess and store it so that for a query curve Q we can more quickly compute the Fréchet distance between P and Q . Moreover, like previous results [3], we focus on the case in which Q is a single line segment. We report on some preliminary results:

- In case of the discrete Fréchet distance we can preprocess P in $O(n \log n)$ time and space so that we can report the discrete Fréchet distance in $O(\log^3 n)$ time.
- In case of the weak Fréchet distance we can preprocess P in $O(n \log n)$ time and space so that we can report the discrete Fréchet distance in $O(\log^2 n)$ time.
- In case the query segments are restricted to be horizontal, we can build a data structure of size $O(n^{3/2})$ that can report the (real) Fréchet distance between P and Q in $O(\log n)$ time. This improves the existing result by de Berg et al. [4] that requires $O(n^2)$ space.
- In case the query segment may have an arbitrary orientation, we can answer queries in $O(\log^2 n)$ time, using $O(n^{4+\varepsilon})$ time. Furthermore, we are hopeful that an extension of our technique for horizontal query segments leads to an improved data structure using only $O(n^{7/2})$ space.

References

- 1 H. Alt and M. Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5:75–91, 1995.
- 2 K. Bringmann. Why walking the dog takes time: Fréchet distance has no strongly subquadratic algorithms unless SETH fails. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 661–670, Oct 2014.
- 3 M. de Berg, J. Gudmundsson, and A. D. Mehrabi. A dynamic data structure for approximate proximity queries in trajectory data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '17*, pages 48:1–48:4, New York, NY, USA, 2017. ACM.
- 4 M. de Berg, A. D. Mehrabi, and T. Ophelders. Data structures for Fréchet queries in trajectory data. In *CCCG*, pages 214–219, 2017.

4.2 Lower Bounds for the Complexity of Knot and Link Equivalence

Benjamin Burton (The University of Queensland – Brisbane, AU), Hsien-Chih Chang (Duke University – Durham, US), Arnaud de Mesmay (University of Grenoble, FR), Francis Lazarus (GIPSA Lab – Grenoble, FR), Maarten Löffler (Utrecht University, NL), Clément Maria (INRIA – Valbonne, FR), Saul Schleimer (University of Warwick – Coventry, GB), Eric Sedgwick (DePaul University – Chicago, US), and Jonathan Spreer (University of Sydney, AU)

License © Creative Commons BY 3.0 Unported license

© Benjamin Burton, Hsien-Chih Chang, Arnaud de Mesmay, Francis Lazarus, Maarten Löffler, Clément Maria, Saul Schleimer, Eric Sedgwick, and Jonathan Spreer

A *knot* is a piecewise-linear embedding of S^1 into \mathbb{R}^3 . A *link* is a piecewise-linear embedding of a disjoint union of S^1 s into \mathbb{R}^3 . Two knots/links are considered *equivalent* if there is a continuous deformation (an *isotopy*) between them that does not induce any crossing.

Open problem: Prove complexity lower bounds for the problem of deciding knot or link equivalence.

A little background: Deciding efficiently the equivalence of two knots or links is arguably one of the biggest problems in knot theory, see for example the survey of Lackenby [8]. All the existing algorithms are terribly inefficient (at least in theory), the only one with an analyzed complexity seems to be brute-forcing Reidemeister moves using the humongous bound of Lackenby and Coward [2]: this algorithm takes time $\exp^{(c^n)}(n)$ Reidemeister moves, where $c = 10^{1000000}$, and the notation $\exp^{(*)}$ denotes a tower of exponential of height $*$.

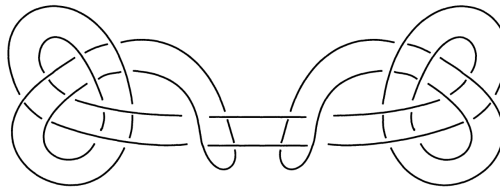
Finding new and improved algorithms probably requires some new insights and breakthroughs – this open problem aims at proving lower bounds instead. Strikingly *no* computational lower bound at all is known for this problem, even for links. While proving **NP**-hardness is a natural goal, even the seemingly humble result of proving that the link equivalence problem is at least as hard as **GRAPH ISOMORPHISM** is open (see Lackenby [10], page 2). Note that testing homeomorphism of 3-manifolds is known to be at least as hard as **GRAPH ISOMORPHISM** [5, 10].

Outputs: The objectives of the group were to investigate the complexity of testing knots and links equivalence, from two perspectives. The first one was about algorithmic complexity, and aimed at proving computational lower bounds for the problem. The second one was combinatorial, and aimed at finding instances of hard knots, i.e., knots for which any simplifying sequence of Reidemeister moves incurs a substantial increase in the number of crossings.

Despite our efforts, we could not progress on the first problem, and focused our attention to the second one, after pondering that it is consistent with current knowledge and not implausible that the knot equivalence problem lies in **NP** and **coNP**, and thus, under classical complexity hypotheses, would not be **NP**-hard, and possibly not even **Graph Isomorphism**-hard.

Hard knot diagrams

We consider the following problem: Are there diagrams of the unknot that require a substantial increase in the number of crossings in order to be simplified by Reidemeister moves to the trivial diagram? One can also consider related simplification problems, such as disconnecting diagrams of split links.



■ **Figure 1** A hard unknot diagram [6].

This question is fundamental in algorithmic knot theory as simplifying diagrams comes as very first step in any computation on knots. The increase in the size of intermediate diagrams measures the difficulty of the search for a diagram with fewer crossings.

Dynnikov’s work on arc presentation [4] implies the existence of a super polynomially long sequence of Reidemeister moves on an unknot diagram with n crossings, leading to the trivial diagram, and for which all intermediate diagrams have at most $(n-1)^2/2$ crossings. A similar result holds for disconnecting the diagram of a split link. More recently, Lackenby [9] proved that unknot and split link diagrams can be simplified in polynomially many Reidemeister moves (specifically $O(n^{11})$ moves) without exceeding a quadratic number of crossings ($O(n^2)$).

In practice however, the “hardest” unknot diagrams known require only 1 extra crossing after which they monotonically simplify. The examples in the literature [7] purporting to require two additional crossings were observed to be erroneous. Further claims of the existence of a family of unknots requiring an arbitrarily large number of additional crossings [7] are suspect.

Constructing hard unknots from any knot K

Figure 1 pictures a classical hard unknot diagram, which has been studied in the context of the energy minimization approach to the unknotting problem [6]. It can be constructed and generalized with the following procedure. For a knot K ,

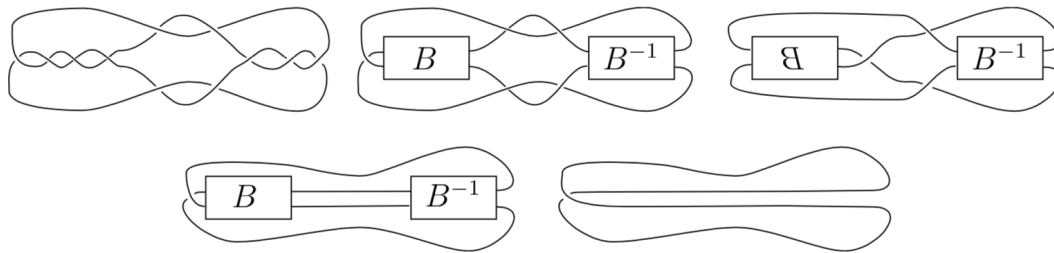
1. double the knot K (along the blackboard framing),
2. cut it open. We now have four ends coming in two pairs.
3. Take two mirror symmetric pairs of ends and stretch them out as two parallel strands,
4. take the remaining pair of ends and wrap them around those two strands before connecting them up,
5. mirror a second copy of this tangle and build the connect sum in the obvious way.

Figure 1 shows the construction for the trefoil knot K . We were able to show, by brute force computation with **Regina** [1], that in the case of the trefoil, this construction gives an unknot that requires 3 extra crossings to be unknotted. The constructions for more complex knots K are under investigation.

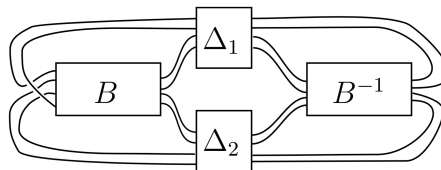
Generalizing the Goeritz unknot

A classical example of a hard unknot due to Goeritz is pictured in Figure 2 on the top left. It can be thought of (see the rest of Figure 2) as the concatenation of two inverse braids with two flypes inserted in between on both of its strands. Undoing these flypes requires turning the braid which can be shown experimentally to require at least one additional crossing.

We generalized this approach to braids with more strands in the braids, in order to increase the number of additional crossings needed. The framework pictured in Figure 3 seems promising, where Δ_1 and Δ_2 are the analogues of flypes, and can be readily generalized to higher number of strands.



