



DAGSTUHL REPORTS

Volume 13, Issue 5, May 2023

Empirical Evaluation of Secure Development Processes (Dagstuhl Seminar 23181) <i>Eric Bodden, Sam Weber, and Laurie Williams</i>	1
Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 23191) <i>Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova</i>	22
Topological Data Analysis and Applications (Dagstuhl Seminar 23192) <i>Ulrich Bauer, Vijay Natarajan, and Bei Wang</i>	71
Regular Transformations (Dagstuhl Seminar 23202) <i>Rajeev Alur, Mikołaj Bojańczyk, Emmanuel Filiot, Anca Muscholl, and Sarah Winter</i>	96
Scalable Data Structures (Dagstuhl Seminar 23211) <i>Gerth Stølting Brodal, John Iacono, László Kozma, Vijaya Ramachandran, and Justin Dallant</i>	114
Designing the Human-Machine Symbiosis (Dagstuhl Seminar 23212) <i>Ellen Yi-Luen Do, Pattie Maes, Florian ‘Floyd’ Mueller, and Nathan Semertzidis</i>	136
Computational Geometry (Dagstuhl Seminar 23221) <i>Siu-Wing Cheng, Maarten Löffler, Jeff M. Phillips, and Aleksandr Popov</i>	165
Novel Scenarios for the Wireless Internet of Things (Dagstuhl Seminar 23222) <i>Haitham Hassanieh, Kyle Jamieson, Luca Mottola, Longfei Shangguan, Xia Zhou, and Marco Zimmerling</i>	182

ISSN 2192-5283

Published online and open access by

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

Publication date

November, 2023

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 <https://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).



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

- Elisabeth André
- Franz Baader
- Daniel Cremers
- Goetz Graefe
- Reiner Hähnle
- Barbara Hammer
- Lynda Hardman
- Oliver Kohlbacher
- Steve Kremer
- Rupak Majumdar
- Heiko Mantel
- Albrecht Schmidt
- Wolfgang Schröder-Preikschat
- Raimund Seidel (*Editor-in-Chief*)
- Heike Wehrheim
- Verena Wolf
- Martina Zitterbart

Editorial Office

Michael Wagner (*Managing Editor*)
Michael Didas (*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
<https://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.13.5.i

Empirical Evaluation of Secure Development Processes

Eric Bodden*¹, Sam Weber*², and Laurie Williams*³

1 Universität Paderborn, DE. eric.bodden@uni-paderborn.de

2 Carnegie Mellon University – Pittsburgh, US. smweber@andrew.cmu.edu

3 North Carolina State University – Raleigh, US. williams@csc.ncsu.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23181 “Empirical Evaluation of Secure Development Processes”. This was the second seminar on this subject. It brought together researchers and practitioners from the fields of software engineering, IT security and human factors, to discuss challenges and possible solutions with respect to empirically assessing secure engineering activities.

Seminar May 1–5, 2023 – <https://www.dagstuhl.de/23181>

2012 ACM Subject Classification Software and its engineering → Software creation and management; Security and privacy → Software security engineering

Keywords and phrases Empirical assessment, Secure development lifecycle


Digital Object Identifier 10.4230/DagRep.13.5.1

1 Executive Summary

Eric Bodden (Universität Paderborn, DE)

Sam Weber (Carnegie Mellon University – Pittsburgh, US)

Laurie Williams (North Carolina State University – Raleigh, US)

License  Creative Commons BY 4.0 International license
© Eric Bodden, Sam Weber, and Laurie Williams

In the past decades, the cybersecurity community has created many principles and practices for developing secure software. However, this knowledge has generally been assembled by the application of common sense and experience, and while individual measures and techniques are often based on real-world data, broader strategies and processes for creating secure software are usually not subjected to rigorous evaluation. This is a serious shortcoming: common sense can be mistaken and experiences over-generalized. Evaluation techniques are necessary to provide a firm scientific basis that can support progress in the field.

Some such techniques do exist for the later software development stages – implementation and testing. Here one enjoys good automation and the mapping between technique and end-product is relatively clear-cut. It is also in these stages where security teams succeed at least partially in providing software developers with concrete prescriptive guidance. Unfortunately, the earlier developmental stages – requirements elicitation, threat modeling, architecture – are just as critical to the security of the final product, yet pose a much greater experimental challenge because of the gap between the process and the product. Experience has shown only limited success at turning software engineers into security experts, particularly so for these crucial initial stages.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Empirical Evaluation of Secure Development Processes, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 1–21

Editors: Eric Bodden, Sam Weber, and Laurie Williams



Dagstuhl Reports

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

Our previous Dagstuhl Seminar 19231 formed a community interested in empirical investigation of secure development practices. This Dagstuhl Seminar now sought to compile a volume merging empirical software engineering and security research to assist the involved communities, including industry and academia, in focusing their research efforts, and to help newcomers to our field find fertile research areas.

The seminar was designed to be highly interactive, with only three introductory presentations on how security researchers, software engineering researchers, and practitioners think about secure software engineering, and which challenges they perceive, particularly with respect to empirical assessment and evidence. Participants then regularly regrouped in altogether one dozen interactive breakout sessions on various topics covering all activities of a prototypical secure development lifecycle, with the intention of eventually gaining the ability to formulate chapters in a to-be-written textbook on the subject.

A special highlight of the seminar was the remote talk by Steve Lipner, former security executive at Microsoft and now executive director of SAFECode, who recapped the most interesting recollections about his introduction of the first secure development lifecycle at Microsoft some 25 years ago, known as the Window Security Push, details about which can be found below.

2 Table of Contents

Executive Summary

<i>Eric Bodden, Sam Weber, and Laurie Williams</i>	1
--	---

Overview of Talks

How do Software Engineering researchers see the world? <i>Eric Bodden</i>	4
What practitioners seek from the community <i>Alex Gantman</i>	4
Inside the Windows Security Push: A Twenty-Year Retrospective <i>Steve Lipner (remotely)</i>	4
How Security People Think of the World <i>Sam Weber</i>	5

Working groups



Breakout Group “Humans in Empirical Evaluation of Secure Development Processes” <i>Yasemin Acar, Robert Biddle, and Sascha Fahl</i>	5
Breakout Session “Software Supply Chains” <i>Eric Bodden and Laurie Williams</i>	6
Breakout Session “Security Metrics” <i>Daniela Soares Cruzes and Akond Rahman</i>	7
Breakout Session “Architecture & Design” <i>Joanna Cecilia da Silva Santos</i>	7
Breakout Session “Adversariality” <i>Olga Gadyatskaya, Robert Biddle, Haipeng Cai, Sven Peldszus, Sam Weber, and Charles Weir</i>	8
Breakout Session “SDL Practices and Budget” <i>Olga Gadyatskaya, Robert Biddle, Eric Bodden, Daniela Soares Cruzes, Alex Gantman, Alessandra Gorla, Henrik Plate, and Sam Weber</i>	9
Breakout Session “Threat Modeling” <i>Kevin Hermann</i>	10
Breakout Session “More Synergy between Software Engineering and Security” <i>Ranindya Paramitha</i>	11
Breakout Session “Code and Design Reviews” <i>Sven Peldszus</i>	12
Breakout Session “Longitudinal Studies” <i>Akond Rahman</i>	17
Breakout Session “Secure Generative AI” <i>Akond Rahman</i>	17
Breakout Session “Tensions between Industry and Academic Objectives” <i>Dominik Wermke and Henrik Plate</i>	18

Participants	21
-------------------------------	----

3 Overview of Talks

3.1 How do Software Engineering researchers see the world?


Eric Bodden (Universität Paderborn, DE)

License  Creative Commons BY 4.0 International license
 Eric Bodden

In this talk I briefly convey my personal experience on how software engineering researchers perceive research on *secure* software engineering. As I will explain, there are a number of related fields such as programming languages and formal verification that also contribute to the goal of securing software. I will contrast software engineering from these fields. Moreover I will discuss the subjects that have received most attention in the field of software engineering and the publishing culture within the community. This culture is very different from that in IT security, and focuses much more on defenses than attacks, and on processes just as much as tools. This also impacts meta-issues such as the quest for more reproducible studies and sharing of artifacts.

3.2 What practitioners seek from the community

Alex Gantman (Qualcomm Research – San Diego, US)

License  Creative Commons BY 4.0 International license
 Alex Gantman

In this brief talk I will highlight my personal view on what challenges software security practitioners face. This includes: (1) Measuring the outcomes and impact of our work (beyond measuring the effort invested), (2) Scaling to large code bases, large organizations, and complex supply chains, and (3) Lack of robust theoretical foundations for our practices.

3.3 Inside the Windows Security Push: A Twenty-Year Retrospective

Steve Lipner (remotely)

License  Creative Commons BY 4.0 International license
 Steve Lipner (remotely)

Main reference Steve Lipner, Michael Howard: “Inside the Windows Security Push: A Twenty-Year Retrospective”, IEEE Secur. Priv., Vol. 21(2), pp. 24–31, 2023.

URL <https://doi.org/10.1109/MSEC.2022.3228098>

This talk discusses a follow-up to an article on the Windows security push in the first issue of IEEE Security and Privacy (January 2003). It provides additional detail on the security push and its results, and describes the creation and evolution of the Security Development Lifecycle (SDL) that integrated software security into Microsoft’s development process.

3.4 How Security People Think of the World

Sam Weber (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Sam Weber

In this talk I briefly share my personal view on how security researchers view the world of secure software engineering, from the early beginnings within the military to the worldwide community we have today. I will explain how we arrived at an attacker mindset and a culture of distrust, and why security is hard to achieve. Lastly, I will highlight some areas that in my opinion require further research, particularly software/system architecture, finding non-generic issues with security tools and certification as well as risk management.

4 Working groups

4.1 Breakout Group “Humans in Empirical Evaluation of Secure Development Processes”

Yasemin Acar (George Washington University, DC, US), Robert Biddle (Carleton University – Ottawa, CA), and Sascha Fahl (Leibniz Universität Hannover, DE)

License  Creative Commons BY 4.0 International license
© Yasemin Acar, Robert Biddle, and Sascha Fahl

This breakout session was motivated by much work in the “Usable Security” research community over the past 20 years (e.g. Lipford and Garfinkel [1]), where the emphasis has been on end-user human factors relating to security, and increasing interest in research on human factors relating to software developers. Our goal was to consider issues specific to human factors relating to developers and topics specific to development of secure software.

We began with an introduction to some key aspects of the area, inviting discussion. Our group consisted of 10 researchers, 4 of whom have been actively engaged in the usable security community for many years. That community had coalesced around such viewpoints articulated by Zurko and Simon as early as 1997 in “user-centered security” [3], and Adams and Sasse in their 1999 paper “User are Not the Enemy” [2], and took positive approaches on better understanding end-users and their contexts, and proposing then evaluating better designs to help users be secure. For software developers, we want to eschew regarding developers as stupid or lazy, and focusing on how we can make secure development easier.

We discussed classic usable security questions such as

- How can we get these people to do secure thing?
- Why are they not doing the secure thing?
- How can we make “doing the secure thing” easier for people to do?
- How do their minds work? What are they (not) worried about?
- How do they use technology (insecurely)?

and discussed that these are generally also researched in software engineering research.


We discussed similarities in methodology across security and software engineering research, and decided to both delve deeper into cognitive frameworks and seminal papers with generally accepted methodology from both fields.

References

- 1 Garfinkel, Simson, and Heather Richter Lipford. Usable security: History, themes, and challenges. Morgan & Claypool Publishers, 2014.
- 2 Adams, Anne, and Martina Angela Sasse. “Users are not the enemy.” *Communications of the ACM* 42.12 (1999): 40-46.
- 3 Zurko, Mary Ellen, and Richard T. Simon. “User-centered security.” *Proceedings of the 1996 workshop on New security paradigms*. 1996.

4.2 Breakout Session “Software Supply Chains”

Eric Bodden (Universität Paderborn, DE) and Laurie Williams (North Carolina State University – Raleigh, US)

License  Creative Commons BY 4.0 International license
© Eric Bodden and Laurie Williams

During our discussion session, we explored various aspects of third-party libraries in software supply chains:

We began by addressing the Log4J issue and the SolarWinds attack, emphasizing the distinction between unintentional programmer errors and deliberate malicious injections. The topic of supply chain attacks was central, with debates on whether known vulnerabilities should be part of these discussions and where to draw the line regarding live downloading or code inclusion. We also touched on challenges, such as defining security in the supply chain, prioritizing vulnerabilities, and handling undocumented behavior in third-party packages. The effectiveness of SBOM was discussed, along with concerns about human factors, like developers making decisions under time constraints and ensuring security in open-source contributions. We explored the role of liability and the importance of interdisciplinary collaboration with legal experts. Self-attestation and the challenges of assessing component behavior were also raised. Cyber-physical impacts of software in supply chains were considered, along with potential risks associated with third-party external attestation. We delved into how to measure supply chain security and identify suitable metrics. Economic considerations were also part of the conversation, including the ongoing cost implications of using open-source software versus in-house development. Our discussion uncovered opportunities related to trust boundaries, assured open-source components, and the security of ecosystems not originally designed for supply chain security. We debated the influence of social proof on library selection and the need for developers to comprehend library features. Additionally, we discussed selective loading of library code and the importance of secure dependency tools. We examined risks associated with solo developers and small teams, including their motivations and interpersonal dynamics. The distinction between bugs and vulnerabilities in software supply chains was clarified. Lastly, there was a suggestion to conduct a case study on a software supply chain and explore interdisciplinary research with legal experts. In summary, our discussion was comprehensive, covering a wide range of topics related to third-party libraries in software supply chains, including challenges, opportunities, human factors, security concerns, and economic considerations. This conversation underscored the intricate nature of supply chain security in the software industry.