■ **Figure 2** The Goeritz unknot.

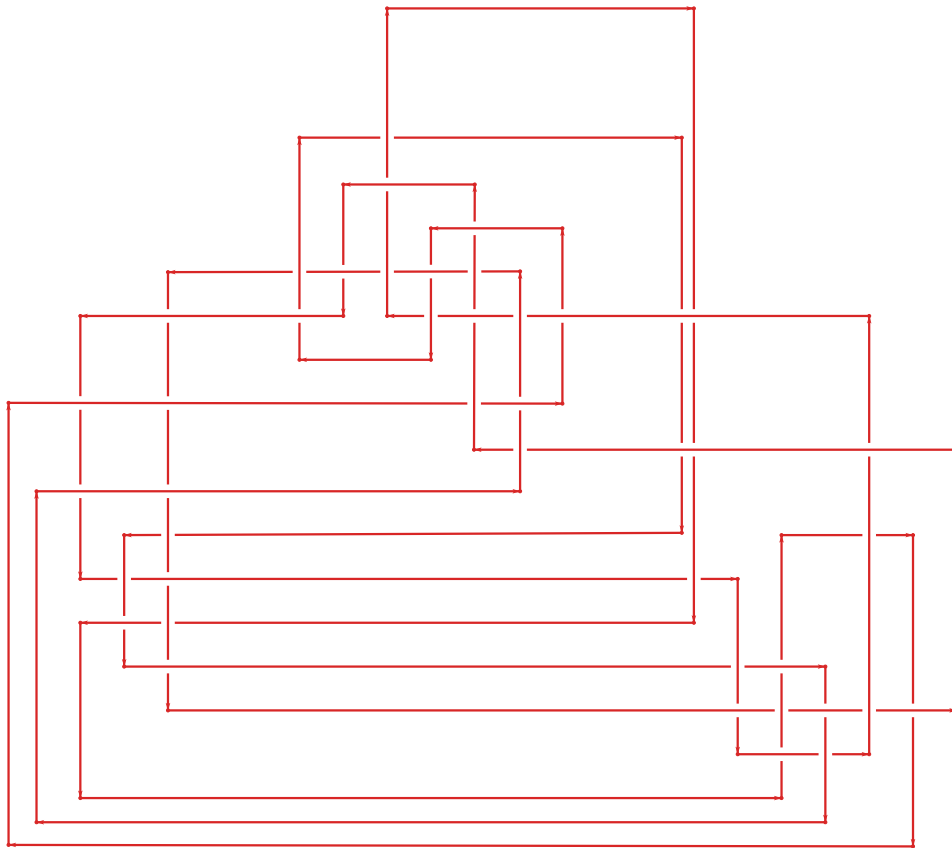


■ **Figure 3** Harder unknots.

An exhaustive computer search with **Regina** [1] shows that for a good choice of a four-strand pseudo-Anosov braid B , more than two additional crossings are required to simplify such unknots. Figure 4 pictures such an example. Proving lower bounds for generalizations of this example as the number of strands goes to infinity is the object of ongoing work.

References

- 1 B. A. BURTON, R. BUDNEY, W. PETERSSON, ET AL., *Regina: Software for low-dimensional topology*. <http://regina-normal.github.io/>, 1999–2017.
- 2 A. COWARD AND M. LACKENBY, *An upper bound on Reidemeister moves*, *American Journal of Mathematics*, 136 (2014), pp. 1023–1066.
- 3 Marc Culler, Nathan M. Dunfield, Matthias Goerner, and Jeffrey R. Weeks. SnapPy, a computer program for studying the geometry and topology of three-manifolds. Available at <http://snappy.computop.org> (2019-11-18).
- 4 I. A. DYNNIKOV, *Arc-presentations of links: Monotonic simplification*, *Fund. Math.*, 190 (2006), pp. 29–76.
- 5 D. EPPSTEIN, *Approaches to gi inspired by knot problem*. Theoretical Computer Science Stack Exchange. URL: <https://cstheory.stackexchange.com/q/12792> (version: 2012-10-02).
- 6 M. H. FREEDMAN, Z.-X. HE, AND Z. WANG, *Möbius energy of knots and unknots*, *Annals of Mathematics*, 139 (1994), pp. 1–50.
- 7 L. H. KAUFFMAN AND S. LAMBROPOULOU, *Hard unknots and collapsing tangles*, *Introductory lectures on knot theory*, Ser. Knots Everything, 46 (2012), pp. 187–247.
- 8 M. LACKENBY, *Elementary knot theory*, *Lectures on Geometry*, Oxford University Press, (2017).
- 9 M. LACKENBY, *A polynomial upper bound on Reidemeister moves*, *Annals of Mathematics*, 182 (2015), pp. 491–564.
- 10 M. LACKENBY, *Some conditionally hard problems on links and 3-manifolds*, *Discrete & Computational Geometry*, 58 (2017), pp. 580–595.



■ **Figure 4** An example of a hard(er) unknot. Figure created with SnapPy [3].

4.3 Combinatorial Homotopies

Hsien-Chih Chang (Duke University – Durham, US), Vincent Despré (INRIA Nancy – Grand Est, FR), Linda Kleist (TU Braunschweig, DE), Francis Lazarus (GIPSA Lab – Grenoble, FR), Anna Lubiw (University of Waterloo, CA), Tim Ophelders (Michigan State University, US), Hugo Parlier (University of Luxembourg, LU), Saul Schleimer (University of Warwick – Coventry, GB), Stephan Tillmann (University of Sydney, AU), Birgit Vogtenhuber (TU Graz, AT), Carola Wenk (Tulane University – New Orleans, US), and Erin Moriarty Wolf Chambers (St. Louis University, US)

License © Creative Commons BY 3.0 Unported license

© Hsien-Chih Chang, Vincent Despré, Linda Kleist, Francis Lazarus, Anna Lubiw, Tim Ophelders, Hugo Parlier, Saul Schleimer, Stephan Tillmann, Birgit Vogtenhuber, Carola Wenk, and Erin Moriarty Wolf Chambers

Main reference J. Erickson and K. Whittelsey. Transforming curves on surfaces redux. Proc. 24th ACM-SIAM Symp. Discrete Alg. (SODA), 2013

URL <https://epubs.siam.org/doi/10.1137/1.9781611973105.118>

A fundamental problem in computational topology is to test if a given loop in a space is contractible, or more generally if two loops are homotopic. The problem is known to be intractable in general but has a relatively simple solution when restricting the space to a 2 dimensional surface: there is a linear time algorithm that takes as input a combinatorial description of the surface and one or two closed paths in its 1-skeleton graph [1, 3]. This algorithm was implemented (and will be available in the next release of CGAL) and can be

considered as a black box that may be used for more complex algorithms. Suppose that we want to convince an outside observer that two curves are indeed homotopic or contractible, without just relying on the answer returned by the above algorithm. It seems that the most intuitive certificate is to exhibit an actual (combinatorial) homotopy. Such a homotopy can be decomposed into elementary homotopies that consist of

- adding or removing a spur,
- replacing a piece of a facial walk by the complementary piece.

Elementary homotopies directly translate into PL homotopies that can be further visualized.

The problem is thus to design an efficient algorithm that outputs a sequence of elementary homotopies.

During this Dagstuhl seminar, we resolved the problem by providing a simple optimal algorithm. The idea is to simulate the linear time algorithm in [1]. Let S be an input combinatorial surface of genus g and let G be its 1-skeleton (vertex-edge graph). In [1, 3] a preprocessing step is to transform S into a quadrangulation Q composed of 2 vertices and $4g$ edges. Any walk γ in G can then be transformed into a homotopic walk c in Q of length at most $2|\gamma|$. Such a walk c admits a canonical representative in its homotopy class that corresponds to the shortest and rightmost homotopic walk in Q . In [1], this canonical representative is obtained by producing a sequence of elementary homotopies for c in Q . It remains to translate those elementary homotopies in G . We propose the following algorithm.

We first compute a representation of Q in G where each edge of Q is represented by a path in G . For this, we compute a system of loops in $O(gn)$ time ($n = |G|$) using a spanning tree and $2g$ additional chordal edges. One can also compute a shortest system of loops in $O(n \log n + gn)$ time as described in [2]. Note that each edge of G appears at most twice in a loop and so at most $4g$ times in the whole system of loops. We then cut S through the loops. We get a polygonal schema with the same number of faces as G , although the boundary has size $O(gn)$. The basepoint of the loops appears $4g$ times along this boundary. Pick any vertex in this polygonal schema and join it to the $4g$ copies of the basepoint using shortest paths. We make those $4g$ paths correspond to the edges of the system of quads. Let Q' be the union of those paths. Q' cuts S into $2g$ quadrilateral regions, each comprising $O(n)$ faces of G .