4.3 Breakout Session “Security Metrics”

Daniela Soares Cruzes (NTNU – Trondheim, NO) and Akond Rahman (Auburn University, US)

- License** © Creative Commons BY 4.0 International license
 © Daniela Soares Cruzes and Akond Rahman
- Main reference** Patrick Morrison, David Moyer, Rahul Pandita, Laurie A. Williams: “Mapping the field of software life cycle security metrics”, *Inf. Softw. Technol.*, Vol. 102, pp. 146–159, 2018.
URL <https://doi.org/10.1016/j.infsof.2018.05.011>
- Main reference** Marcus Pendleton, Richard Garcia-Lebron, Jin-Hee Cho, Shouhuai Xu: “A Survey on Systems Security Metrics”, *ACM Comput. Surv.*, Vol. 49(4), pp. 62:1–62:35, 2017.
URL <https://doi.org/10.1145/3005714>

The goal of our breakout session was to understand the challenges and opportunities for security metrics in the context of empirical evaluation of secure software processes. We started the discussion by discussing existing reviews of security metrics [1, 2]. During this breakout group we discussed a wide range of topics related to challenges and opportunities for future work related to security metrics. Some of the highlights of the breakout session are: (i) the usefulness of metrics is dependent on context and stakeholder preferences, (ii) a lack of a metric suite that provides a holistic overview of the systems, and (iii) we derived a set of research questions, which included questions related with incentives and evaluation measures. We believe that this breakout session provides the groundwork for synthesizing the challenges, opportunities, and open research questions for the secure software development community. Examples of open research questions that came out of this session include but are not limited to: (i) how do we evaluate metrics that do not have a ground truth?; (ii) how to determine the usefulness of metrics?; (iii) how do we measure risk for interface designs?; (iv) what approaches can we use to evaluate metrics?; and (v) how much evidence is required to demonstrate initial viability of a security metric?

4.4 Breakout Session “Architecture & Design”

Joanna Cecilia da Silva Santos (University of Notre Dame, US)

- License** © Creative Commons BY 4.0 International license
 © Joanna Cecilia da Silva Santos

Software architecture design plays a crucial role in ensuring that security requirements are addressed. It allows for the identification and mitigation of potential security risks early in the development lifecycle. Given this importance, the Architecture & Design working group discussed the disjoint between software engineering and software security community, the need for mapping studies between the well established practices in software architecture and software security. Moreover, the group deliberated about secure software design and architecture (including architecture design principles and practices), and the challenges of implementing secure design principles. The key challenges identified were

1. Architectural models may not always be available: how to analyze software architectures regarding security?
2. If models are available, how to conduct these analyses in a scalable and automated fashion?
3. Software architecture descriptions may have different formats. How to analyze such a heterogeneous set of architecture descriptions?

4. The discussions about (secure) design decisions tend to be more qualitative than quantitative. How to measure these decisions? That is, how to evaluate the effectiveness of software designs for security?
5. How to tame with architectural drift?
6. How do you integrate design practices into the software development process of an organization?
7. Secure coding tend to be small low-level practices that are not tied to higher-level secure design decisions. How to connect secure coding practices to these design decisions?
8. How to integrate commercial-off-the-shelf (COTS) products? How to re-evaluate security properties after integrating other components?

Given these challenges, the group discussed the idea of creating a body of knowledge to guide empirical evaluation. The envisioned body of knowledge would include:

- Anti-patterns, which would focus on what not to do;
- Attack surfaces;
- Architectural styles and patterns along with their security implications;
- A minimum set of secure design decisions captured in a checklist.

This body of knowledge can start with anecdotal stories about (in)security by design that leads to (severe) vulnerabilities, which would be used to highlight the importance of constructing software systems that are secure by design.

The group proposed several research ideas, such as an evaluation experiment to assess the effectiveness of design decisions, observational studies to evaluate design changes in response to security incidents, and a study of mental checklists used by practitioners.

4.5 Breakout Session “Adversariality”

Olga Gadyatskaya (Leiden University, NL), Robert Biddle (Carleton University – Ottawa, CA), Haipeng Cai (Washington State University – Pullman, US), Sven Peldszus (Ruhr-Universität Bochum, DE), Sam Weber (Carnegie Mellon University – Pittsburgh, US), and Charles Weir (Lancaster University, GB)

License  Creative Commons BY 4.0 International license

© Olga Gadyatskaya, Robert Biddle, Haipeng Cai, Sven Peldszus, Sam Weber, and Charles Weir

Adversaries are a core part of the security process. Yet, we often struggle to understand and to predict them. In this session we discussed what we know about adversaries and what could help us to protect our systems from unknown miscreants.

One of the key discussion points was that adversariality is an inherent part of nature that drives evolution. Parasites exist in all ecosystems, and no barriers to stop them work perfectly. At the same time, resilience of ecosystems to parasites is ensured by diversity. Yet, in software engineering monocultures flourish because it is easier to develop and maintain them, and it is very often that one system dominates a whole market. For example, Chrome is the leading browser today, and, moreover, all popular browsers but Firefox are based on WebKit. We have seen how big the cost of a single vulnerability can be in a monoculture with the Heartbleed bug in the OpenSSL library.

So the solution seems to be in diversity. Competition ensures that our ecosystems are more robust. In many areas alternatives do exist: operating systems, programming languages, network protocol stacks, cryptography libraries, machine architectures, and others.

At the same time, at the level of individual organizations monocultures are appreciated as they reduce maintenance and monitoring costs. Ensuring diversity also comes with its own challenges, not least the scale and depth of supply chains, where software diversity might be evident at one level, but depend on the same components at deeper levels. Thus, we need to come up with new methods to ensure diversity cost-effectively and assess the security protections it affords to organizations.

We have also discussed that the security game is about costs. Defensive mechanisms make costs higher for the attacker. For example, slowing down the password check routine helps tremendously in preventing bruteforcing and other kinds of attacks. Economic models are important for understanding how adversaries operate and how to disrupt them. We observed that other disciplines, such as criminology, study attackers as well. Joining forces can be a way forward.

4.6 Breakout Session “SDL Practices and Budget”

Olga Gadyatskaya (Leiden University, NL), Robert Biddle (Carleton University – Ottawa, CA), Eric Bodden (Universität Paderborn, DE), Daniela Soares Cruzes (NTNU – Trondheim, NO), Alex Gantman (Qualcomm Research – San Diego, US), Alessandra Gorla (IMDEA Software Institute – Madrid, ES), Henrik Plate (Endor Labs – Palo Alto, US), and Sam Weber (Carnegie Mellon University – Pittsburgh, US)

License © Creative Commons BY 4.0 International license
© Olga Gadyatskaya, Robert Biddle, Eric Bodden, Daniela Soares Cruzes, Alex Gantman, Alessandra Gorla, Henrik Plate, and Sam Weber

In this breakout session we discussed cost-effectiveness of individual practices in secure development lifecycle (SDL), how it can be defined and improved.

One of the points discussed was measurability: some practices yield results that are inherently more measurable than others. For example, it is easier to measure the number of vulnerabilities found in fuzzing compared to measuring aggregated outcomes of code review. Cost is another aspect that is inherently hard to measure: it is difficult to estimate the cost of using a library that might need to be patched later.

Organizations might be able to make decisions about cost-effectiveness if they know their return on security investment and can predict their security and business risks. Yet, this is very challenging, especially for the early phases of SDL. One method to understand cost-effectiveness of early SDL phases is to review completed projects and analyze what could have been done differently. However, this does not allow to fully grasp the situation, as SDL processes usually address what has already been encountered by that organization, but not yet-unknown challenges. Similarly, there is a lack of understanding of the cost-effectiveness of such activities as awareness and training for developers.

Moving forward, it would be interesting to conduct a large industry survey on distribution of investments over different SDL phases. We hypothesized that for many organizations most of the effort is spent on implementation and testing phases because all developers are involved there, while all other phases would have much smaller effort allocated to them. It would be interesting to see how this distribution changes with increasing maturity of organizations, and whether we can prove that spending more effort and budget in the early phases will lead to decrease of effort required for the later phases, especially the maintenance phase. Organizations are willing to invest money in security, but they need to know how to spend it better.

4.7 Breakout Session “Threat Modeling”

Kevin Hermann (Ruhr-Universität Bochum, DE)

License  Creative Commons BY 4.0 International license
© Kevin Hermann

The second discussion on threat modeling delved into the concept of a threat model product line, which involves developing distinct threat models for shipping a product to different companies or countries. It emphasized the importance of traceability and variability in threat models to identify areas for improvement and assess associated costs.

Threat Model Product Lines

Threat models vary based on the specific context, country, and product variants. Chipsets and cloud-based platforms are shipped or offered in multiple countries which may have different threats. Applying a change for one customer to mitigate a threat in one context can lead to disabling functionalities for other customers. Therefore, challenges arise by addressing vulnerabilities in one variant without disrupting functionality in others. Building threat models during the product stage and incorporating them into incident response are valuable, given that vulnerabilities can still emerge. Propagating changes across different variants is a complex task, often requiring modifications to existing threat models.

Challenges

Evaluating this approach is a major challenge, as no metrics to assess the effectiveness of threat models have been established, yet. However, considering factors beyond the STRIDE model, such as attacker resources and utilizing attack tree models are useful to estimate risks and costs in variant development. As an example, if an IoT device which has no value suddenly enters the White House, its value increases and therefore the potential for attacks. Additionally, the selection of relevant factors for creating effective threat models is difficult, as modelling irrelevant factors can lead to overestimation.

Research Directions

Creating a tool to derive multiple threat models from a single model, potentially through the use of STRIDE tools, for which Data Flow Diagrams (DFDs) are required, could be the first step to present the idea of threat model product lines. However, difficulties arise on validating the DFDs created for a threat model. Comparing graphical models is hard, as it often requires human interpretation. Instead, the use of natural languages seems promising, as they are simple to compare.

Conclusion

In conclusion, the breakout session provided valuable insights into the complexities of threat modeling in different contexts, challenges associated with merging and propagating models, the necessity of multiple threat models for the same product, and the importance of credible and validated threat models for effective security measures. Finally, first ideas for research directions for threat model product lines were discussed.

4.8 Breakout Session “More Synergy between Software Engineering and Security”

Ranindya Paramitha (University of Trento, IT)

License © Creative Commons BY 4.0 International license
© Ranindya Paramitha

Joint work of Yasemin Acar, Evan Austin, Haoipeng Cai, Sascha Fahl, Ben Hermann, David Lo, Alena Naiakshina, Ranindya Paramitha, Henrik Plate, Dominik Wermke, Laurie Williams

Software Engineering and Security research are two different yet intersecting worlds whose intersections have been discussed through decades [1]. Having the two communities together sitting in the same room brought some interesting questions: (1) How are Software Engineering and Security research similar and different in general? (2) Is it possible/ how to borrow methodologies from each other? (3) How to do something impactful with this synergy?

Security paper is famous for being related to something “scary”. There are several “kinds” of security research: (1) Attacks: the type of security research that focuses on the discovery of the “bad”, eg. [2] in CCS’22. The focus is to show the attacks’ interesting impacts against many targets or small but important targets. (2) Defenses: this kind of research tends to be harder on getting the paper published. The reason for this is that finding holes in defense is considered easier than criticizing attacks. (3) Security measurement: including manual analysis, human factor/ usable security. One interesting challenge is to understand how to increase the cost of attack: how to make it expensive enough to attack so people do not attack a system, which can be economically or psychologically. (4) Tooling (to support attacks/ defenses). In general, security research focuses more on finding new attacks (the discovery of the “bad”, security fatalism) and generalizing them.

On the other hand, Software Engineering research focuses more on the fundamental issues, and not directly finding something: applying methodologies to improve practices in software engineering. Back in 2000, Software Engineering focused more on (1) process modeling (laying out a process, eg. agile software model [3], SDLC) with less validation (no validation/ toy problems were common). These days there is more research on (2) empirical analysis: observation studies and generating theories from it, eg. finding the pattern of how people collaborate in the software ecosystem. This includes (3) experiments with a negative result, as the community believes that interesting questions and well-designed experiments are still valuable even with a negative result, eg. the Replications and Negative Results (RENE) track in ICPC’22 collocated with ICSE’22. There is also (4) human factors research, which uses methods such as systematized user surveys to practitioners (eg. [4]) in order to understand what and how developers think in software engineering, eg. why they work in one way and not the other. On these latest years, there has been an emerging track called (5) Registered Report track (eg. in ESEM’22), which allows researchers to submit 6 pages experiment design, get peer-reviewed, present it, and then have a maximum 1-year period to conduct the experiment and submit the full paper to a journal with continuity acceptance.

Software Engineering and Security have intersections in several ways, eg. the concept of “bugs” in Software Engineering is similar to “vulnerabilities” in Security. “Novelty” in Software Engineering research is valued like “attack” in Security. Nowadays, papers from one community can also be accepted in others, eg. tooling papers that can find a class of vulnerability for a lot of packages can be accepted in both Software Engineering and Security conferences. The possible synergy or “bridge” to get the best of both worlds is to have research with great fundamental methodology (Software Engineering) but with a big “splash” impact (Security). However, there is still a need for the systematization of knowledge in the intersection between Software Engineering and Security, both from different papers but also

other artifacts (ie. gray literature). This can bring scientific contribution to finding the gap, which area research has been done, which area where more research is still needed, and even which research area is not promising.

References

- 1 Mouratidis, H., & Giorgini, P. (Eds.). (2006). Integrating Security and Software Engineering: Advances and Future Visions: Advances and Future Visions.
- 2 Mingrui Ai, Kaiping Xue, Bo Luo, Lutong Chen, Nenghai Yu, Qibin Sun, and Feng Wu. 2022. Blacktooth: Breaking through the Defense of Bluetooth in Silence. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22). Association for Computing Machinery, New York, NY, USA, 55–68. <https://doi.org/10.1145/3548606.3560668>
- 3 Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J. & Thomas, D. (2001). Manifesto for Agile Software Development Manifesto for Agile Software Development.
- 4 Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J. & Thomas, D. (2001). Manifesto for Agile Software Development Manifesto for Agile Software Development.