We next “push” the given walk γ into Q' , thus getting a homotopic curve γ' (corresponding to the above c) composed of at most $2|\gamma|$ concatenations of the paths in Q' . We finally simulate the moves in Q by sweeping quadrilateral regions. We thus obtain a canonical representative in G that can be further homotoped to a curve δ , applying the reverse sequence of moves that brings δ to the canonical representative.

Analysis: Sweeping a single quad costs $O(n)$ elementary homotopies. We observe by a simple analysis of the algorithm in [1] that the total length (number of quads) swept by the canonisation of c in Q is $O(|c|) = O(|\gamma|)$. Indeed, referring to the terminology in [1], it appears that the shortcut part of a bracket cannot be part of another bracket oriented the same way as this would imply a degree 4 vertex in Q . The linear bound on the number of swept quads easily follows. We thus get a sequence of $O(n|\gamma|)$ elementary homotopies to obtain a “canonical” form for γ in G . This number of elementary homotopies is tight: if the surface S has a subpart with the shape of a big “mushroom” and if γ surrounds its foot many times, we have to sweep over that mushroom that many times, enforcing a sequence of $\Omega(n|\gamma|)$ moves. A similar lower bound can be obtained from two homotopic walks winding around the ends of a long cylinder in S .


The above observation seems to imply that the computation of a canonical form does not require the run length encoding of the curve used in the paper by Erickson and Whittlesey [1]. A simple forward traversal of the curve should allow to shorten it without backtracking more than once when we shortcut brackets.

References

- 1 J. ERICKSON AND K. WHITTELEY, *Transforming curves on surfaces redux.*, Proc. 24th ACM-SIAM Symp. Discrete Alg. (SODA), 2013, pp. 1646–1655.
- 2 J. ERICKSON AND K. WHITTELEY, *Greedy optimal homotopy and homology generators.*, Proc. 16th ACM-SIAM Symp. Discrete Alg. (SODA), 2005, pp. 1646–1655, pp. 1038–1046.
- 3 F. LAZARUS AND J. RIVAUD, *On the homotopy test on surfaces.*, Proc. 53rd IEEE Symp. Found. Comput. Sci. (FOCS), 2012, pp. 440–449.

4.4 Simple Graph Cycle in Homotopy Class

Hsien-Chih Chang (Duke University – Durham, US), Arnaud de Mesmay (University of Grenoble, FR), Vincent Despré (INRIA Nancy – Grand Est, FR), Francis Lazarus (GIPSA Lab – Grenoble, FR), and Erin Moriarty Wolf Chambers (St. Louis University, US)

License  Creative Commons BY 3.0 Unported license
© Hsien-Chih Chang, Arnaud de Mesmay, Vincent Despré, Francis Lazarus, and Erin Moriarty Wolf Chambers

Problem definition

Let Σ be a compact surface and G be a graph embedded in Σ . Let $\gamma = \{\gamma_1, \dots, \gamma_k\}$ be a collection of closed curves on Σ . Is there a polynomial-time algorithm to compute a set of simple closed walks $C = \{C_1, \dots, C_k\}$ in G , such that each C_i is homotopic to γ_i ?

Result

It turns out that Schrijver [1] had studied this problem before and obtained the following characterization, which is conjectured by Lovász and Seymour [2, Section 76.7]. So instead of reporting on our discussions, we present a sketch of Schrijver’s proof.

Given an embedded graph G and two closed curves γ and δ on Σ , define $\mathbf{cr}(G; \delta)$ to be the number of crossings between G and δ , and $\mathbf{mincr}(\gamma; \delta)$ be the minimum number of crossings between γ' and δ' , where γ' is homotopic to γ and δ' is homotopic to δ . A closed curve δ is **doubly-odd** with respect to G and multicurve γ if $\delta = \delta_1 \cdot \delta_2$ for some closed curves δ_1 and δ_2 , satisfying

$$\begin{aligned} \mathbf{cr}(G; \delta_1) &\not\equiv \sum_i \mathbf{cr}(\gamma_i, \delta_1) \pmod{2}, \text{ and} \\ \mathbf{cr}(G; \delta_2) &\not\equiv \sum_i \mathbf{cr}(\gamma_i, \delta_2) \pmod{2}. \end{aligned}$$

The **double point** of δ is equal to $\delta_1(0) = \delta_2(0)$.

► **Theorem 1** (Schrijver [1]). *Let Σ be a compact surface and G be a graph embedded in Σ . Let $\gamma = \{\gamma_1, \dots, \gamma_k\}$ be a collection of closed curves on Σ . There are simple closed walks $C = \{C_1, \dots, C_k\}$ in G where each C_i is homotopic to γ_i if and only if the all the following properties hold:*

1. there are pairwise disjoint simple closed curves $\gamma'_1, \dots, \gamma'_k$ on surface Σ (not necessarily in G),
2. for each closed curve δ on Σ ,

$$\text{cr}(G; \delta) \geq \sum_i \text{mincr}(\gamma_i, \delta); \text{ and}$$

3. for each doubly-odd closed curve $\delta = \delta_1 \cdot \delta_2$ whose double point is not on G ,

$$\text{cr}(G; \delta) > \sum_i \text{mincr}(\gamma_i, \delta).$$

In the book by Schrijver [2, Section 76.7] it is also claimed that such a characterization leads to a polynomial-time algorithm. However this is not immediately clear from the characterization. We give a short exposition on Schrijver’s proof using modern language in discrete and computational topology.

Sketch of proof

Necessity. First we prove the easy direction that the three conditions in Theorem 1 are all necessary. Let C_1, \dots, C_k be the simple closed walks in G promised by the theorem. Condition 1 is satisfied easily by taking $\gamma'_1, \dots, \gamma'_k$ to be C_1, \dots, C_k . Condition 2 is also straightforward:

$$\text{cr}(G; \delta) \geq \sum_i \text{cr}(C_i, \delta) \geq \sum_i \text{mincr}(\gamma_i, \delta),$$

where the second inequality follows by the fact that C_i is homotopic to γ_i for all i . As for Condition 3, let $\delta = \delta_1 \cdot \delta_2$ be any doubly-odd closed curve whose double point is not on G . For both $j \in \{1, 2\}$, one has $\text{cr}(G; \delta_j) \geq \sum_i \text{cr}(C_i, \delta_j)$ immediately, and the inequality must be strict because

$$\text{cr}(G; \delta_j) \not\equiv \sum_i \text{cr}(\gamma_i, \delta_j) \equiv \sum_i \text{cr}(C_i, \delta_j) \pmod{2}$$

because δ is doubly-odd and parity of number of intersections is unchanged under homotopy. Thus,

$$\begin{aligned} \text{cr}(G; \delta) &= \text{cr}(G; \delta_1) + \text{cr}(G; \delta_2) \\ &> \sum_i \text{cr}(C_i, \delta_1) + \sum_i \text{cr}(C_i, \delta_2) \\ &= \sum_i \text{cr}(C_i, \delta) \\ &\geq \sum_i \text{mincr}(C_i, \delta). \end{aligned}$$

Sufficiency. Here we present the proof when γ is a single closed curve and the goal is to find one simple closed walk C ; the proof for the general case is a rather straightforward extension. Without loss of generality we can assume that Σ is orientable and not topologically a sphere [1, Claim 2], and G is cellularly embedded in Σ [1, Claim 1]. We can also safely assume that the closed curve δ in conditions of Theorem 1 does not cross G or its dual G^* at the edges. This implies that $\text{cr}(G; \delta) = \text{cr}(G^*; \delta)$. We can recast the problem to the dual graph G^* ; the existence of disjoint simple closed walk C in G is equivalent to the following:

There is a simple closed curve C^* not intersecting the vertices of G^* , such that C^* is homotopic to γ and each face of G^* is traversed by C^* at most once.

Now we describe how to construct the set of simple closed walk C^* satisfying the above properties. Let γ be the input closed curve. By Condition 1 we can assume without loss of generality that γ itself is a *simple* closed curve and not intersecting any vertices of G^* ; we also assume that γ has transverse intersections with G^* . In other words, γ is a *normal curve* with respect to G^* . Consider a face f of G^* together with the portion of γ intersecting f ; we call the collection of arcs from the intersection the *curve parts* of f . For each component κ of the curve parts, draw an infinitesimally short segment crossing κ in a way that all short segments are disjoint. We call the collection of endpoints of all short segments as *terminals* and denoted as T . The two endpoints of the same segment are called a *terminal pair*.