4.9 Breakout Session “Code and Design Reviews”

Sven Peldszus (Ruhr-Universität Bochum, DE)

License  Creative Commons BY 4.0 International license
© Sven Peldszus

The code and design review breakout group focused on the discussion of how to evaluate tools and human factors in group and design reviews. Reviews are an essential part of security evaluations during the development of software systems. Following the principle of security by design, security has to be considered already at design time. The planned design must be reviewed concerning the security of the system as well as its implementation later. Despite such reviews can be supported by tooling, the reviewer has an essential role.

Effectiveness of static code analysis

In practice, there is a significant difference in what different reviewers look for in code reviews and their performance. The job of security experts is to find the one critical bug. In contrast to this, the general developer is likely to accept reviews that consist of more than 70% true positive findings.

While it is essential to find as many security bugs as possible, a significant practical barrier is false positive findings that hinder addressing security effectively. False positive findings have to be resolved by humans that come with their individual preferences and might be more effective in handling specific kinds of false positive findings. To optimize the outcomes of code reviews, it is important to tailor these closer to the target audience of the review. To this end, it could be beneficial to learn what are the usual most favorite false positive findings of different stakeholders.

A general approach for optimizing tool-assisted reviews could be to learn from current false positives to mitigate them in the future. Such learning must be prepared by intensive data mining. However, to optimize the reviews, we must identify the false positive findings to

avoid in the future, we must rank all the identified false positives according to some criteria that have yet to be identified. One possibility could be a rating of how critical a specific kind of finding is to fix. The intuition is that developers will start to work on the most critical finding kinds and benefit there the most from fewer false positives. Another possibility could be to consider manual downranks of developers.

For all of these metrics, one practical barrier is to get hands on the necessary data, which is usually not publicly available. One would have to massively collect practical feedback from developers that label the findings of review tools. Since it is not feasible to request a single developer to label all findings of each analysis tool, one could divide the work among multiple parties.

This huge effort in labeling, which is a huge barrier to academic studies to improve the analysis tools, is also a huge barrier in practice when a new tool should be deployed. After the deployment, the outlined task has to be performed to identify the findings of the new analyzer that must be fixed. Here, the application of the new analyzer on only newly written code can help to reduce the number of findings. Developers can focus on the new code they are working on and improve it while reducing the risk of the outcomes of the new analysis tool being ignored entirely. Still, there will be less precise results than expected, since many details have to be trained, including optimizing the configuration of the new tool and also the developers themselves.

Despite these outlined challenges in getting hands with the number of findings and false positives, incremental scans and synchronizing scans across branches are huge barriers. In particular, a false positive identified and labeled in one branch should be also labeled on the other branches on which it is present as a false positive finding.

While these individual observations and ideas sound reasonable, it remains to gather data in an empirical analysis to precisely identify how large the effort of reviews for developers is. Thereby, we have to be careful about what exactly to measure. Among others, there are findings in published works that show that developers struggle with configuring checks, which increases their effort in the end. Also, the findings themselves will not change just by tweaking their distribution. Another question is whether we can transfer our own experiences from applying security analysis tools on open-source projects to the integration of the tools in a large toolchain. In the second scenario, one has to consider multiple stakeholders participating in the review. The security team can help in configuring the analysis tools but others have to run them.

When it comes to the effectiveness of a security review, experiences from the industry have shown that success metrics are essential. Since having a non-exploitable system is usually the major goal, one could consider rating findings according to their exploitability. While fixing a buffer overflow might address a true positive finding, this does not necessarily improve the system's security. However, not having an exploit does not prove the effectiveness of static analysis since we might just not be aware of an exploit, yet. Also, it has been shown that there is a correlation between the pure number of issues in a file and the likelihood of a vulnerability, not taking any additional metrics such as exploitability or criticality into account. Still, even such simple metrics as the plain number of findings might not be applicable due to changes. Therefore, a purely empirical approach by counting the number of findings is probably not the right one.

Comparison of static analysis tools

Even assuming a success metric, the issue of insufficient resources for fixing all issues remains and it is questionable whether it makes sense to deploy tools that can find more or other kinds of vulnerabilities when even we are practically not able to fix all the existing critical vulnerabilities. When analysis tools are applied in classes, students report many findings and question the usefulness of the tool for identifying real issues.

Since the pure deployment of additional analysis tools is not the solution, we have to be more selective in which analysis tools we deploy for what purpose. We need means to properly compare analysis tools to decide on which ones to deploy. To this end, we have to focus more on an analytical comparison of what the tools do for a qualitative comparison of different analysis tools.

Such a comparison of security analysis tools could be based on databases of labeled findings and the rules of the tools themselves. One issue in this regard is the availability of such information which is usually only partly publicly accessible and if so only for a single tool. Comparing rulesets among tools and estimating the overlap between tools is a yet unsolved challenge. Comparable to mutation testing, an assessment of static analysis tools could be realized by artificially introducing bugs. Whether vulnerable samples generated by ChatGPT are suitable as a baseline is currently unclear. In the end, we still rely on people building a limited number of ground truths. Recent work mainly focuses on a quantitative comparison of how many bugs in a known data set can be detected by which tool.

Fixing vulnerabilities

The detection of possible vulnerabilities and their rating is only one of the steps in effectively securing a system. After deciding on what are the concrete vulnerabilities that threaten a system, these have to be fixed. Here, plenty of research has been done in the direction of automated fix generation. While these fix generators are effective in generating fixes, they are yet not used in practice. In the end, the generated fixes can be seen as a blueprint for fixing a detected issue but manual checking of the generated fixes is non-neglectable. In practice, simple automated dependency updates are often rejected by developers which raises the question of how good such tooling has to get.

As in the reviews themselves human peculiarities seem to play an essential also in fixing identified vulnerabilities. Developers might not be satisfied by just accepting proposed fixes while their fix was compatible with the proposed one. Further, even when a vulnerability is fixed by a developer in one team, most likely it will not be fixed in other teams working on the same code in another branch. Often organizational overhead but also lack of communication prevent the effective propagation of security fixes.

Alike to this manual problem in fixing multiple versions of a system, also static analysis works only on a single version but the fix is needed on all versions. While developers have this often in their minds at least for the variants for which they are responsible, tools currently entirely lack such features.

Besides identical duplicates of one bug, we also have to consider additional instances of bugs in similar locations. In the end, humans tend to do the same mistake more than once. Currently, we rely on them to remember these additional locations after one instance has been detected.

One of the challenges in finding such additional locations of bugs is the significant impact of context information that makes a bug probably only so some variants or branches. Therefore, it is essential to check bug-fixing code before it is pushed to other branches.

Design reviews

While static code analyzers have concrete instances on which they can work, design reviews are more challenging. While effective formal methods exist, they are often only feasible for some systems due to their huge overhead. One fundamental problem is usually the lack of design specifications such as models on which a design review can be performed.

While the recovery of models to use in a design review is in principle possible, the main question is how does such a review process look like. In the end, static analysis is well-integrated into today's development processes, while this is not the case for model-based design reviews. While it is favorable to also have such integration for design reviews, there is the danger of the process becoming more important than the product itself.

While static code checks immediately work on the concrete artifacts, for design reviews we have to identify a suitable degree of abstraction. Models can range here from a single very detailed model that is close to the source code to an abstract component diagram with data flows comparable to data flow diagrams in threat modeling.

Depending on the degree of abstraction, different security issues might be identified but most likely not detailed security vulnerabilities such as those identified by static code analyzers. Still, given suitable traceability between the models and code even the low-level static code analysis can benefit from the integration of design models into the process. Among others, design models contain information about elements that are not part of the code but with which it interacts. This information can be leveraged to tailor static code analysis such as taint analysis based on the planned interaction with external entities.

Altogether, design models allow us to systematically structure systems, divide them according to security concerns by creating insulation capsules around security-critical parts, and plan concrete security protections according to the division. Due to possibly non-trivial constraints such structurings and analysis have to be supported by tools. We have to provide architects with guidelines on how to design a secure system. In this regard, anti-pattern catalogs and associations between design patterns and suitable security patterns could help. However, the extraction of such patterns is still open work. Here, pattern mining from repositories could be suitable.

To get more benefit out of such design reviews than an initial plan, their integration into the development process is necessary. Whenever there is any change, we have to find means to reflect it as well in the code as in the design models. However, this integration would allow us also to provide developers with easy-to-comprehend information about changes such as explicitly showing how dependencies among components change when a specific call is added to the implementation. On the downside, an attacker might use exactly this benefit of easily accessible information about security measures to plan an attack.

To be practicable and applicable, we need easy and cheap processes that allow us to build such models incrementally. Here, in particular, scalability is a major concern. While we have had such security-by-design approaches already for a relatively long time, it is unclear what exact improvements we need for bringing them into practice. Nevertheless, examples such as formal methods demonstrate that this is possible.

Summary of tool-assisted security reviews

In security reviews one has to take the perspective of an attacker which is a special case compared to other domains such as safety. Attackers are actively looking for opportunities instead of things happening accidentally. To this end, as a reviewer, you have to always keep all possibilities in mind, while usually lacking the necessary context knowledge. The main task of tooling for design reviews and static code analysis is to help in identifying and presenting unknown dangers.

When considering entirely manual code reviews, reviewers usually tend to report the more obvious findings and the completeness is usually questionable. Still, practices such as pair programming and reviewing the commits of others have been proven to be effective. Here, tool support can help in facilitating these practices, e.g., by suggesting reviewers based on the touched artifacts.

In contrast to manual reviews, tools are often more complete but usually at the cost of precision. Targeting this issue, the question is which low-level findings should be shown to developers. Among others, tools should not only provide huge lists of low-level findings but automatically derive suggestions of suitable security patterns based on the identified dangers. For example, the static analysis identifies what a component is doing and suggests patterns corresponding to that. But even just highlighting the use of critical APIs could be beneficial.

The ultimate goal of the deep integration of tools in the planning and review process is to allow the development of security mechanisms that would be infeasible to plan only manually. For example, we could target more fine-grained rights management. Starting on a high abstraction at the system level, this should be systematically pushed into the applications. This would allow us to make sure in the application code that some parts only have certain permissions and would allow for more control over third-party libraries. The main problem could be getting developers to use the more complex structures that probably result from this. Here, tooling can make such rights management or other security measures easier to include.

Open research problems

We conclude with a summary of open research problems that we identified in the discussion above.

Despite static analysis tools already being widely deployed, their configuration is still an open problem that hinders their effective use. We have to identify simpler ways of configuring static analysis tools. Therefore, it is essential to judge the quality of the results of a tool with a specific configuration. We need to identify suitable measures of security to support such a comparison. But we not only have to be able to compare configurations of an individual tool, but we need means to systematically compare different analysis tools to allow effective tool selection.

We need a deeper investigation of the effectiveness of security measures. To suggest suitable security measures, we have to know if specific measures that have been realized prevented the vulnerabilities for which they have been planned. Related to this it should also be checked what are the cost of specific measures and what is relation to their benefits. Additional measures should be mined from repositories and relations to other principles such as design patterns should be extracted.

Since we aim at integrating design models and design reviews into the development process, we have to identify suitable levels of abstraction to do so. The question is what impact do different views on security have on planning and checking secure designs? This integration and effectiveness could be increased further by creating relations with other security artifacts such as CVEs.

4.10 Breakout Session “Longitudinal Studies”

Akond Rahman (Auburn University, US)

License  Creative Commons BY 4.0 International license
 Akond Rahman

A longitudinal study is a research study that employs repeated and continuous measures to follow entities over prolonged periods of times. As part of one of the breakout sessions participants discussed if longitudinal studies could aid in empirical evaluation of secure development processes. The discussion started with a controversial argument from one of the participants that longitudinal studies are often convoluted with mining software repositories where some researchers frame empirical studies as longitudinal studies for ‘marketing purposes’. With the proper definition of longitudinal studies in context participants discussed teased out multiple challenges in conducting longitudinal studies, which included a lack of motivation amongst academics, the time required to publish results, availability of data, and availability infrastructure and management resources.

While the participants acknowledged the challenges inherent to conducting a longitudinal study, they also agreed on the value this research study can bring to empirical evaluation of secure development processes. The participants laid out the following research questions that can be answered in the context of secure software development using longitudinal studies:

- How does security practices change over time for organizations and across organizations?
- How do industry practitioners use static analyzers?
- What areas in secure software development can benefit from execution of longitudinal studies?
- What are the long-term impacts of using generative artificial intelligence (AI) on secure coding practices?

In all, the session triggered great interest amongst participants, many of whom are now collaborating in conducting a longitudinal study in the domain of secure generative artificial intelligence (AI).

4.11 Breakout Session “Secure Generative AI”

Akond Rahman (Auburn University, US)

License  Creative Commons BY 4.0 International license
 Akond Rahman

Generative artificial intelligence (AI) is the discipline of using unsupervised or semi-supervised machine learning techniques to generate human artifacts, such as software source code, movie reviews, and book summaries. In recent times, the most popular generative AI technique in the context of software engineering is use of large language models (LLMs), such as ChatGPT to automate software engineering tasks. As part of this session, participants discussed their experiences is using ChatGPT for software engineering research. Participants mentioned the use of technique called prompt engineering to use LLMs for generating source code.

All participants agreed that while generative AI helps in automating software engineering tasks, there are some shortcomings. One participant discussed one of their recent paper [1], where they found LLM-generated code to include compilation concerns (e.g., 90% of code not compiling), security smells [2], which provide evidence to the perception of “garbage in,

garbage out”. The participants further stated that 90% of the datasets that LLMs use to train have quality concerns. All of these shortcomings further provided insights on what could be possible research directions for securing generative AI.

The open research questions that were discussed are:


- How can LLMs help in writing secure code?
- How should we engineer prompts to generate secure code?
- What strategies should we use to improve the quality of LLMs without re-training?

References

- 1 M. L. Siddiq, S. H. Majumder, M. R. Mim, S. Jajodia and J. C. S. Santos, “An Empirical Study of Code Smells in Transformer-based Code Generation Techniques,” 2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM), Limassol, Cyprus, 2022, pp. 71-82, doi: 10.1109/SCAM55253.2022.00014.
- 2 A. Rahman, C. Parnin and L. Williams, “The Seven Sins: Security Smells in Infrastructure as Code Scripts,” 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), Montreal, QC, Canada, 2019, pp. 164-175, doi: 10.1109/ICSE.2019.00033.