We are going to set up a linear program with variables $\psi(\cdot)$ indexed by T , and set the constraints properly so that a solution to the linear program gives us hint on how to find the correct homotopy to turn γ into C^* . Consider the following linear program:

$$\begin{array}{ll} \min & \sum_{t \in T} |\psi(t)| \\ \text{s.t.} & \psi(t) + \psi(\bar{t}) = 0 \quad \text{for each terminal pair } t \text{ and } \bar{t}, \\ & \psi(t) + \psi(t') \leq \lambda(tt') \quad \text{for each pair of } t \text{ and } t' \text{ in } T, \end{array}$$

where $\lambda(tt') := \min_{\pi} \text{cr}(G^*; \pi) - 1$, with the minimum taking over all paths π on Σ connecting t to t' , such that the two endpoints of the lift $\hat{\pi}$ of π in the universal cover $\hat{\Sigma}$ connects different lifts of γ , and $\hat{\pi}$ lies in the same component of $\hat{\Sigma}$ subtracting $\hat{\gamma}$. Schrijver showed that the existence of a solution to the above linear program is equivalent to the properties in Theorem 1 [1, Section 2 and Claim 3].

Now here comes the important definitions. Recall that $\hat{\Sigma}$ is the universal cover of Σ and everything with a hat is a lift of the corresponding object in Σ . Let $\hat{\Sigma}_{\gamma}$ be the *cyclic cover* of Σ with respect to γ . Lifts \hat{G} and $\hat{\gamma}$ are defined accordingly. Notice that $\hat{\gamma}$ is a simple closed curve in $\hat{\Sigma}$. The notion of terminals and function $\psi(\cdot)$ associated with the linear program can be lifted and defined with respect to the universal cover and cyclic cover as well.

For any lift ℓ of γ and any lift \hat{f} of a face f of G^* , define

$$\Pi_{\ell}(\hat{f}) := \min_{\hat{\pi}} (\text{cr}(\hat{G}^*; \hat{\pi}) - \psi(\pi(0))),$$

where $\hat{\pi}$ ranging over all paths in Σ that starts at some lift of a terminal corresponding to a curve part from ℓ that projects to $\pi(0)$ and ends in \hat{f} , and crosses ℓ an even number of times. Notice by definition if some \hat{f} intersects ℓ then $\Pi_{\ell}(\hat{f}) \leq 0$. Define *zero (dual) faces* $F_0(\ell)$ to be the collection of faces \hat{f} in \hat{G}^* with $\Pi_{\ell}(\hat{f}) = 0$. Define *non-positive (dual) vertices* $V_{\leq 0}$ to be the collection of vertices v of \hat{G}^* that has a path $\hat{\pi}$ starting from a terminal in $\hat{\Sigma}$ that projects to $\pi(0)$ and ending at v that crosses $\hat{\gamma}$ an even number of times, such that

$$\text{cr}(\hat{G}^*; \hat{\pi}) - \psi(\pi(0)) \leq 0.$$

Based on the upper bound on $\text{cr}(\hat{G}^*; \hat{\pi})$, the number of non-positive vertices in $V_{\leq 0}$ must be finite.

Equipped with these definitions we are now ready to describe the construction of C . Let $[\hat{\gamma}]$ be the representative of the *homology class* of $\hat{\gamma}$ over \mathbb{Z}_2 ; as γ does not intersect vertices of G^* , $\hat{\gamma}$ can be viewed as a closed walk in \hat{G} . Here we abuse the type and refer to $[\hat{\gamma}]$ as a subset of edges of \hat{G} . Define E_0 to be the symmetric difference between $[\hat{\gamma}]$ and the (dual)

edge cut formed by $V_{\leq 0}$ (which is a collection of cycles, a valid homology class). Among the simple cycles of E_0 in \hat{G} , there is at least one simple cycle \hat{C} homotopic to $\hat{\gamma}$ on $\hat{\Sigma}$ [1, Claim 6]. Project \hat{C} back to Σ to obtain the desired cycle C in G . Important properties of these definitions are

1. if \hat{C} passes through some face \hat{f} in \hat{G}^* , then any lift of \hat{f} must be in $F_0(\ell)$ [1, Claims 6 and 7]; and
2. each pair of $F_0(\ell)$ and $F_0(\ell')$ are disjoint if $\ell \neq \ell'$ [1, Claim 5].

Finally it is sufficient to show that C is indeed simple (or, vertex-disjoint); in other words, C^* never passes through any dual face in G^* more than once. The construction guarantees that \hat{C} is simple, so it must be the case that two faces \hat{f} and \hat{g} of \hat{G}^* passed by \hat{C} projects to the same face in G^* . Now $F_0(\ell)$ must contain both \hat{f} and \hat{g} by Property (1). Consider the deck transformation ϕ on $\hat{\Sigma}$ that maps face \hat{f} to \hat{g} . Now $\phi(\ell)$ is another lift of γ . Since $\hat{g} = \phi(\hat{f}) \in \phi(F_0(\ell)) = F_0(\phi(\ell))$ as well, by Property (2) above ℓ and $\phi(\ell)$ must be the same lift of γ . This implies that after projecting to $\hat{\Sigma}$ the two faces \hat{f} and \hat{g} must be identical. Therefore, C^* never passes through any dual face in G^* more than once.

Efficient implementation

To carry out a polynomial-time algorithm, a few questions remain.


- How do we compute every $\lambda(tt')$ in polynomial-time to set up the linear program?
- How to compute cycles \hat{C} and C ? In particular, how to construct non-positive vertices $V_{\leq 0}$?

References

- 1 Alexander Schrijver. Disjoint circuits of prescribed homotopies in a graph on a compact surface. *Journal of Combinatorial Theory, Series B* 51(1):127–159. Elsevier, 1991.
- 2 Alexander Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Algorithms and Combinatorics 24. Springer-Verlag, 2003.

4.5 Beautiful curves on beautiful surfaces

Saul Schleimer (University of Warwick – Coventry, GB), Vincent Despré (INRIA Nancy – Grand Est, FR), Francis Lazarus (GIPSA Lab – Grenoble, FR), Hugo Parlier (University of Luxembourg, LU), Stephan Tillmann (University of Sydney, AU), and Erin Moriarty Wolf Chambers (St. Louis University, US)

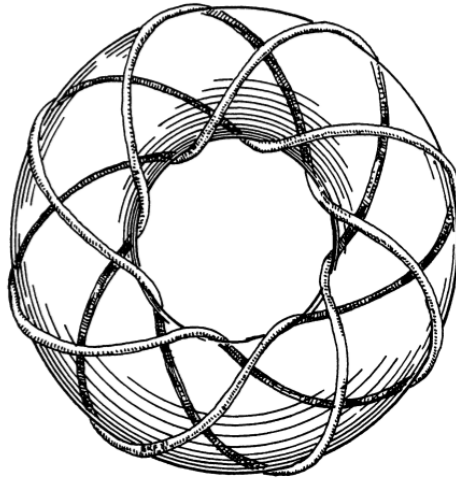
License  Creative Commons BY 3.0 Unported license
© Saul Schleimer, Vincent Despré, Francis Lazarus, Hugo Parlier, Stephan Tillmann, and Erin Moriarty Wolf Chambers

We considered the problem of how to draw “beautiful curves on beautiful surfaces”.

That is, suppose we are given a closed connected oriented surface S in the form of a very nice mesh in \mathbb{R}^3 . Suppose that we are also given a simple closed curve α in S , say as the (circular) sequence of mesh edges it crosses. We wish to find a “beautiful” representative of α .

Curves on the torus

For the torus, we can argue that the Clifford torus (stereographically projected from the three-sphere) is the “best” geometric torus. Now, for any curve α there is an optimal position,



■ **Figure 5** From page 38 of *A topological picturebook* by George Francis [1].



■ **Figure 6** From page 421 of *On the geometry and dynamics of diffeomorphisms of surfaces* by William Thurston [2].

making α both as straight as possible and also as well-spaced as possible. To see this, consider the Clifford torus in S^3 and stereographically project it to \mathbb{R}^3 . Since the Clifford torus is square, there are “obvious” representatives (straight, going through the origin). These are best both in terms of straightness and well-spacedness.

Higher genus

In higher genus things are less clear-cut. It is much harder to argue for a particular “best” geometric surface. Note that in the picture by Thurston the surface is a bit pinched along the outer two non-separating curves.

It is also now much harder to decide what constitutes a “best” representative of a given isotopy class of curve. Thurston’s curve shows some of the issues involved, especially when compared to the standard short curves. Note that in Thurston’s curve it is not so easy to “see” the number of components, or the topological type.

Also, Thurston has preferred to make the curve “well-spaced” rather than “straight”. This gives a few places where the geodesic curvature is “too high”. Thus we see that there is a tension between drawing the curve straight and having good “interarc” spacing. The former takes us towards geodesic laminations, the latter towards quadratic differentials.

Teruaki

One implemented solution (in genus two) is given by Kazushi Ahara as part of his game Teruaki exploring the mapping class group.

<http://www.aharalab.sakura.ne.jp/teruaki.html>

This is makes for a very interesting game! Note that if you perform large powers of Dehn twists, then the programme gets confused; curves may crash into themselves and may also drift off of the surface.