4.12 Breakout Session “Tensions between Industry and Academic Objectives”

Dominik Wermke (CISPA – Saarbrücken, DE) and Henrik Plate (Endor Labs – Palo Alto, US)

License  Creative Commons BY 4.0 International license
© Dominik Wermke and Henrik Plate

4.12.1 Introduction

- What: A discussion breakout session with academics from software engineering, security, and related fields, as well as people from industry about the (research-related) tensions and opportunities between their objectives.
- Why: Being research fields generally close to industry, both Software Engineering and Security often rely on direct interaction with and feedback from industry to push and adapt research ideas further.
- Outcomes: A number of entities and approaches were identified as impactful opportunities for “bridging the gap” between industry and academia, namely institutions like Fraunhofer, conferences like SecDev, and events like developer summits and industry workshops.

4.12.2 Industry Problems

The discussion was initiated with a question from the academic side: What problems does industry face, why don’t they share (them with researchers)? To which industry responded that they do share their problems, but academia considers most of them to be not interesting problems. Academia brought up that they are restricted in problem / topic selection by a number of factors, namely that they need to be publishable (and fit specific venue requirements) and enable a career both for junior researchers and involved graduate students. Based on this response, the discussion moved on to if industry problems without solutions are still interesting to academia. As an example, the initially low adoption of test generation in industry because of flaky tests was brought up, a problem which was then picked up by

academics. There was agreement that beautiful / interesting problems from industry can be picked up by academia and that both industry and academia consider their respective problems to be interesting.

It was then discussed whether the discussion framing is part of the problem, namely that industry asks themselves what academia can do for them vs. academia asks how they can solve industry problems. Industry pointed out that the academia framing in reality appears to them to be more along the lines of “show me your problems so I can solve the interesting ones”. It was proposed that academics might not accept research (problems from industry) because they might not fit their mental model. Based on that, the general question on whether the academic side is open to expanding what they work on was posed. Boundaries imposed by existing publishing venues were identified as a potentially limiting factor, highlighting opportunities to develop better-suited venues targeting industry problems.

4.12.3 Applicable Results

Following the statement “Always go after the practical problems”, the discussion turned to how to actually turn research results into applicable results in industry and what or who is required to bridge that gap. It was pointed out that there are organizations with the specific goal of bridging that gap (e.g., Fraunhofer in Germany). It was also discussed that bridging the gap likely involves efforts from all involved parties: Industry needs to understand the current state of research and be willing to apply the published and applicable findings. Researchers need to have a mind about what is practical and be willing to receive industry feedback on the applicability of their findings.

4.12.4 Innovation

The next discussion point was based on a slide from Alex Stamos’ 2019 USENIX talk about “Tackling the Trust and Safety Crisis” (a pyramid putting what is actually talked about at USENIX at the top vs. other, more impactful areas of InfoSec for industry below it). Based on that slide, the point was brought up that a lot of the innovation in the InfoSec area actually comes from industry and that academia runs after industry. This situation posed the question on whether academia might have been trapped in the valley of industry impact, resulting in pushing researchers away from more fundamental research approaches. As an example, more foundational theories in other areas were brought up, like in the fundamental theories of databases, compilers, and operation systems.

Developer Summits were brought up for a way to bring industry developers together to discuss hard problems. It was postulated that these summits work (for both industry and present researchers) because industry people love to talk to industry people about all the cool stuff they have been doing, and researchers are present soaking it up. A question was posed on how to publish findings from these summits, with the discussion leaning more towards opinion pieces or initial exploratoritive approaches to identify industry’s problems. As a research idea for validating perception, “100 questions from/for industry” was introduced. As challenges for hosting developer summits the discussion included: who to contact and how to bring smaller companies together (that can’t send someone to summits. It was mentioned that including developers from smaller companies also presents an opportunity, because if smaller companies can send someone, they probably have greater oversight and decision power over the tech in the company.

Another discussed challenge was that people love to talk about things they plan on implementing in the future, not the experiences they actually made, with a suggested solution for research being to carefully listen and have the technical background to discern these

answers. After the break, the discussion continued about Workshop / Discussion events (~ 2 h, Chatham house rules), with patterns for success of these events being identified as networking, critical mass, and no shortcuts. A challenge was brought up in that it is not always possible to directly trace the impact of conversations you had at these events and not every idea actually being able to be traced back.

4.12.5 Why Do Research Groups Fail?

The next discussion point was around the question of why industry research groups / departments / teams fail. The discussion centered around if there actually has to be a good probability of failure in research, that even the failed cases have to be spun as successes to publish in academia, and the potential cultural split for research departments vs. other teams because they focus on long term problems.

Based on this potential cultural split, the next discussion point focused on the differences in objectives for industry and academia, namely that “research” in academia and industry (might) not mean the same thing. Points included the need to differentiate between short term solutions and long term solutions, with industry required to provide value immediately and academia being more oriented towards long term. Another point was how to define a good goal or bad goal, with research trying to address real problems that are relevant for the industry, having at least a minimum impact.

4.12.6 Recap

The final discussion point focused on recapping both breakout sessions, with tension points between industry and academia including: Academia only grabbing the problems they think are interesting and then leaving, with industry internships having the potential to better bridge this gap. Industry problems are uninteresting for academia because they often are just constraints for business reasons with obvious solutions and research can not help with that. Messy/complex industry systems in general, with the problem that if the better / correct solution doesn't lead to better outcomes, is it really better? The breakout closed with a recap of the potential of places for academia and industry collaboration such as (industry) conferences, Fraunhofer, and developer workshops.

Participants

- Yasemin Acar
George Washington University,
Washington, DC, US
- Evan Austin
NRL – Washington, US
- Alexandre Bartel
University of Umeå, SE
- Thorsten Berger
Ruhr-Universität Bochum, DE
- Robert Biddle
Carleton University –
Ottawa, CA
- Eric Bodden
Universität Paderborn, DE
- Haipeng Cai
Washington State University –
Pullman, US
- Michael Coblenz
University of California –
San Diego, US
- Daniela Soares Cruzes
NTNU – Trondheim, NO
- Joanna Cecilia da Silva Santos
University of Notre Dame, US
- Sascha Fahl
Leibniz Universität
Hannover, DE
- Olga Gadyatskaya
Leiden University, NL
- Matthias Galster
University of Canterbury –
Christchurch, NZ
- Alex Gantman
Qualcomm Research –
San Diego, US
- Alessandra Gorla
IMDEA Software Institute –
Madrid, ES
- Ben Hermann
TU Dortmund, DE
- Kevin Hermann
Ruhr-Universität Bochum, DE
- Johannes Kinder
LMU München, DE
- Jacques Klein
University of Luxembourg, LU
- Piergiorgio Ladisa
SAP Labs France – Mougins, FR
- David Lo
SMU – Singapore, SG
- Tamara Lopez
The Open University –
Milton Keynes, GB
- Fabio Massacci
VU University Amsterdam, NL
- Tim Menzies
North Carolina State University –
Raleigh, US
- Mehdi Mirakhorli
Rochester Institute of
Technology, US
- Alena Naiakshina
Ruhr-Universität Bochum, DE
- Ranindya Paramitha
University of Trento, IT
- Liliana Pasquale
University College Dublin, IE
- Sven Peldszus
Ruhr-Universität Bochum, DE
- Henrik Plate
Endor Labs – Palo Alto, US
- Akond Rahman
Auburn University, US
- Awais Rashid
University of Bristol, GB
- Brad Reaves
North Carolina State University –
Raleigh, US
- Heather Richter Lipford
University of North Carolina –
Charlotte, US
- Daniel Votipka
Tufts University – Medford, US
- Sam Weber
Carnegie Mellon University –
Pittsburgh, US
- Charles Weir
Lancaster University, GB
- Dominik Wermke
CISPA – Saarbrücken, DE
- Laurie Williams
North Carolina State University –
Raleigh, US



Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics

Timothy Baldwin^{*1}, William Croft^{*2}, Joakim Nivre^{*3},
Agata Savary^{*4}, Sara Stymne^{†5}, and Ekaterina Vylomova^{†6}

- 1 MBZUAI – Abu Dhabi, AE. tbaldwin.net
- 2 University of New Mexico – Albuquerque, US. wacroft@icloud.com
- 3 Uppsala University, SE. joakim.nivre@lingfil.uu.se
- 4 University Paris-Saclay, CNRS – Orsay, FR.
agata.savary@universite-paris-saclay.fr
- 5 Uppsala University, SE. sara.stymne@lingfil.uu.se
- 6 The University of Melbourne, AU. ekaterina.vylomova@unimelb.edu.au

Abstract

The Dagstuhl Seminar 23191 entitled “Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics” took place May 7–12, 2023. Its main objectives were to deepen the understanding of language universals and linguistic idiosyncrasy, to harness idiosyncrasy in treebanking frameworks in computationally tractable ways, and to promote a higher degree of convergence in universalism-driven initiatives to natural language morphology, syntax and semantics.

Most of the seminar was devoted to working group discussions, covering topics such as: representations below and beyond word boundaries; annotation of particular kinds of constructions; semantic representations, in particular for multiword expressions; finding idiosyncrasy in corpora; large language models; and methodological issues, community interactions and cross-community initiatives. Thanks to the collaboration of linguistic typologists, NLP experts and experts in different annotation frameworks, significant progress was made towards the theoretical, practical and networking objectives of the seminar.

Seminar May 7–12, 2023 – <https://www.dagstuhl.de/23191>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases computational linguistics, morphosyntax, multiword expressions, language universals, idiosyncrasy

Digital Object Identifier 10.4230/DagRep.13.5.22

* **Editor / Organizer**

† **Editorial Assistant / Collector**



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 22–70

Editors: Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova



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

1 Executive Summary

Agata Savary (University Paris-Saclay, CNRS – Orsay, FR)

Timothy Baldwin (MBZUAI – Abu Dhabi, AE)

Joakim Nivre (Uppsala University, SE)

William Croft (University of New Mexico – Albuquerque, US)

License © Creative Commons BY 4.0 International license
© Agata Savary, Timothy Baldwin, Joakim Nivre, William Croft

The Dagstuhl Seminar 23191 entitled “Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics” was an accomplishment of long-standing efforts, initiated as early as in October 2018. We submitted at that time a Dagstuhl Seminar proposal which was selected to take place in Dagstuhl on June 21–26, 2020. Due to the Corona/COVID-19 pandemic, the event was first re-scheduled to August 29 to September 3, 2021, and finally transformed into a reduced online seminar under the same title on August 30–31, 2021 [1]. Despite its very reduced format, the seminar achieved part of its objectives and provided a proof of concept of the initial proposal. Following the encouragement from the participants, we re-submitted roughly the same proposal in November 2021 for a full-fledged on-site event. It was then selected to take place in Dagstuhl on **May 7–12, 2023**.

The objectives, following the initial 2018 proposal, were threefold:

- **Theoretical:** To deepen the understanding of language universals, and of linguistic idiosyncrasy in particular, so as to further promote unified modelling while preserving diversity.
- **Practical:** To harness idiosyncrasy in treebanking frameworks, in computationally tractable ways and, thus, to foster high quality NLP tools for very many languages.
- **Networking:** To promote a higher degree of convergence to universalism-driven initiatives, while focusing on three main aspects of language modelling: morphology, syntax, and semantics.

The program of the event followed the Dagstuhl model:

- A list of **recommended readings** was published prior to the event.
- Recordings from the **introductory talks**, given by the 4 organizers at the 2021 online seminar, ensured common understanding of the terminology, scope and challenges to address.
- **Personal introductions** of all participants helped achieve a community building effect.
- Six outstanding speakers were invited to give plenary **inspirational talks**.
- **Working groups** (WGs) were built in a bottom-up manner on the basis of discussion issues submitted by the participants. WGs ran in parallel, were coordinated and minuted by two co-leaders each, and were organized in the following settings.
- For **days 1 and 2** (Monday-Tuesday) the discussion issues were submitted by the participants prior to the event. On this basis 5 WGs were formed:
 - WG1 – *Below and beyond word boundaries* (co-leaders: Daniel Zeman and Reut Tsarfaty)
 - WG2 – *Annotation of particular kinds of constructions* (co-leaders: Manfred Sailer and Nathan Schneider)
 - WG3 – *Representing the semantics of MWEs* (co-leaders: Dag Haug and Nianwen Xue)
 - WG4 – *Finding idiosyncrasy in corpora* (co-leaders: Francis Bond and Nurit Melnik)
 - WG5 – *Methodological Issues and community interactions* (co-leaders: Amir Zeldes and Gosse Bouma)

- Day 3 was dedicated to **reporting**, collecting new issues and re-designing the WGs.
- As a result, 5 other WGs were formed for **days 4 and 5** and reported on on day 5:
 - WG6 – *Below and beyond word boundaries* (co-leaders: David Yarowsky and Omer Goldman), continuation of WG1
 - WG7 – *Construction grammar meets Universal Dependencies* (co-leaders: Lori Levin and Peter Ljunglöf), continuation of WG2 and WG4
 - WG8 – *To semantics and beyond!* (co-leaders: Archana Bhatia and Kilian Evang)
 - WG9 – *Cross community/formalism discussions (big, hairy problems)* (co-leaders: Chris Manning and Laura Kallmeyer)
 - WG10 – *Large language models (and other NLP tools)* (co-leaders: Francis Tyers and Mathieu Constant)¹
- Wednesday afternoon featured a **hike** in the surrounding countryside.
- The evenings were dedicated to **socializing**. This included a piano-violin duet, a guitar duet, a jazz improvisation, a swing dancing duet, and a choir singing songs suggested by the participants, in English, Georgian, German, and Latin (for the sake of language diversity!).

All the inputs and instantaneously produced outcomes (minutes, slides, useful links) are downloadable from our Wiki space.²

The event attracted **37 participants**. Their feedback during and after the event was mostly enthusiastic. At least one group formed at the event continues online meetings to further discuss the scientific challenges (representation of constructions in the Universal Dependencies framework).

Based on the reports submitted by the WG co-leaders and by individual proposers of discussion issues, we can estimate the extent to which the event achieved its initial objectives.

- On the **networking** side, the seminar brought together several pre-existing communities and allowed them to achieve synergies:
 - Linguistic experts specialized in analyzing constructions and collecting them in so-called constructicons, intensely collaborated with NLP experts, notably over the problem of how to represent constructions formally and query them in corpora.
 - While the community of typology experts was unfortunately under-represented (despite the best efforts of the organizers), the few attending experts were frequently consulted, which yielded several enlightening discussions.
 - The communities of Universal Dependencies (UD) and Universal Morphology (UniMorph) converged, even further than initially expected, around the problems of annotating subword units.
 - The communities of UD and PARSEME, which had started aligning objectives prior to the seminar, further strengthened coordination.
 - New links were established between PARSEME and the Universal Meaning Representation (UMR) community. This effect is important since the former models lexical and morpho-syntactic properties of MWEs, while the latter offers a framework for representing their semantics.
 - An unplanned networking effect also occurred between our seminar and Dagstuhl Seminar 23192 on “Topological Data Analysis and Applications”, running in parallel on the same site. Bei Wang Phillips (University of Utah – Salt Lake City, US) gave an evening invited talk to our invitees on the applications of topological methods to interpretability of word embeddings in distributional semantics.

¹ No report was provided for this group, which only met for a short session before splitting into other groups.

² <https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023/-/wikis/home>

- On the **theoretical** side, the seminar focused even more than expected on the notion of construction, which is broader and harder to capture than multiword expressions, and has been defined in wildly divergent ways across different communities. The confluence of different communities led to theoretical results including the following:
 - Steps were taken towards a formal definition of construction, as an expression in a formal graph language (similar to the one supported by the Grew-match corpus browser)
 - Advances in formalizing the notion of an “interesting” construction, which relates to the notion of idiosyncrasy, a core concept in a narrower guise in the multiword expression community
 - Formalizing the task of searching for “a similar but different construction” as an instance of the theoretical problem of approximate tree/graph matching
 - Progress towards understanding the notion of idiosyncrasy as an instance of rule breaking which is “creative” and “has a purpose”, as opposed to, for instance, plain grammar/spelling errors (rule breaking with no purpose)
 - Understanding idiosyncrasy via cross-linguistic triangulation – what is seen as idiosyncratic in one language can be systematic across many languages/language families (e.g. kinship terms)
 - Progress towards formalizing the annotation of semantics of UD and multiword expressions, especially for temporal and negation expressions

We also addressed a major challenge in language technology, which is a universal definition of the notion of a word. Namely, proposals emerging from WG1 and WG6 suggest that the difficult challenges for defining wordhood across languages should be alleviated by lifting the constraint of a rigid segmentation of a sentence into words prior to linguistic analysis. Instead, proposals of formats allowing different granularity of description items (below and beyond the word level) were suggested and discussed.

- On the **practical** side, discussions at the seminar led to a number of proposals for tools, procedures or practices to support interdisciplinary research. Some of these were tested out already in Dagstuhl, while others are being realized in follow-up activities to the seminar. The following is a non-exhaustive list of examples:
 - Practical steps were taken towards improved UD guidelines for multiword expressions, which will facilitate interfacing UD and PARSEME in the future.
 - Concrete guidelines were drafted for representing subword units in UD, which will facilitate the integration of resources from UD and UniMorph.
 - Discussions of construction-oriented UD guidelines (based on “a Swadesh list for morphosyntax”) resulted in a prototype implementation with links to annotation examples in different languages.
 - Discussions of future extensions of UD explored concrete proposals for new feature mechanisms to incorporate notions of constituency.
 - Practical exercises demonstrated how the grew-match system (and other search tools) can be used to search for constructions in linguistic corpora.
 - Participants discussed concrete proposals for automatically identifying idiosyncratic phenomena in corpora.

The survey organized by the Dagstuhl Officers shortly after the event shows very encouraging results (in most categories it was ranked higher than the average of the Dagstuhl Seminars from the past 60 days). The major drawbacks noticed by the participants were the insufficient number of experts in typology (less than 5%),³ and of young researchers (about 32%).

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7), pages 89–138.

³ This was notably due to the last minute cancellation, for personal reasons, by William Croft, one of the 4 co-organizers of the event.

2 Table of Contents

Executive Summary

Agata Savary, Timothy Baldwin, Joakim Nivre, William Croft 23

Overview of Talks

Adrian's Fish Tail: Compounds and Adnominal Possession Across Languages
Maria Koptjevskaja-Tamm 29

What Kinds of Parts do Multi-part Expressions Have?
Lori Levin 30

Universal Dependencies: Its Multilingual NLP Successes and Other Surprising Impacts
Christopher Manning 30

Indigenous Voices from the Past: Opening up the Florentine Codex to Modern Digital Scholarship
Francis M. Tyers 30

Working groups

WG1: Above and Below Word Level
Daniel Zeman and Reut Tsarfaty 31

WG2: Annotation of Particular Constructions
Nathan Schneider and Manfred Sailer 35

WG3: Semantics of Multi-Word Expressions
Dag Haug and Nianwen Xue 37

WG4: Finding Idiosyncrasy in Corpora
Nurit Melnik and Francis Bond 39

WG5: Methodological Issues and Community Interactions
Gosse Bouma and Amir Zeldes 44

WG6: Above and Below Word Level
David Yarowsky and Omer Goldman 46

WG7: UniCoDeX (Universal Construction Dependency Xrammar)
Peter Ljunglöf and Lori Levin 47

WG8: To Semantics and Beyond
Archana Bhatia and Kilian Evang 53

WG9: Fostering Corpus-based Typology ["Big Hairy Problems"]
Laura Kallmeyer and Christopher Manning 57

Open Problems

Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus
Alexandre Rademaker, Francis Bond, and Daniel Flickinger 60

NLP-based Study of Universals of Linguistic Idiosyncrasy
Agata Savary 64


Subword Relations, Superword Features
Daniel Zeman 67

Participants 70

3 Overview of Talks

3.1 Adrian’s Fish Tail: Compounds and Adnominal Possession Across Languages

Maria Koptjevskaja-Tamm (Stockholm University, SE)

License  Creative Commons BY 4.0 International license
© Maria Koptjevskaja-Tamm

As is well known, adnominal possession is not restricted to possession *stricto sensu*, but can cover many other relations, e.g., *Adrian’s house / sister / finger / school* etc. Typical possessors act as *anchors* or *reference point entities* for identification of the head, and the whole construction can therefore be said to denote *anchoring relations*. In many languages these are clearly distinguished from expressions used for *typifying relations*, i.e., for *qualifying* classes of entities via their relations to other entities. To give a couple of examples, typifying relations are expressed by noun phrases with adjectives derived from nouns in Russian (e.g., *kofe-jn-aja čaška* “coffee-ADJ-F.SG.NOM cup” = “a coffee cup” and *ryb-ij xvost* “fish-ADJ.M.SG.NOM tail” = “a fish tail”) and by noun-noun compounding in Swedish (*fisk+stjärt* “fish+tail” and *kaffe+kopp* “coffee+cup”), whereas the standard possessive construction in both languages contains the possessor in the genitive case. Other languages, however, utilize identical or, at least, very similar constructions for both anchoring and typifying relations. This is the case with adnominal dependents in the genitive case in Lithuanian, e.g., *Adrian-o namas* “Adrian-GEN house” = “Adrian’s house” and *kavos puodelis* “coffee:GEN cup” = “a coffee cup”. The Lithuanian phrase *žuvies uodega* “fish:GEN tail” may therefore refer to a tail of a particular fish, but also denote a class of tails that share certain properties without necessarily being a part of a fish (as those belonging to mermaids). The cross-linguistic variation exemplified by Russian, Swedish and Lithuanian is not surprising. The rationale for a similar treatment of anchoring and typifying relations is obvious – both types of adnominal dependents characterize entities via their relations to other entities. On the other hand, typifying adnominals differ in that 1. the dependent is not individualized; 2. the dependent-head combination refers to a subclass of a broader class and often functions as a classificatory label for it, suggesting that the dependent and the head together correspond to one concept; 3. the head cannot be identified via its relation to the dependent. In my talk I present a typology of the formal ways in which European languages deal with the distinction between anchoring and typifying relations and suggest several generalizations on the form-function correlations in this area. The insights gained from the talk may have consequences for syntactic and semantic annotation of multilingual language resources and tools, including the perennial issue of the border between words, multi-word expressions and regular syntactic phrases.

References

- 1 Maria Koptjevskaja-Tamm. 2005. Maria’s ring of gold: adnominal possession and non-anchoring relations in the European languages. In Kim, Ji-yung, Yu. Lander, and B. H. Partee (Eds.), *Possessives and Beyond: Semantics and Syntax*, pages 155–181. Amherst, MA: GLSA Publications.
- 2 Maria Koptjevskaja-Tamm. 2002. Adnominal possession in the European languages: form and function. *Sprachtypologie und Universalienforschung (STUF)*, 55(2), pages 141–172.

3.2 What Kinds of Parts do Multi-part Expressions Have?


Lori Levin (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Lori Levin

This talk distinguished multi-word expressions from multi-part expressions, where the parts are not necessarily words. The English causal excess construction (e.g., It was so big that it fell over) was presented as an example of a multi-part expression where the parts are words, parts of speech, morphological features, and abstract syntactic processes. The talk also addressed morphosyntactic strategies, specifically the issue of representing the meaning of constructions in strategy-neutral semantic frames.

3.3 Universal Dependencies: Its Multilingual NLP Successes and Other Surprising Impacts

Christopher Manning (Stanford University, US)

License  Creative Commons BY 4.0 International license
© Christopher Manning

Joint work of Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman
Main reference Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: “Universal Dependencies”, *Computational Linguistics*, Vol. 47(2), pp. 255–308, MIT Press, 2021.
URL http://dx.doi.org/10.1162/coli_a_00402

This talk outlines some of the design goals of the morphosyntactic annotation framework Universal Dependencies (UD), emphasizing how they differ from previous NLP practice and the natural predilections of linguists. In particular, for broad usage, “less is more”. In NLP, UD enables new tools that work well from raw text on dozens of languages and these tools are still improving and used, despite the dominance of Large Language Models in the field this decade. But a new arc of work is using UD resources for typological linguistic and psycholinguistic research. This work has somewhat different design goals, motivating a more surface structure oriented representation, as has been explored in Surface Syntactic Universal Dependencies (SUD), and there is room to do more here. But, overall, the success of UD and its wide adoption represents a success for linguistics – and a win for non-linguists who want simple, useful language processing tools.

3.4 Indigenous Voices from the Past: Opening up the Florentine Codex to Modern Digital Scholarship

Francis M. Tyers (Indiana University – Bloomington, US)

License  Creative Commons BY 4.0 International license
© Francis M. Tyers

Joint work of Francis M. Tyers, Robert Pugh, Valery A. Berthoud F.

The Florentine Codex is a bilingual text produced in Mexico in the 16th century. It describes the lives and beliefs of the Indigenous people who were living in the Valley of Mexico before the arrival of the Spanish. The text is in Nahuatl, the lingua franca of the area, and is accompanied by a translation or summary in Spanish. In this talk we describe the processing of the Nahuatl text and some linguistic issues that will need to be addressed in the annotation,

including relational nouns, functional incorporation and subordination and clause structure. We also describe efforts to translate from the Nahuatl of the era to modern varieties of Nahuatl spoken in the Sierra of Puebla.

4 Working groups

4.1 WG1: Above and Below Word Level

Daniel Zeman (Charles University – Prague, CZ) and Reut Tsarfaty (Bar-Ilan University – Ramat Gan, IL)

License © Creative Commons BY 4.0 International license
© Daniel Zeman and Reut Tsarfaty

Joint work of Daniel Zeman, Reut Tsarfaty, Omer Goldman, Sylvain Kahane, Sara Stymne, Francis Tyers, Ekaterina Vylomova, David Yarowsky

Goals

- To extend the UD representation in a way that can accommodate complex morphosyntactic phenomena (further discussed in WG6)
- To provide a proof of concept annotation for languages with no 1:1 mapping between segments and morphosyntactic nodes
- To improve parallel representation of argument structure between languages with radically different realization mechanisms (polysynthetic vs isolating)
- To discuss the role / contribution of derivational morphology (further discussed in WG6)
- To discuss how morphological marking complements compound and MWE
- To help field linguists who want to represent segmentation down to morphs
- To ultimately propose guidelines for cross-linguistically consistent annotation of polysynthetic (and in particular polypersonal agreement, noun-incorporating) languages

Representing phrase-level information in phrase-level features in UD, somehow comparable to representations for languages in which the same information is expressed in a single word (or a sublexical morph) was also discussed in WG4 and WG9.

Mechanisms that have been proposed

- “Empty nodes” (= abstract nodes) for indicating pronominal feature bundles
- “Empty nodes” carrying the lemma of an incorporated noun (not necessarily a linear segment)
- A layered representation of phrase-level features for each (lexical) node which is distinct from the word-level features that the lexical item contributes. (“Layered” as in layered features, which exist as language-specific in UD, and have also been introduced in UniMorph 4.0.)
- Linear segmentation (of clitics etc) is optional, not mandatory

Implication

To allow for the free use of these mechanisms to express non-explicit morphological phenomena, and still stay faithful to the UD guidelines, it has been proposed that the kind of representation we develop is a part of the enhanced, rather than basic, UD trees.