Plans of attack

We can think about two ways to approach the problem. Fix a good triangular mesh Δ of the surface $S \subset \mathbb{R}^3$. Our first approach concentrates on making the curves straight.

Hyperbolic geometry: Cut open Δ ; this gives a fundamental domain D . Lay D out conformally in the hyperbolic plane. Compute the resulting Fuchsian group. Given an element of $\pi_1(S)$ draw all of its intersections D . Push these forward to $\Delta \subset \mathbb{R}^3$.

Our second approach concentrates on making the curves well-spaced.

Quadratic differentials: For any simple closed curve $\alpha \subset S$, the *Strebel differential* $q = q_\alpha$ is the unique quadratic differential in the conformal class of S , of area one, so that all non-singular flowlines are isotopic to α . So the periods of q are given and we must solve for the differential. Now compute the “core-curve” of the maximal cylinder in q .

References

- 1 George K. Francis. *A topological picturebook*. Springer, New York, 2007. Reprint of the 1987 original.
- 2 William P. Thurston. On the geometry and dynamics of diffeomorphisms of surfaces. *Bull. Amer. Math. Soc. (N.S.)*, 19(2):417–431, 1988.

4.6 Minimum Area Homotopies

Carola Wenk (Tulane University – New Orleans, US), Hsien-Chih Chang (Duke University – Durham, US), Vincent Despré (INRIA Nancy – Grand Est, FR), Francis Lazarus (GIPSA Lab – Grenoble, FR), Anna Lubiw (University of Waterloo, CA), Tim Ophelders (Michigan State University, US), Hugo Parlier (University of Luxembourg, LU), and Erin Moriarty Wolf Chambers (St. Louis University, US)

License © Creative Commons BY 3.0 Unported license

© Carola Wenk, Hsien-Chih Chang, Vincent Despré, Francis Lazarus, Anna Lubiw, Tim Ophelders, Hugo Parlier, and Erin Moriarty Wolf Chambers

Given a closed curve γ embedded in the plane, can we compute a “nice” homotopy in polynomial time that minimizes the homotopy area contracting γ to a point?

Chambers and Wang have introduced the notion of minimum-area homotopies as a distance measure between two curves [1]. They presented a polynomial-time dynamic programming algorithm for the case that the winding numbers, induced by the closed curve composed of the two open curves, are consistent (i.e., all non-negative or all non-positive). For the general case, Nie has presented a polynomial-time algorithm [2] that is algebraic in nature and results in homotopies with degeneracies. Recently it has been shown that one

can express Nie’s algorithm entirely using such (degenerate) homotopies that either collapse a face or cancel around a face [3]. On the other hand, it has been shown that there is always a minimum-area homotopy that can be represented as a decomposition of self-overlapping curves [4]. Contracting each self-overlapping curve in such a decomposition to a point then results in a “nice” homotopy. However, the algorithm in [4] to compute such a decomposition is exponential. A related paper on combinatorial properties of self-overlapping curves and interior boundaries [5], that introduces the notions of obstinance and wrapping, helps in understanding the relationships between minimum area homotopies and self-overlappingness of curves in the plane.

Some of the insights and related problems discussed during the Dagstuhl seminar are as follows:

1. Can we check whether a curve is self-overlapping in less than $O(N^3)$ time? Here, N is the number of vertices of a polygonal input curve. Shor and Van Wyk’s dynamic programming algorithm [6] runs in $O(N^3)$ time.
2. If we know how many times a face is swept in a minimum-area homotopy (e.g., from Nie’s algorithm), can we compute an area-optimal self-overlapping curve decomposition? Alternatively, is computing an area-optimal self-overlapping decomposition NP-hard?

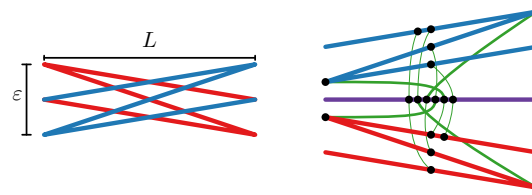
Insights:

- Blank’s algorithm [7] for computing whether a curve is self-overlapping runs in time quadratic to the depth-sum of the curve, and is therefore faster than Shor and van Wyk’s for shallow curves.
- All $2 \rightarrow 0$ moves and two of the three $0 \rightarrow 2$ moves preserves self-overlappingness; the last $0 \rightarrow 2$ move requires the local potentials to be positive.
- Any tree-like self-overlapping curve [8, 9] is simple. A possible strategy: First add a “shell” to the input curve, then use regular homotopy to shrink the curve. With enough shelling all $2 \leftrightarrow 0$ moves can be executed safely.

Other questions that were discussed were how to decide self-overlappingness of a curve that is embedded on a surface, for a example on a sphere or on a higher genus surfaces. For curves in the plane it was discussed whether the casings by Eppstein and Mumford [10] help in determining that a curve is self-overlapping. In this case a 2-move must stay consistent, both over-over or under-under; is there a relation to Morse theory to find a height function?

References

- 1 E.W. Chambers and Y. Wang. Measuring similarity between curves on 2-manifolds via homotopy area. *ACM Sympos. Comput. Geom. (SoCG)*, 2013.
- 2 Z. Nie. On the Minimum Area of Null Homotopies of Curves Traced Twice. *arXiv: 1412.0101*, 2014.
- 3 B.T. Fasy, B. McCoy, D. Millman, C. Wenk. A Geometric Interpretation of the Cancellation Norm. In preparation, 2019.
- 4 B.T. Fasy, S. Karakoc, C. Wenk. On Minimum Area Homotopies of Normal Curves in the Plane. *arXiv: 1707.02251*, 2017.
- 5 P. Evans and C. Wenk. Combinatorial Properties of Self-Overlapping Curves Through Minimum Homotopy Area. In submission, 2019.
- 6 P.W. Shor and C.J. Van Wyk. Detecting and decomposing self-overlapping curves. *Computational Geometry: Theory and Applications* 2(1): 31-50, 1992.
- 7 Samuel J. Blank. Extending Immersions and regular Homotopies in Codimension 1. *Ph.D. Dissertation, Brandeis University*, 1967.
- 8 V. Arnold. Plane curves, their invariants, perestroikas and classifications. *Advances in Soviet Mathematics* 21, 1994.



■ **Figure 7** The optimal isotopy for the simple zigzag. (The green trajectories are actually straight-line; they are drawn as curves for better visibility.) For further details see [7].

- 9 F. Aicardi. Tree-like curves. *Advances in Soviet Mathematics* 21, 1994.
 10 D. Eppstein and E. Mumford. Self-overlapping curves revisited. *20th ACM-SIAM Symp. Discrete Algorithms*: 160-169, 2009.

4.7 Nice Morphs and Isotopic Fréchet Distance

Erin Moriarty Wolf Chambers (St. Louis University, US), Vincent Despré (INRIA Nancy – Grand Est, FR), Linda Kleist (TU Braunschweig, DE), Maarten Löffler (Utrecht University, NL), Anna Lubiw (University of Waterloo, CA), Tim Ophelders (Michigan State University, US), Hugo Parlier (University of Luxembourg, LU), Stephan Tillmann (University of Sydney, AU), Birgit Vogtenhuber (TU Graz, AT), and Carola Wenk (Tulane University – New Orleans, US)

License © Creative Commons BY 3.0 Unported license

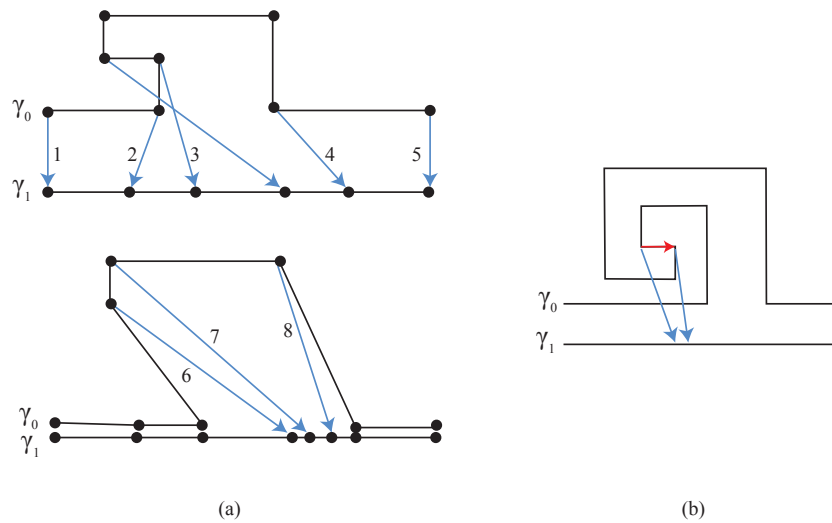
© Erin Moriarty Wolf Chambers, Vincent Despré, Linda Kleist, Maarten Löffler, Anna Lubiw, Tim Ophelders, Hugo Parlier, Stephan Tillmann, Birgit Vogtenhuber, and Carola Wenk

Introduction

Fréchet distance is a way of measuring the distance between two simple (i.e., non-self-intersecting) curves γ_0 and γ_1 in the plane. Informally, we want the minimum leash length that allows a person to travel along γ_0 while their dog travels along γ_1 . Either one may vary their speed but neither may back up. There is a polynomial time algorithm to compute Fréchet distance between two curves [2]. Any travel plan that realizes the minimum leash length yields a *Fréchet matching* between the points of γ_0 and the points of γ_1 . This in turn determines a continuous deformation from γ_0 to γ_1 , as each point moves along the taut leash (say at uniform speed) between the matched points. However, in the course of this continuous deformation, the curve will not remain simple in general.