Empty (abstract) nodes are part of the enhanced UD graph (but not of the basic UD tree). (Note: They are called *empty* nodes in the current UD documentation but it is a misnomer because they often are not really empty, they may have a word form, UPOS tag etc. So we propose that UD should switch to a different term, such as *abstract* nodes.) Enhanced UD already contains abstract nodes that represent elided predicates in gapping constructions. If we now add abstract nodes for a different purpose, namely to represent segments of surface words, we need a way of distinguishing different types of abstract nodes. We also need to identify the surface word to which the segment (abstract node) belongs. Therefore, the segment-abstract nodes could be identified by `PartOf=ID` in the MISC column, where `ID` is the ID of a regular node in the basic UD tree.

Enhanced UD seems to be a suitable area where segments and their relations could be represented. It is still part of the UD specification (as opposed to add-ons built on top UD, using the CoNLL-U Plus format), meaning that such data could be part of official UD releases. At the same time, the Enhanced UD guidelines are less developed and frozen than the Basic UD guidelines, so it should be easier to add new guidelines here. Enhanced annotation is considered optional in UD corpora and can be easily separated from the basic annotation, hence the additional complexity can be completely transparent for users who are not interested in it. These are the main reasons why we propose to use the enhanced representation for subword relations, as opposed to the multi-word token mechanism (MWT), which would lead to the new annotation being visible also in the basic representation. (Moreover, the assumption about MWT is that an orthographic word is split into multiple morphosyntactic words. If we also use it to further split morphosyntactic words into morphs, we will have to solve the problem of distinguishing the different levels of granularity and different types of units.)

We have not come to a conclusion about the labels of the relations between subword units. We have identified two levels of granularity that might be of interest:

- Decomposition to lexical units: compounds and incorporation
- Complete segmentation to morphs

Technical details of the proposal

- Phrase-level features are put to the MISC column of the node whose subtree contains all nodes that belong to the phrase. Not all nodes in the subtree belong necessarily to the phrase, so the phrase has to be specified using node ids. It can be discontinuous. A possible representation is like this: `Phrase=1,3-5,7`
- The features have to be distinguished from other things that may be present in the same MISC cell. For UD-style features, prefixing the feature name with “Phrase” might work: `PhraseAspect=Prog|PhraseTense=Pres`. Datasets that use UniMorph-style features instead might need just one string: `PhraseUniMorph=V;PROG;PRS`.
- UD documentation should retire the term “empty node” and switch to “abstract node”, as these nodes often are not empty in the sense of not having any lexical or morphological value.
- If the dataset contains segmentation of syntactic words to smaller units that are not syntactic words, the abstract nodes should be used. Consequently, the segmentation is only visible in the enhanced representation while the basic tree stays reasonably simple.
- As abstract nodes are already used for other purposes than morphological segmentation, the nodes resulting from segmentation should be distinguished from other abstract nodes. Specifically, an abstract node (of the old kind) is not considered to correspond to any part of any surface token. The rule for the abstract nodes resulting from segmentation would

be that they appear between the node corresponding to the surface token they are part of, and the next node (abstract or regular). Each abstract node resulting from segmentation would have in MISC a reference to its corresponding surface token or syntactic word: `PartOf=2`.

- While an abstract segment-node knows to which surface token it belongs, it does not have to declare precisely which substring of the surface token it represents. Consequently, the order of the abstract nodes is not prescribed, although annotators are encouraged to follow the ordering of the corresponding morphs where it is observable.
- The abstract segment-nodes have to be connected in the enhanced graph, if for nothing else, then to maintain compatibility with Enhanced UD. It is yet to be seen whether and to what extent it is useful to define “syntactic” relations between the segments. As a minimum, the main lexical root has to be declared as the head and the other segments can be attached directly as its dependents. The relation labels (deprels) have to be taken from the UD repository. In some cases `compound` could be used. Cases with no better option could use a subtype of `dep`, such as `dep:infl`.
 - Languages with incorporation will attach the incorporated segments as core arguments of the verb: `obj`.
 - Another possible usage of the relations between segments is to show the order of derivation. CCG-like categories might then be added to MISC to signal that this morph combines with a `VERB` and once combined, the result is an `ADJ`.
 - Also note that some languages seem to have examples where another word would modify just one segment of the current word (Turkish *mavi arabadakiler* “those in the blue car”; lit. “blue car-in-those”). Here the enhanced graph would have an `amod` relation between *araba* and *mavi*, while in the basic tree it would go directly between *arabadakiler* and *mavi*.
- Two possible levels of segmentation are envisioned (both are optional, so people do not have to segment if they do not want to): 1. just split compounds (including incorporated nouns); 2. segment all the way down to morphs. Splitting compounds is useful for cross-linguistic parallelism. Complete segmentation is useful for field linguists who want to represent it, including features and glosses. And of course, there is a third level of segmentation, which already exists in UD: multi-word orthographic tokens are split to syntactic words; this one is done in basic UD and does not require abstract nodes.
- The `FORM` of an abstract segment can be empty (underscore), but it can be non-empty where it makes sense. The `LEMMA` should have a canonical form of that segment (e.g. English prefixes *in-* and *im-* would share the canonical form *in*).
- It is possible to say that some forms in a paradigm table are segmentable while others are not. For instance, one could say that the English verb *closed* is segmented to *close + d*, where the suffix is the bearer of the feature `Tense=Past`, but if the verb is irregular like *made*, it can stay unsegmented and bear the feature `Tense=Past` as a whole.
- If a user/application only wants to work with basic trees, the segmentation will be transparent for them. However, it is also possible that they want to work with the enhanced graph for other reasons but they still do not want to see the segmentation (or they want to see the compound level but not the complete decomposition to morphs). For that purpose we need an algorithm that will only extract the enhanced graph over full syntactic words. This has to be worked out. (Also, both the head of the subtree of segments and the original unsegmented word need to be attached somewhere in the enhanced graph.)

- In general it would be useful if one can say what is the backbone tree in the enhanced graph, and which edges are extra (creating reentrancies and cycles). This is currently not possible; one would have to modify the labeling schema for enhanced relations. Note that the backbone enhanced tree is not necessarily identical to the basic UD tree, as it can contain abstract nodes. The current proposal does not (yet) say how this should be done.
- Note that an abstract node may be also needed to represent a participant of an event (subject, object, oblique) which is not overtly represented as a word. This may correspond to an abstract segment-node under the verb, if the participant is referenced by the verbal morphology, but it is also possible that there is no trace of it in morphology and we still need to represent it e.g. to annotate coreference. (Conversely, in some languages a participant may be overtly referenced multiple times in the same clause: clitic doubling, full noun phrase, and verbal morphology.)

References to related UD issues

- <https://github.com/UniversalDependencies/docs/issues/701>
- <https://github.com/UniversalDependencies/docs/issues/703>
- <https://github.com/UniversalDependencies/docs/issues/704>

Future work

- Do a pilot segmentation of compounds in Parallel UD treebanks (PUD, 1000 news and Wikipedia sentences per language). Try a subset of PUD languages, especially those with lots of compounding, such as German and Swedish. Examine parallelism in word (morph) alignment.
- Convert UD morphological features in UD treebanks to UniMorph (update the conversion procedure, originally tested with UniMorph 2.0, to the latest set of UD features and the latest version of UniMorph (4.0). For each word form, compare the UD annotation with the corresponding entry in UniMorph word lists, if available. Possibly improve UD and/or UniMorph data based on the comparison. Prepare a new version of UniMorph where a new column will say for each word form its frequency in UD treebanks.
 - Expand the UD-UniMorph comparison to phrase-level morphology, such as periphrastic tense, aspect or voice.

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7), pages 89 – 138.
- 2 Martin Haspelmath. 2022. Draft. Defining the Word.
- 3 Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.
- 4 Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül özateş, Suzan üsküdarlı, Arzucan özgür, Tunga Güngör, and Balkız öztürk. 2022. *Enhancements to the BOUN treebank reflecting the agglutinative nature of Turkish*. arXiv preprint arXiv:2207.11782.
- 5 Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

4.2 WG2: Annotation of Particular Constructions

Nathan Schneider (Georgetown University – Washington, DC, US) and Manfred Sailer (Goethe-Universität Frankfurt am Main, DE)

License © Creative Commons BY 4.0 International license

© Nathan Schneider and Manfred Sailer

Joint work of Nathan Schneider, Manfred Sailer, Christopher Manning, Maria Koptjevskaja-Tamm, Jörg Bucker, Gosse Bouma, Lori Levin, Timothy Baldwin, Amir Zeldes

WG2 discussed challenges in syntactic annotation, with particular attention to the Universal Dependencies (UD) framework. The discussion featured issues with current UD annotation policies/guidelines as well as some broader challenges. The UD guidelines issues were (i) two possible changes to the relation inventory, and (ii) mischievous nominal constructions. The other main issue (not specific to UD) was (iii) negation and idiomaticity. discussion also touched briefly on the relationship between Construction Grammar and UD, but this was deemed more appropriate for a new working group. For this topic see the report from WG7.

Possible Changes to the UD Relation Inventory

Two specific objections to the current (UDv2) relations were discussed. First, there is a top-level **indirect object** relation (`iobj`), but it has been a source of confusion, owing in part to the diverse range of uses of the term “indirect object” in different linguistic traditions. We concluded that `iobj` should be considered for removal in UDv3, with multiple `obj` for double object constructions (different phrases meeting the UD criteria for `obj` could optionally distinguished via subtypes). Second, UD’s broad interpretation of the **adverbial clause** relation (`advcl`) lumps together adjunct clauses and complement clauses with oblique-like marking.⁴ The group agreed to endorse the idea of a new `cobl` relation in UDv3 that would target oblique subordinate clauses [1].

Mischievous Nominal Constructions

Various specialized patterns in names, numbers, dates, and measurements are known to be challenging for syntactic annotation [2, 3]. We discussed cases like “Eminent linguist Mr. Bill Croft”, where the phrase “eminent linguist” and the title “Mr.” are arguably modifiers as they can be freely omitted. We agreed that the UD guidelines should be changed to treat these as modifiers, adopting the suggested label `nmod:desc` (for “descriptor”) to distinguish these from other kinds of `nmod` [2].

More controversial were date expressions (“*July 30, 1980*” – should this be considered headed?), numbered entities (“*Room S108*”), and other names with an entity type (“*Michigan Street*”, “*Lake Michigan*”) or suffix (“*Richard III*”, “*BMW Inc.*”). Some of these cases are currently listed as `flat` in the guidelines. It emerged from the discussion that there were two different interpretations of `flat` held by members of the group. One interpretation is that `flat` is an “escape hatch” for tricky cases related to names (etc.) which do not easily fit general-purpose syntactic constructions. This is the expansive view of `flat`. Another interpretation is that `flat` should be reserved for expressions that are truly **headless**, defying all possible attempts to identify one part as the syntactically most important element. We leave these issues to further discussion.

⁴ Consider the sentence “As it was raining, we worked on improving annotation.” The second subordinate clause, “on improving annotation”, is not *adverbial* in the ordinary use of the term, but rather is a complement clause with prepositional marking. “As it was raining” is a true adverbial clause adjunct, and should remain `advcl`.

Negation and Idiomaticity

Negation is a grammatical category in all human languages. However, there is no explicit uniform encoding of negation in treebanks or in UD. Negation is also often expressed through a combination of different morphosyntactic elements, but, being a grammatical category, it is usually not considered a multiword expression or a phraseme. The purpose of the discussion point was to raise awareness for this phenomenon and to explore where to locate the representation of negation in annotated corpora. The following data can serve to illustrate the phenomenon.

Examples of morphosyntactic strategies for clausal negation include:

- In German, a clause can be negated by simply adding the negative adverb *nicht* “not”.
- In standard French, clausal negation is expressed by a combination of a pre-verbal particle, *ne*, and post-verbal *pas*, as in *Il ne pleut pas*. “It isn’t raining.” In this example, negation is expressed through two elements, both of which only express the category of negation.
- We also find so-called neg-words such as English *nothing*, i.e., indefinites that fulfill another dependency in a clause but, in addition, also mark negation, as in English *Alex said nothing*.

It is common that languages can use more than one of these elements to express a single negation, as in the French example *Personne ne fait confiance à personne*. (gloss: nobody NE makes confidence to nobody) ‘Nobody trusts anybody.’

Negation itself is not a uniform phenomenon on the meaning and usage side either, i.e., morphosyntactic negation marking can encode clausal negation, but also constituent negation, meta-linguistic negation, or expletive negation. A negative meaning can be explicit, or can be inferred. In addition, we find idiomatic combinations that express negation of various types: The German expression *einen Dreck* (lit.: “a dirt”) “nothing” marks a sentence as morphosyntactically negative. The English expression *I’ll be damned if . . .* has the effect that the *if* clause is semantically, though not morphosyntactically, negative. Finally, we find items that mark only pragmatically inferable negativity like sarcasm or irony. An example is German clause final, intonationally separated *. . . – also nicht!* (lit.: “dots thus not”) or *. . . und ich bin der Kaiser von China* “. . . and I am the emperor of China,” as in *Alex ist echt schlau – also nicht/und ich bin der Kaiser von China!* “Alex is really clever – certainly not!”

The various types of semantic negativity are also relevant for the treatment of MWEs, as we find lexical expressions that are restricted to occur in clauses with a particular type of negation, so-called *negative polarity items*. In the simplest case, there is a fixed expression with a particular negator, such as the German bound word *Unterlass* “stop”, which only occurs in the combination *ohne Unterlass* “without stop”. The German modal verb *brauchen* “need” requires a semantically negative clause, but the choice of the negation strategy is irrelevant. From a collocational or MWE perspective, *brauchen* would not be a single word, but rather form a collocation or MWE together with the semantic category of negation. Finally, NPIs such as English *lift a finger* or *fine* in cases with an explicit semantic negation, but also in some cases of (conventionalized) pragmatic negation, such as in denial (*But I DID lift a finger*). However, sarcasm and irony (i.e., only conversational pragmatic negation) do not seem to license any known NPIs.

Turning back to the seminar theme: Multiple exponence of negation is not marked as a dependency in UD, nor as a MWE in PARSEME. If Negation is a morphosyntactic category, shouldn’t it be marked? Yes, but probably as a grammatical category at the clausal level. As for the other phenomena, negation and negation-related phenomena might be a good motivation and testing ground for adding constructional information (as proposed by WG7) in addition to syntactic dependencies and classical MWEs to corpus annotation.

References

- 1 Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of COLING*, pages 3837–3852. Santa Fe, New Mexico, USA.
- 2 Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of UDW*, pages 160–172. Sofia, Bulgaria.
- 3 Daniel Zeman. 2021. Date and time in Universal Dependencies. In *Proceedings of UDW*, pages 173–193, Sofia, Bulgaria.

4.3 WG3: Semantics of Multi-Word Expressions

Dag Haug (*University of Oslo, NO*) and Nianwen Xue (*Brandeis University – Waltham, US*)

License © Creative Commons BY 4.0 International license
© Dag Haug and Nianwen Xue

Joint work of Dag Haug, Emily B. Bender, Archana Bhatia, Kilian Evang, Jan Hajic, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue

- In compositional frameworks MWEs are defined by non-compositionality.
- Also the approach taken by PARSEME: “Probably the most salient property of MWEs is semantic non-compositionality.”
- In PARSEME the distinction remains intuitive, because not tied to a particular mapping theory, while (non-)compositionality presupposes to a mapping from syntax to semantics (but not a specific syntactic or semantic framework).
- Even in settings where you have compositionality, there will be borderline cases like “*white wine*”, “*dry wine*”, “*red hair*”, etc.
- No clear way to distinguish a non-compositional analysis from one where *white*, *dry*, *red* has a special meaning only used in certain domains (subsecutive adjective)
- UMR annotates meaning directly, so MWEs cannot be defined by non-compositionality. Instead it is cases where several lexical items map to a single node in the semantic graph. It becomes an alignment issue between the syntactic structure and the UMR annotation.
- But arguably there are cases, where you would want to represent the idiom with several concepts, because it is internally modifiable.
- We decided to go through the categories of MWEs in PARSEME with regard to how they should (in principle) be represented semantically
- Two main questions:
 1. whether they are decomposable, i.e., which morphosyntactic components contribute to (distinguishable) meaning components;
 2. what these meaning components look like.

PARSEME has contributed considerably to MWE identification. In this context, semantic properties of MWEs have also been discussed. But a number of issues have been left open concerning the semantics of MWEs. In particular the actual semantic representations of MWEs in the context of annotating and processing MWEs have not been tackled so far.

We started from the MWE typology in the PARSEME annotation guidelines⁵ and discussed how reasonable semantic representations for the different types could be built.

⁵ https://parseme.fr/lis-lab.fr/parseme-st-guidelines/1.3/?page=030_Categories_of_VMWEs

LVC.full

In this type, the noun denotes an event and the verb contributes TAM features (in the sense of UMR). Argument structure is not modified, but “shared” between the LVC and the eventuality described by the second argument.

LVC.cause

The noun denotes an event or state, sometimes figuratively. Its first argument is not the first argument of the light verb. The noun does not provide a nominalization of the eventuality described by the entire sentence.

We discussed the following cases from the PARSEME guidelines:

- “*give a headache*” is actually at the same time LVC.cause and idiomatic MWE, at least in many cases. Question: How to annotate this? “*headache*” is an metaphor, which is embedded in a LVC.cause.
- “*grant rights*”: it is not clear whether this is a LVC. Maybe not, since the meaning of “*grant*” is relatively rich: transfer, causing somebody to have something ...
- “*provoke the destruction*”: rather not an LVC since the verb is a full verb with a specific meaning.
- “*give a bath*” might be a better example, compared to “*take a bath*”, which is not causative, but “*give*” still seems to carry some real meaning here?

It is an open question whether we consider “*take a bath*” and “*give a bath*” to be both instances of “*bath*” or do we want to say that the latter is “cause to take a bath”?

In sum, we are a little skeptical of the LVC.cause type, since in many of the examples, the verb carries too much meaning to be a light verb. The discussion was then partially continued in WG8.

VIDs (verbal idioms)

A big challenge to find an appropriate concept. Do we combine concepts, or do we make new concepts? For example for the German idiom “*ein kleines Vöglein hat mir gezwitschert*”: Is the concept “little-bird-tweeting” or “being-told-in-secret”?

A second important question is how to deal with modification of idioms? E.g. “*jump on the bandwagon*”, “*jump on the AI bandwagon*”, “*jump on the latest AI bandwagon*”. UMR creates a new concept that takes arguments like “AI”, “latest”, but this doesn’t work well if the process is truly recursive. The Düsseldorf group decomposes such idioms so that “jump” means “join”, “bandwagon” means “fad”. Other similar examples are “*pull strings*”, “*pull family strings*”, “*take the project under its federally funded wing*”, “*who made them kick their respective buckets*”.

Here are more modification examples, from Riehemann 2001 (found in the North American News Corpus):

- Meanwhile modern navigation and transport ensured that **no significant stone on the planet was left unturned**, no nation or tribe undiscovered or undocumented.
- King and Alexander, who sued each other after their bitter 1992 divorce, **buried the legal hatchet** in May.
- Russia cannot be allowed to **call NATO’s shots**.

The problem in such cases is to make it possible for the modifier to access the slot that it modifies. But it can also be problematic to construct the correct semantics when we have cases where a modifier appears relatively low in the structure of the idiom, but actually modifies the idiom as a whole (external modification):

- “leave a (very, extremely, horridly) bad taste in someone’s mouth”: Is the position of the modifier wrt the whole idiom interesting or difficult here?
- “they want to have their political cake and eat it too”

WG4 also looked at internal modification of verbal idioms.

IRVs: inherently reflexive verbs

The semantics of this group seems clear: the reflexive (by definition of inherently reflexive) does not contribute a participant.

More interesting is the case of reciprocals such as “*sich treffen*” vs. English “*meet*”, where you can have both “We met” and “We met each other”. Should all of these have the same argument structure, i.e. should there be an implicit second argument in the cases like “We met”? The consensus of the group was yes.

We also discussed examples like “*find oneself in a difficult situation*”. These are syntactically idiosyncratic, but not semantically not idiomatic.

In sum, we saw little need to deal with inherently reflexive verbs as MWEs.

Verb particle constructions

“*run over*”. In MRS, the verb is the concept and the particle is selected. In UMR, the verb and the particle are concatenated and an aspect feature is (possibly) added. In DRT, the particle has empty semantics and there is just one semantic unit.

Multi-verb constructions

“*make do*”. These are very rare, typically opaque and form one unit with just one semantic contribution which cannot be internally modified.

Inherently adpositional construction

“*rely on*”. In both UMR and DRT, the verb provides the concept and the selected adposition mediates a thematic role which is governed by the verb.

4.4 WG4: Finding Idiosyncrasy in Corpora

Nurit Melnik (The Open University of Israel – Raanana, IL) and Francis Bond (Palacký University Olomouc, CZ)

License © Creative Commons BY 4.0 International license
© Nurit Melnik and Francis Bond

Joint work of Timothy Baldwin, Archana Bhatia, Nina Böbel, Francis Bond, Mathieu Constant, Daniel Flickinger, Maria Koptjevskaja-Tamm, Peter Ljunglöf, Nurit Melnik, Alexandre Rademaker, Agata Savary, Leonie Weissweiler

Introduction

The working group was formed to address three main discussion topics:

- Discovering linguistic idiosyncrasy (Nurit Melnik)
- Identifying non-compositional MWEs in text (Francis Bond)
- NLP-based study of universals of linguistic idiosyncrasy (Agata Savary)

Over the course of six sessions, the group engaged in intense discussions which included but were not limited to these issues. This document aims to provide a comprehensive overview of our discussions and suggest potential directions for future research.

Discussion topics

The central theme of our discussions revolved around the concept of idiosyncrasy. We were fortunate to have a diverse group of participants with expertise in various fields such as NLP, computational linguistics, theoretical linguistics (particularly Head-driven Phrase Structure Grammar and Construction Grammar), typology, grammar engineering, and more. This diverse range of perspectives greatly enriched our discussions.

Our conversations were structured around three interrelated topics.

Idiosyncratic vs. regular. Although we did not want to delve into the question of what is the precise definition of idiosyncrasy the issue hovered over our discussions.

How to find idiosyncrasy. We brainstormed different methods, automatic and manual, for finding cases of idiosyncrasy within a single language as well as across different languages.

Accounting for mismatches between levels. We mostly discussed challenging cases of mismatches involving multi-word expressions (MWEs).

Idiosyncratic vs. regular

Our initial strategy was not to spend time trying to come up with an precise definition of idiosyncrasy. Instead, we began by looking for phenomena which we would intuitively identify as idiosyncratic. We realized that such phenomena generally stood in opposition to what is “regular”.

One type of idiosyncratic phenomena are syntactic structures which are not part of a regular core. These structures are often discussed in the Construction Grammar literature. For example, the English *the X-er the Y-er* construction has a very unique morphosyntactic pattern.

- (1) The more the merrier. [en]

Another example is the *do-be* construction, which is subject to various idiosyncratic constraints that are not derived from general properties of the language.

- (2) What you have to do is *(to) get ready. [en]

Some constructions have a regular syntactic structure, but they can host lexical items which are not expected to appear in them.

- (3) I sneezed the foam off my cappuccino. [en]

Regular syntactic structures may also have idiosyncratic meanings. Thus, for example, in the following exchange, the meaning of the coordination of the two identical Swedish adjectives meaning ‘happy’ is ‘not so good’.

- (4) Are you happy? [en] Happy and happy [sw]

Another type of idiosyncrasy involves exception to rules. For example, although adjectives in English precede the nouns that they modify, the adjective *enough* can only appear post-nominally.

- (5) a. That sounds good enough. [en]
 b. *That sounds enough good.

We also considered the notion of idiosyncrasy from a typological cross-linguistic perspective and asked whether there are particular domains which are more susceptible to idiosyncrasy. One phenomenon that we focused on was what is referred to as *pro*-drop, namely the ability to “drop” pronominal subjects, which is found in Japanese but not in English.

- (6) a. 着いた tuita “ ϕ arrived” [ja]
 b. I arrived. [en]

The common wisdom is that *pro*-drop depends on the richness of the morphology. However as the Japanese example indicates, this is not necessarily the case and this property is more idiosyncratic than is believed.

Other domains which were mentioned as potentially exhibiting idiosyncrasy were: temporals, kinship terms, negation, existentials, possessives, inherently reflexive verbs. This topic was later developed in the discussions of WG7, where they proposed to compile a typologically informed collection of “meta-constructions” (or domains) and the “morphosyntactic strategies” which languages employ to realize them.

How to find idiosyncrasy

Written sources. The most obvious resource for finding out about language are grammar books and language teaching materials. This type of literature often expands on the distinction between phenomena which can be accounted for by rules and phenomena which constitutes exception to these rules, i.e., idiosyncrasy. In addition to language teaching material, errors in learners’ output may also indicate idiosyncratic phenomena.

Published papers in linguistics, mainly in Construction Grammar and related frameworks, present and analyze constructions (e.g., the English *the X-er the Y-er, do-be* and *way* constructions). Naturally, these papers can be used as resources for finding idiosyncratic constructions. Moreover, it was suggested that it could also be possible to automatically parse and mine linguists’ papers for examples of such phenomena.

Albeit not written per se, an additional resource for finding idiosyncrasies are online constructicons which are developed for various languages.⁶ Some examples are:

- Brazilian Portuguese constructicon: <https://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main>
- Swedish constructicon: (SweCcn): <https://spraakbanken.gu.se/eng/sweccn>
- Russian constructicon:
<https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus&lang=swe&advanced=false&searchTab=special&hpp=25&extended=and%7Crus-construction%7Cequals%7C&page=1>
- English constructicon:
<http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/index.html>

⁶ This topic was further discussed in WG7 meetings.

Parsing corpora. Written sources are useful for finding *known* cases of idiosyncrasy. A bigger challenge is discovering new ones. For this purpose we discussed methods of using NLP tools to agnostically explore corpora and identify language use which cannot be accounted for by “regular” grammar rules.

Broad-coverage precision grammars as rule-based computational grammars are a useful tool for identifying cases of grammatical “rule breaking” phenomena. We discussed an experiment performed by Baldwin [1], who ran the English Resource Grammar over a sample of the British National Corpus and conducted a thorough error analysis. One example of an idiosyncratic construction that was identified by virtue of its rejection by the grammar is the *do-be* construction, e.g., “the thing we should do is buy a new car”.⁷

Recent neural/deep learning approaches for NLP do not share this characteristic; they are capable of parsing everything – from the most regular to the most idiosyncratic, as well as ungrammatical. However, there may be a way to probe the “black box” and to look at probabilities, surprisal, entropy, perplexity scores or confidence levels in order to identify instances that challenge models which are based on statistic regularity. It may be possible to use an incremental parser and look for spikes which indicate unexpected semantic or syntactic co-occurrences. Some group members expressed an interest in further exploring these methods.

Accounting for mismatches between levels

One general type of idiosyncrasy that we discussed is mismatches between levels (phonology, morphology, semantics, syntax), particularly in the domain of MWEs. Following are some more specific cases that were presented as particularly challenging for NLP.

MWEs can encode single predications. In other words, MWEs can appear in the sense hierarchy in the same way as a single word. For example, the meaning of the English MWE *look up* is similar to *phseek*, yet it is subject to idiosyncratic syntactic constraints.

- (7) a. I looked up the word. [en]
 b. I looked the word up.
 c. *I looked up it.
 d. I looked it up.

Moreover, some MWEs may look like “regular” phrases but as MWEs their parts do not exhibit “regular” syntactic behavior. For example, the relationship between the noun and adjective in the MWE *hot dog* is not intersective modification; *hot dog* is not a $\text{dog}_{n:1}$ that is $\text{hot}_{a:1}$, but a kind of sausage ($\text{hot_dog}_{n:1}$). This effects the syntax:

- (8) a. # I ate a very hot dog. [en]
 b. I ate a very hot pizza.