Adding the further constraint that the curve should remain simple throughout the deformation yields the notion of *isotopic Fréchet distance*, which can be defined as follows. A *morph* from γ_0 to γ_1 is a continuous family of simple curves γ_t , indexed by time t , $0 \leq t \leq 1$ so that the curve at time $t = 0$ is γ_0 and the curve at time $t = 1$ is γ_1 . A morph determines a trajectory $p(t)$, $0 \leq t \leq 1$ of each point $p = p(0)$ in γ_0 to its destination point $p(1)$ in γ_1 . The minimum over all morphs of the maximum trajectory length is the *isotopic Fréchet distance*, first introduced in [3]. An example is shown in Figure 7.

We explored the problem of finding a “nice” morph between γ_0 and γ_1 to realize or approximate the isotopic Fréchet distance.



■ **Figure 8** (a) γ_0 morphs to γ_1 as vertices travel on the blue trajectories in the order specified; (b) no morph is possible for these trajectories because the segment of γ_0 drawn in red must rotate 360° degrees but the trajectories do not allow that.

Related Work

When the two curves do not cross each other but share the same start and end points, they determine a simple polygon. Efrat et al. [4] gave an algorithm for the version of Fréchet distance where the leash must remain inside the polygon. In this situation, the leash paths will not cross each other so the Fréchet mapping yields a morph between the curves.

Another relevant related problem is to morph between two simple polygons while preserving edge lengths as much as possible [5]. For results on morphing planar graphs, see [1] and references therein.

Morphing along the Fréchet leashes

One approach we explored was to morph by moving each vertex along its Fréchet “leash” but not necessarily at uniform speed. Given a mapping (perhaps not even a Fréchet mapping) between γ_0 and γ_1 , subdivide γ_0 and γ_1 so that each vertex of one maps to a vertex of the other. This gives a set of “leashes”, each one a straight line segment, between [the expanded set of] vertices of γ_0 and γ_1 . Now we try to move each vertex along its leash path so that the curve remains simple at all times, using the freedom that vertices need not travel at uniform speed.

This is not always possible. However, in the special case when both curves are x -monotone (increasing) then it is possible as shown by Tim Ophelders in his thesis [7]. We examined the case where the two curves do not intersect and found a counter-example—see Figure 8.

It would be interesting to prove NP-hardness for the decision problem (can vertices be moved on the given leash paths so that the curve remains simple). One related result is that it is NP-complete to decide if we can morph some segments from initial to final positions with the restriction that the segments never cross and every vertex moves along a straight line from initial to final position [10, Chapters 6,7]. As above, the freedom is that vertices need not travel at uniform speeds.

A lower bound

Suppose a subpath ab of γ_0 should morph to a subpath $a'b'$ of γ_1 . It may need to turn 360° (as in Figure 8), or some multiple $k \cdot 360^\circ$. Imagine morphing the subpath ab to a point x , rotating it by $k \cdot 360^\circ$, and then morphing to $a'b'$. To minimize the maximum trajectory length, we should choose x so that $\max(|ax| + |xa'|, |bx| + |xb'|)$ is minimized. We show that this provides a lower bound on the isotopic Fréchet distance.

Let α and β be the paths from a to a' and from b to b' obtained from an optimal morph, and assume that the path $t \mapsto \beta(t) - \alpha(t)$ winds (clockwise or counterclockwise) around the origin for at least 180° . We claim that the cost of the morph is at least the minimum value of $\max(|ax| + |xa'|, |bx| + |xb'|)$ over all x . The path α lies inside the ellipse A with foci a and a' and a major axis of length $\|\alpha\|$, and β lies inside the ellipse B with foci b and b' and a major axis of length $\|\beta\|$. For any $x \in A \cap B$, we have $\|\alpha\| \geq |ax| + |xa'|$ and $\|\beta\| \geq |bx| + |xb'|$, so it suffices to show that A intersects B . Since the morph is an isotopy, we have $\alpha(t) \neq \beta(t)$, so we can define $\theta(t)$ to be the angle of the (directed) segment from $\alpha(t)$ to $\beta(t)$. Because of the winding, there exist t and t' such that $\theta(t) = \theta(t') + 180^\circ$. This defines a trapezoid with one diagonal connecting $\alpha(t)$ and $\alpha(t')$, and the other diagonal connecting $\beta(t)$ and $\beta(t')$. By convexity of A and B , the first diagonal lies in A and the second lies in B , and since diagonals of a trapezoid intersect, A intersects B .

We believe that there exist mappings between curves that result in an optimal morph that is more expensive than both the lower bound described above, and the (homotopic) Fréchet distance. In particular, the above lower bound essentially requires α to have only a single bend, namely infinitesimally close to the unique point x that minimizes $\max(|ax| + |xa'|, |bx| + |xb'|)$. It is likely that we can simultaneously force the α leash to bend at a unique point $y \neq x$ that minimizes $\max(|ay| + |ya'|, |cy| + |yc'|)$, which violates the lower bound. It is unclear whether this violation can actually be realized in an optimal Fréchet isotopy.

Three-dimensional approaches

If we view time as the third dimension then we have two curves, say γ_0 in the $z = 0$ plane and γ_1 in the $z = 1$ plane, and we wish to find a “nice” surface joining them. The morph is then obtained by slicing the surface in successive planes. For morphs of planar graphs, this interpretation is explicitly discussed by Surazhsky and Gotsman [9]. We discussed soap bubbles (fascinating but tricky) and PL-minimal surfaces [6].

One idea is to triangulate the top and bottom planes with many triangles, and then triangulate the region between them with “nice” tetrahedra. We need to then find a normal surface with boundary equal to the two curves, for which techniques as in [8] might be used. One complication is that the minimal surface that solves this discrete version of the Plateau problem may not be an annulus bounded by the two curves, but have higher genus. Thus, either one needs a triangulation that is closely adapted to the local geometry of the curves, or a more sophisticated clean-up step would be required in order to obtain the desired morph.

A piecewise linear approach

An alternative approach, coming from classical PL topology, is to relate the curves by a sequence of elementary moves. An elementary move replaces two edges of a triangle with the third, or vice-versa. The main problem with this approach is to determine an objective function defining a “nice morph”. One possibility is to ask for a minimal sequence of elementary collapses.

Conclusions


It seems that finding an isotopy (a simplicity preserving morph) that minimizes the maximum trajectory (i.e., isotopic Fréchet) is really at odds with finding a “nice” morph. This is because solutions to isotopic Fréchet shrink spiraling parts of the curve infinitesimally small and then unspiral them.

References

- 1 Soroush Alamdari, Patrizio Angelini, Fidel Barrera-Cruz, Timothy M Chan, Giordano Da Lozzo, Giuseppe Di Battista, Fabrizio Frati, Penny Haxell, Anna Lubiw, Maurizio Patrignani, Vincenzo Roselli, Sahil Singla, and Bryan T. Wilkinson. How to morph planar graph drawings. *SIAM J. Computing*, 46(2):29 pages, 2017.
- 2 Helmut Alt and Michael Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91, 1995.
- 3 Erin W. Chambers, David Letscher, Tao Ju, and Lu Liu. Isotopic Fréchet distance. In *CCCG*, 2011.
- 4 Alon Efrat, Leonidas J. Guibas, Sarel Har-Peled, Joseph S. B. Mitchell, and T. M. Murali. New similarity measures between polylines with applications to morphing and polygon sweeping. *Discrete & Computational Geometry*, 28(4):535–569, 2002.
- 5 Hayley N. Iben, James F O’Brien, and Erik D Demaine. Refolding planar polygons. *Discrete & Computational Geometry*, 41(3):444–460, 2009.
- 6 William Jaco, J. Hyam Rubinstein. PL minimal surfaces in 3-manifolds. *Journal of Differential Geometry*, 27(3):493–524, 1988.
- 7 Tim Ophelders. *Continuous Similarity Measures for Curves and Surfaces*. PhD thesis, Eindhoven University of Technology, Netherlands, 2018.
- 8 Jonathan Spreer. Normal surfaces as combinatorial slicings. *Discrete Mathematics*, 311(14):1295 – 1309, 2011.
- 9 Vitaly Surazhsky and Craig Gotsman. High quality compatible triangulations. *Engineering with Computers*, 20(2):147–156, 2004.
- 10 Hamideh Vosoughpour. *Straight Line Movement in Morphing and Pursuit Evasion*. PhD thesis, University of Waterloo, Canada, 2017.

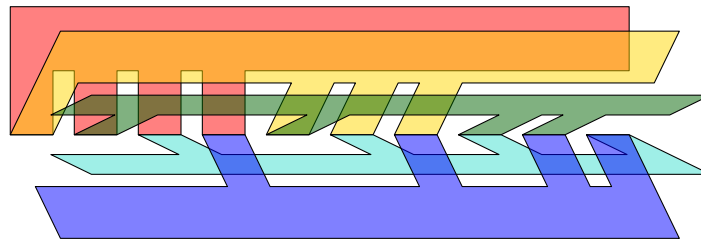
4.8 Representing Graphs by Polygons with Edge Contacts in 3D

Alexander Wolff (Universität Würzburg, DE), Elena Arseneva (St. Petersburg State University, RU), Arnaud de Mesmay (University of Grenoble, FR), Linda Kleist (TU Braunschweig, DE), Boris Klemz (FU Berlin, DE), Maarten Löffler (Utrecht University, NL), André Schulz (FernUniversität in Hagen, DE), and Birgit Vogtenhuber (TU Graz, AT)

License  Creative Commons BY 3.0 Unported license
 © Alexander Wolf, Elena Arseneva, Arnaud de Mesmay, Linda Kleist, Boris Klemz, Maarten Löffler, André Schulz, and Birgit Vogtenhuber

Evans et al. [8] showed that every graph has a contact representation in 3D in which each vertex is represented by a (flat) convex polygon, two polygons touch *in a corner* if and only if the corresponding two vertices are adjacent, and the interiors of any two polygons are disjoint. In this short note, we investigate representations where polygons that correspond to adjacent vertices must share an edge, rather than a corner.

We first allow our polygons to be nonconvex. In this case, we can easily represent every graph; see Figure 9.



■ **Figure 9** A realization of K_5 by nonconvex polygons with edge contacts.



■ **Figure 10** The Szilassi polyhedron realizes K_7 by nonconvex polygons with edge contacts [3].

► **Observation 1.** Every graph can be realized by polygons with edge contacts in 3D.

Proof. We sketch how to obtain a realization. To represent a graph G with n vertices, we start with n rectangles such that the intersection of all these rectangles is a line segment s . We then cut away parts of each rectangle thereby turning it into a comb shaped polygon; see Figure 9. The goal of this step is to ensure that for each pair (P, P') of polygons, there is a subsegment s' of s such that s' is an edge of both P and P' that is disjoint from the remaining polygons. The result is a representation of K_n . To obtain a realization of G , it remains to remove edge contacts that correspond to unwanted adjacencies, which is easy. ◀

If we additionally insist that each polygon shares all of its edges with other polygons, the polygons describe a closed volume. In this model, K_7 can be realized as the Szilassi polyhedron; see Figure 10. The tetrahedron and the Szilassi polyhedron are the only two known polyhedra in which each face shares an edge with each other face [3]. Which graphs other than complete graphs can be represented by edge contacts of non-convex polygons where every edge must be shared, remains an open problem.

We now consider the setting where each vertex of the given graph is represented by a convex polygon in 3D and two vertices of the given graph are adjacent, if and only if their polygons have edge contact. (So far, most people have only insisted that the edge of one polygon is contained in the edge of the adjacent polygon. For example, Duncan et al. [1] showed that in this model every planar graph can be realized by hexagons in the plane and that hexagons are sometimes necessary.) Note that it is allowed to have edges that do not touch other polygons. We start with some simple observations.

► **Observation 2.** Every planar graph can be realized by convex polygons with edge contacts in 2D.

Proof. Let G be a planar (embedded) graph. Add to G a new vertex r and connect it to all vertices of some face. Let G' be a triangulation of the resulting graph. Then the dual of G' , G^* , is a cubic 3-connected planar graph. Using Tutte's barycentric method, draw G^* into a regular polygon with $\deg_{G'}(r)$ corners such that the face dual to r becomes the outer

face. Note that the interior faces in this drawing are convex polygons; the polygon that corresponds to a vertex v of G' has $\deg_{G'}(v)$ corners. To convert this contact representation of $G' - r$ into a contact representation of G , we may need to remove some edge contacts, which can be easily achieved.

The same can be shown as follows. Using the classical result of Koebe, take a contact representation of the given planar graph by touching disks. For each pair of touching disks, place a very short line segment on their common tangent such that the line segment is centered on the touching point. Then represent every vertex of the given graph by the convex hull of the line segments that touch its disk. If the line segments are short enough, every two of the resulting convex polygons are interior disjoint. By construction, the polygons of adjacent vertices share an edge. Each polygon has twice as many edges as the degree of the corresponding vertex. ◀

So for planar graphs vertex and edge contacts behave similarly. For nonplanar graphs (for which we need the third dimension), the situation is different. Here, edge contacts are more restrictive. We introduce the following notation. In a 3D representation of a graph G by polygons, we denote by p_v the polygon that represents vertex v of G .

► **Lemma 1.** *Let G be a graph. Consider a 3D edge-contact representation of G with convex polygons. If G contains a triangle uvw , polygons p_v and p_w lie on the same side of the plane that supports p_u .*

Proof. Due to the convexity of the polygons, p_v and p_w either both lie above or both lie below the plane that supports p_u , otherwise p_v and p_w cannot share an edge. In this case, the edge vw of G would not be represented; a contradiction. ◀

► **Observation 3.** For $n \geq 5$, K_n is not realizable by convex polygons with edge contacts in 3D.

Proof. Assume that K_n admits a 3D edge-contact representation. Since every three vertices in K_n are pairwise connected, by Lemma 1, for every polygon of the representation, its supporting plane has the rest of the complex on one side. In other words, the complex we obtain is a subcomplex of a convex polyhedron. Consequently, the dual graph has to be planar, which rules out K_n for $n \geq 5$. ◀

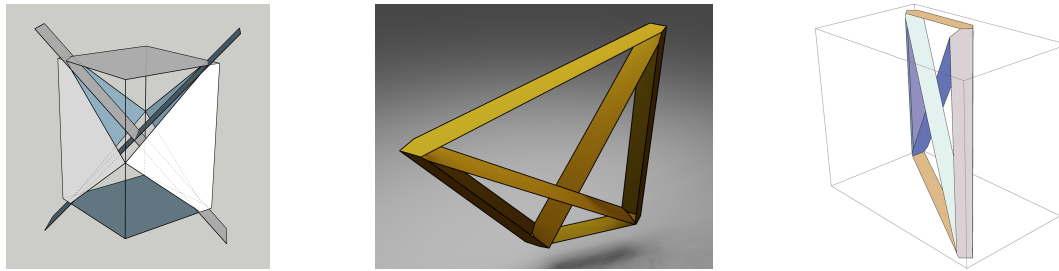
► **Observation 4.** $K_{4,4}$ is realizable by convex polygons with edge contacts.

Proof. We sketch how to obtain a realization. Start with a box in 3D and intersect it with two rectangular slabs as indicated in Figure 11 on the left. We can now draw polygons on the faces of this complex such that every vertical face contains a polygon that has an edge contact with a polygon on a horizontal or slanted face. The polygons on the slanted faces lie in the interior of the box and intersect each other. To remove this intersection we pull out one corner of the original box (see Figure 11). ◀

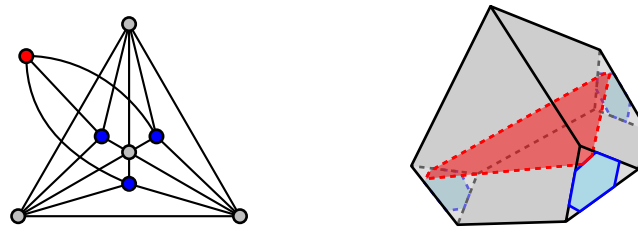
In contrast to Observation 4, we believe that the analogous statement does not hold for all bipartite graphs, i.e., we conjecture the following:

► **Conjecture 1.** There exist values n and m such that the complete bipartite graph $K_{m,n}$ is not realizable by convex polygons with edge contacts.

By Observation 2, all planar 3-trees can be realized by convex polygons with edge contacts (even in 2D). When switching to 3D, also nonplanar 3-trees can have a realization by convex polygons with edge contacts; see for example Figure 12. However, this is not the case for all nonplanar 3-trees.



■ **Figure 11** A realization of $K_{4,4}$ by convex polygons with edge contacts.



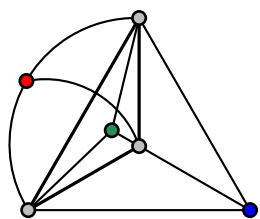
■ **Figure 12** A nonplanar 3-tree with a realization by convex polygons with edge contacts. The gray vertices form a K_4 .

► **Observation 5.** Not all 3-trees can be realized by convex polygons with edge contacts.

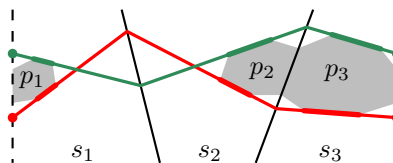
Proof. Consider the 3-tree in Figure 13, which consists of $K_{3,3}$ plus a cycle that connects the gray vertices of one part of the bipartition. The other part of the bipartition consists of three colored vertices (red, green, blue). For the sake of contradiction, assume that there is a representation by convex polygons with edge contacts and distinguish two cases: Either the three polygons are coplanar or not.

If the gray polygons are coplanar, then all edge contacts must lie in the same plane. This, however, contradicts the fact that $K_{3,3}$ is not planar.

If the three edges are not coplanar, the gray polygons form a prism-like shape. Note that every colored vertex together with the gray vertices forms a K_4 . Hence, by Lemma 1, all gray polygons must lie on one side of the supporting plane of a colored polygon. Each supporting plane of a colored polygon intersects the gray triangular prism in a triangle. Two of these triangles must intersect (otherwise one of the outer colored polygons would be cut off by the middle colored polygon and would not have any contact with the gray polygons). The two intersecting triangles (say, the red and the green) cross each other exactly twice, and the two points of intersection lie on two distinct sides s_1 and s_2 of the gray prism; see Figure 14. Let s_3 denote the side of the gray prism that does not contain any of these intersection points. Each of the two crossing triangles intersects s_3 in a line segment. These two line segments partition s_3 into three regions. For $i \in \{1, 2, 3\}$, let p_i denote the gray polygon that lies on side s_i . Polygon p_3 lies in the middle (bounded) region of s_3 , otherwise it cannot have an edge contact with both the red and the green polygon. The two crossing triangles partition each of the remaining two sides s_1 and s_2 into four regions. Two of these regions are unbounded; the other two are bounded and triangular. To realize edge contacts with p_3 , polygons p_1 and p_2 have to be located in the triangular region of s_1 and s_2 , respectively, that is adjacent to the middle region of s_3 . However, in this case p_1 and p_2 cannot possibly touch; a contradiction. ◀



■ **Figure 13** A 3-tree that is not realizable by convex polygons with edge contacts. The gray vertices form a 3-cycle.



■ **Figure 14** Schematic drawing of a potential realization. Net of the three gray polygons and traces of the planes that contain the red and green polygons, which must touch each of the gray polygons. The line of intersection between two of the gray polygons is drawn twice (dashed).

We can realize n -vertex graphs with $4n - O(1)$ edges by stacking horizontal polygons into a tetrahedron whose bottom face is horizontal and whose bottom corners have been cut off, but we can do better.

► **Theorem 2.** *There is a family of graphs $(G_k)_{k \geq 4}$ such that, for each $k \geq 4$, G_k has $n_k = k^2 + k - 3$ vertices and $5n_k - O(\sqrt{n_k})$ edges, and admits a realization by convex polygons with edge contacts.*

Proof. Let $k \geq 4$ be arbitrary but fixed. We construct G_k in several steps:

1. We start with a maximally triangulated planar graph H_k with m vertices, and designate a root vertex r of degree $c = 3$. Note that, for every $n \geq 4$, there exists a triangulation with a vertex of degree 3, e.g., a planar 3-tree.
2. Create a convex polytope P_k with H_k as its dual such that the face of r is *largest*, i.e., there exists a direction in which we can project the polytope such that the image is planar and r is the outer face.
3. Delete the face of r , and scale P_k until it is almost completely flat.
4. Imagine a k -sided cylinder. We will assume the axis of the cylinder is vertical.
5. Create k copies of P_k and place them at the faces of the cylinder, with the hollow side facing the inside of the cylinder. Each copy is rotated such that no edges are horizontal, and they are rotation-symmetric (around the cylinder axis) copies of each other.
6. For every vertex v of P_k not adjacent to r (corresponding to a face of H_k), create a regular k -gon with the k copies of v as its vertices. The k -gon will be horizontal.
7. Since v has (at least) three incident edges, either two go up and one down, or the other way around. In the first case, move the k -gon up by ε . In the second case, move the k -gon down by ε .
8. Cut a corner of each face between the two upgoing or downgoing edges, and glue the resulting horizontal edge to the k -gon.

We now have $n_k = k(m - 1) + (2m - c)$ convex polygons with $k(3m - 6 - c) + k(2m - c) = 5km - (6 + 2c)k$ adjacencies. Using $c = 3$ and $m = k$ yields $n_k = k(m - 1) + (2m - 3) = k^2 + k - 3$ and $k(3m - 6 - 3) + k(2m - 3) = 5k^2 - 12k = 5n_k - 17k + 15 = 5n_k - O(\sqrt{n_k})$. ◀

Evans et al. [2] also considered representing *hypergraphs* in 3D. In their model, each hyperedge is represented by a (flat) convex polygon, two polygons share a corner if and only if the two hyperedges share a vertex, and any two polygons have disjoint interiors. They showed that the two smallest Steiner triple systems $S(2, 3, 7)$ and $S(2, 3, 9)$ admit such a

representation (using triangles). In addition, they conjectured that no Steiner quadruple system (SQS) can be realized by using quadrilaterals in 3D.

We show that their conjecture is true for *convex* quadrilaterals. Assume that a SQS can be represented using convex quadrilaterals in 3D. Then the intersection of these quadrilaterals with a small sphere around a vertex is a planar graph. In a SQS $S(3, 4, n)$, each vertex is incident to $(n-1)(n-2)/6$ convex quadrilaterals, which can be split into $(n-1)(n-2)/3$ triangles. This yields a planar graph on $(n-1)$ vertices with $(n-1)(n-2)/3$ edges, which is impossible for $n > 9$. The same argument also precludes arbitrary topological embeddings into arbitrary 3-manifolds. Since Evans et al. [2] showed that $S(3, 4, 8)$ has no realization with quadrilaterals in 3D (and there is no SQS for $n = 9$), no SQS can be realized using convex quadrilaterals.

References

- 1 Christian A. Duncan, Emden R. Gansner, Y. F. Hu, Michael Kaufmann, and Stephen G. Kobourov. Optimal polygonal representation of planar graphs. *Algorithmica*, 63(3):672–691, 2012. doi:10.1007/s00453-011-9525-2.
- 2 William Evans, Paweł Rzażewski, Noushin Saeedi, Chan-Su Shin, and Alexander Wolff. Representing graphs and hypergraphs by touching polygons in 3D. In Daniel Archambault and Csaba Tóth, editors, *Proc. Int. Symp. Graph Drawing & Network Vis. (GD'19)*, LNCS, 2019, to appear. URL: <http://arxiv.org/abs/1908.08273>.
- 3 Szilassi polyhedron. Wikipedia entry. Accessed 2019-10-08. URL: https://en.wikipedia.org/wiki/Szilassi_polyhedron.

Participants

- Elena Arseneva
St. Petersburg State
University, RU
- Maïke Buchin
Ruhr-Universität Bochum, DE
- Benjamin Burton
The University of Queensland –
Brisbane, AU
- Hsien-Chih Chang
Duke University – Durham, US
- Arnaud de Mesmay
University of Grenoble, FR
- Vincent Despré
INRIA Nancy – Grand Est, FR
- Linda Kleist
TU Braunschweig, DE
- Boris Klemz
FU Berlin, DE
- Francis Lazarus
GIPSA Lab – Grenoble, FR
- Maarten Löffler
Utrecht University, NL
- Anna Lubiw
University of Waterloo, CA
- Clément Maria
INRIA – Valbonne, FR
- Tim Ophelders
Michigan State University, US
- Hugo Parlier
University of Luxembourg, LU
- Saul Schleimer
University of Warwick –
Coventry, GB
- Lena Schlipf
FernUniversität in Hagen, DE
- André Schulz
FernUniversität in Hagen, DE
- Eric Sedgwick
DePaul University – Chicago, US
- Rodrigo I. Silveira
UPC – Barcelona, ES
- Jonathan Spreer
University of Sydney, AU
- Frank Staals
Utrecht University, NL
- Stephan Tillmann
University of Sydney, AU
- Ivor van der Hoog
Utrecht University, NL
- Birgit Vogtenhuber
TU Graz, AT
- Carola Wenk
Tulane University –
New Orleans, US
- Erin Moriarty Wolf Chambers
St. Louis University, US
- Alexander Wolff
Universität Würzburg, DE