However, even if we think of a MWE as a single predicate, bits of it can still be accessed and modified. In some cases, internal syntactic modification becomes semantically external. For example, although *Texan* is modifying the *dust* part of the MWE in (9a), it is interpreted as modifying the entire meaning of the MWE (9b).

- (9) a. He bit the Texan dust. [en]
 b. He died in Texas.

⁷ See discussion in [2].

It should be noted that internal modification being interpreted externally is not only a feature of MWEs. Consider the following example.

- (10) a. I have an occasional drink. [en]
 b. I drink occasionally.

Syntactically, the adjective *occasional* is modifying the noun *drink*, but semantically, what is occasional is not the drink but rather the entire drinking event. Accounting for mismatches between levels was also discussed in WG1, WG3, WG9.

Future work

Following our WG discussions several of the participants joined forces with members of WG2 (*Annotation of particular kinds of constructions*) to discuss the relation between Construction Grammar and Universal Dependencies, thus forming WG7, initially named “CxG meets UD”. Some of the topics which we identified as ideas for future work were discussed in WG7 sessions.

- Create a cross-linguistic idiosyncrasycon/constructicons even the ‘rare’ constructions are often cross-linguistic
- Sense-tag corpora

Other ideas for future work were related to the challenge of automatic discovery of idiosyncrasy.

- Investigate the feasibility of detecting surprisal in automatic parsing
- Do 10-fold cross-validation and mine the errors
 (But how do we distinguish errors, creativity and idiosyncrasy?)

Finally, one issue that came up and prompted ideas for future research addressed the notion of idiosyncrasy with respect to large language models (LLMs): How do LLM-generated texts compare to natural texts in terms of the frequency and distribution of various idiosyncratic phenomena such as idioms, MWEs and *that*-less relative clauses?

References

- 1 Timothy Baldwin, John Beavers, Emily M Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2005. Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar—and the corpus. In *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 49–70.
- 2 Dan Flickinger and Thomas Wasow. A corpus-driven analysis of the do-be construction. 2013. In Philip Hofmeister and Elisabeth Norcliffe, editors. *The core and the periphery: Data-driven perspectives on syntax inspired by Ivan A. Sag*, pages 35–63. Centre for the Study of Language and Information.

4.5 WG5: Methodological Issues and Community Interactions

Gosse Bouma (University of Groningen, NL) and Amir Zeldes (Georgetown University – Washington, DC, US)

License  Creative Commons BY 4.0 International license

© Gosse Bouma and Amir Zeldes

Joint work of Gosse Bouma, Amir Zeldes, Carlos Ramisch, Agata Savary, Emily Bender, Joakim Nivre, Lori Levin, Sara Stymne, Teresa Lynn

The group discussed several issues which roughly fit into the following topics:

- **UD Maintenance:** Current UD treebank maintenance is approaching a crisis, as more and more resources are neglected, and active developers are in charge of amounts of data they cannot update by themselves when guidelines are revised. The group discussed strategies to recruit and teach new UD annotators, with several recommendations emerging:
 - We must motivate newcomers to contribute, for example by:
 - * creating live public leaderboards reflecting committed contributions
 - * highlighting the status of up-to-date resources on the main UD page
 - * finding venues (workshops, special issues) to publish papers about maintenance efforts (Findings of UD?)
 - * creating a designation for UD contributors that can easily be put on CVs (“UD editor” or similar)
 - * approaching motivated potential contributors, incl. native speaker linguists, retired linguists, Master’s program students looking for projects and others
 - * organizing tutorials at venues like ESSLI or the Linguistic Institute, advertising at the Linguistic Olympiad, or possibly leveraging networks like Unidive
 - Use GitHub issues more clearly to recruit maintenance workers (“help wanted”)
 - More challenging possibilities with new developments:
 - * set up easy-to-use interfaces where UD repo contents can be easily imported for editing, so that willing contributors only need to receive a login
 - * figure out a gamified environment where multiple contributors compete for inter-annotator agreement or other score metrics
- **A Swadesh list of constructions:** A list of abstract typologically widespread constructions with examples of UD annotations in multiple languages would be useful for a variety of purposes, including didactic venues, documentation, validation and consolidation of practices for new languages joining UD (this was also outlined in WG7 and WG9). Such constructions could include “predicative possession” (x had y) or “property comparison” (x is ADJ_{er} than y), and the examples would span strategies from across languages. The group discussed several aspects of the creation of such a resource:
 - Data collection should be simplified as much as possible to lower the barrier for entry and increase likelihood of contributions by asking contributors to simply supply a Grew Match query for each construction and to assert that, e.g. the first 3 hits are correct examples (otherwise, they should specify in a separate column the match numbers of some correct hits)
 - Problems raised included the likelihood that examples would omit interesting variants in each language which exceed 3 basic examples, or cases about which the list simply does not ask (Lori pointed out that for some questionnaires, e.g. the Lingua checklist used by Comrie and Smith, possibly thousands of permutations of morphological categories would need to be considered)

- The group converged on the idea that asking for maximal diversity in the example types was desirable, with the understanding that more variants will inevitably be missed, but we should still focus on getting some, rather than no information
- Strategies to select the constructions were also discussed:
 - * A grammar description framework, such as the Grammar Matrix (<https://matrix.ling.washington.edu/>) could be used to create an outline of a language type, and each distinct type implicates distinct constructions of interest
 - * The UD Cairo corpus of 20 example sentences potentially contains a good list of candidate constructions
 - * The numbered list of constructions in Croft’s book can be used as a starting point as well
- **NLP for Typology:** There is a growing interest in using UD treebanks for typology (witness work by Levshina [2], token-based typology, and others). This raised various discussion points:
 - UD treebanks are often limited in size, and may not be comparable in genre and register across languages. Can we use NLP (i.e. automatic annotation or methods for selecting and annotating comparable fragments across languages)?
 - UD annotation was originally not (or not exclusively) designed with this application in mind, and it misses some dimensions that could be very valuable for typology. Treebanks could be made more valuable for typologists by adding construction level annotation (ie explicit annotation of clause types such as questions or passives, even if this is sometimes implicitly encoded) and/or annotation beyond syntax (semantic dimensions, information packaging)
 - Using UD for cross-lingual comparison does presuppose that we know when phenomena are comparable across languages (see Haspelmath’s discussion of comparative concepts [4]), but this may not always be the case for decisions made in UD annotation.
- **Unifying UD and PARSEME:** This topic is discussed in some detail in Savary [3], PARSEME meets Universal Dependencies. Even though PARSEME and UD are two orthogonal annotation layers, they can be merged, e.g. by adding a ParseM column to UD CONLL-U format. Discussion points:
 - The notion “MWE” is used somewhat sloppy in UD to group compound, fixed, and flat relations, where compound definitely should not be under the rubric MWE, and fixed and flat may be better represented as “head-less” rather than MWE.
 - PARSEME covers some MWE types that do not fall under one of the UD relations compound, fixed, or flat.
 - In the future, PARSEME aims to include nominal constructions as well. This could be coordinated with proposals in WG2 for the analysis of mischievous nominal constructions [1].

References

- 1 Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*
- 2 Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23, pages 533–572
- 3 Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions. *Northern European Journal of Language Technology*, 9, 1
- 4 Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, v.86, pages 663–687

4.6 WG6: Above and Below Word Level

David Yarowsky (Johns Hopkins University – Baltimore, US) and Omer Goldman (Bar-Ilan University – Ramat Gan, IL)

License © Creative Commons BY 4.0 International license

© David Yarowsky and Omer Goldman

Joint work of Goldman, Omer; Yarowsky, David; Zeman, Daniel; Stymne, Sara; Vylomova, Ekaterina; Kahane, Sylvain

WG6 was a direct continuation of WG1 from the 1st half of the seminar and was tasked with proposing enrichment for Universal Dependencies (UD) above and below the word level. While WG1 discussed both addition of phrase level features above the word level and morphological segmentation below it, WG6 focused solely on the morphological segmentation, hashing out the different possible implementations and discussing the concrete steps needed to be taken to achieve it. In essence, the participants proposed a two-step process of alignment between UniMorph and UD that will allow usage of UniMorph data in UD trees.

UD-UniMorph alignment

In the current state of affairs, both UD and UniMorph contain morphological data but they tag it according to different although generally compatible annotation schemas. Therefore we suggest an effort to align these two resources. This effort will probably not include any changes needed to be done in UD.

As a first step we propose to map the differences between UD and UniMorph, both in terms of the features used and the structure of the features. The mapping will be done as part of an effort to extract morphological data from UD to UniMorph. Once it'll be done, we could produce a list of the changes we believe are required and approve through the relevant committees in the management of both datasets.

Having both projects using the same annotation schema for morphological data will make it the cooperation between the project organizers smoother.

Morphological Segmentation in UD

Currently, morphological information in UD is confined to the “feats” column where features are attributed to the entire node, and it makes languages seem different depending on the frequency of white space usage. Decomposing words into morphemes and marking the relations between morphemes with dependency arcs as done between words, will equate the structure of different languages and will make it easier to typologically compare languages.

The group proposed segmenting words into morphemes, mostly using the segmentation files from UniMorph, currently existing for about a dozen of languages. The lemma of each morpheme will be its canonical form and the features be associated only with the relevant morpheme rather than with the entire word.

There were 2 main implementation options discussed: one where the content of each “morphological node” is a morpheme, and one where it is a truncated word. For example, a word like “*industrialization*” will be decomposed into “*industry*”, “*-al*”, “*-ize*” and “*-ation*” according to the first option, and to “*industry*”, “*industrial*”, “*industrialize*” and “*industrialization*” according to the second. The benefits of the latter is that it does not require the annotators to decide on the “canonical form” and whether it even exists and it is more closely aligned with the UniMorph derivational morphology data, but on the other hand this option is less intuitive and diverges to some extent from the structure of UD for words.

This discussion should be revisited after the completion of the first phase.

4.7 WG7: UniCoDeX (Universal Construction Dependency Xrgrammar)

Peter Ljunglöf (University of Gothenburg, SE) and Lori Levin (Carnegie Mellon University – Pittsburgh, US)

License © Creative Commons BY 4.0 International license

© Peter Ljunglöf and Lori Levin

Joint work of Baldwin, Timothy; Bhatia, Archana; Böbel, Nina; Bond, Francis; Bucker, Jörg; Constant, Mathieu; Flickinger, Daniel; Kahane, Sylvain; Levin, Lori; Ljunglöf, Peter; Lynn, Teresa; Melnik, Nurit; Nivre, Joakim; Rademaker, Alexandre; Sailer, Manfred; Savary, Agata; Schneider, Nathan; Weissweiler, Leonie; Zeldes, Amir

Introduction

This is a summary of the discussions that took place in Working Group 7 during the Dagstuhl Seminar 23191 *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics*.

WG7 was formed from the initial working groups 2 (*Annotation of particular kinds of constructions*) and 4 (*Finding idiosyncrasy in corpora*), who independently of each other realised that they wanted a group that was focused on the relation between Construction Grammar and Universal Dependencies. So for the second half of the seminar we formed WG7 with the initial name “CxG meets UD”, but we changed the name of the group to the more catchy UniCoDeX (*Universal Construction Dependency Xrgrammar*).

The group met for four sessions and had intense discussions which resulted in the formation of three interconnected tasks. This document tries to reflect the results of our discussions and the way forward.

Overview of discussion topics

The overall question of WG7 was how to connect Construction Grammar (CxG) with Universal Dependencies (UD) in a way that both Construction Grammar projects and Computational Linguistics projects can benefit from. Our discussions were organized around three interrelated topics:

A. Documenting morphosyntactic diversity in UD.⁸ How can UD annotations and guidelines be improved to better reflect typological differences between languages (also discussed in WG1)?

B. Standard for a construction annotation layer in UD. How could construction annotations be added to augment UD treebanks? What annotation standard would be needed (also discussed in WG5 and WG9)?

C. Searching for constructions in UD treebanks. How could UD treebanks be queried for interesting examples of a given construction (the topic was also discussed in WG4, WG5, WG9)?

Topic A. Documenting morphosyntactic diversity in UD

The general goal of this topic is to verify the typological coverage of UD, increase its consistency, and advise users on how to analyze constructions in their languages. There are two important concepts: *meta-constructions* framed in comparative terms based on [1]; and

⁸ Previous name: “Typologically valid UD annotation guidelines”

morphosyntactic strategies that specific languages may employ. The group working on this topic will create a collection of annotation examples and guidelines for different languages and meta-constructions:⁹

- This collection corresponds to the “Swadesh list” for morphosyntax from WG2 – we dub it the *Nivre list* since it was proposed by Joakim Nivre. Using this checklist, treebankers can determine which morphosyntactic strategies are used in their languages for each meta-construction and how to annotate them.
- One component of this collection is a spreadsheet with languages in the rows and meta-constructions in the columns, where the cells contain links to queries yielding annotation examples and notes indicating which morphosyntactic strategy is illustrated in each example.
- Another component is a table of meta-constructions and morpho-syntactic strategies, with examples from different languages and language families.

This collection will be used to promote consistent and typologically informed coverage of morphosyntactic strategies in UD and update the general UD annotation guidelines from a typological perspective, including morphosyntactic strategies as defined by [1].

- This will help UD to find areas where we can improve current explanations and analyses to be more typologically oriented.
- It will also be of help when extending the guidelines to improve the coverage of explanations.
- It will ensure that it is possible to represent all (or at least most) known typological diversity in UD.
- It will facilitate using UD for research in language typology.

Example(s). Object predication is a cross-linguistic meta-construction which uses different strategies in different languages. In this meta-construction a semantic object is information-packaged as a predicate. (This is conventionally called a predicate nominal.)

- English: verb copula strategy (“Dani is a student”)
- Russian: zero strategy (“Dani student”)
- Hebrew: pronoun copula strategy (“Dani hu student”)
- Classical Nahuatl: inflect the noun as a verb: (“ni-ticitl”, 1sg-doctor)

We want annotation guidelines for each language/strategy so that an annotator will not necessarily go to an English treebank by default. The table of morphosyntactic constructions and strategies should be able to guide a treebanker to the right strategy. The spreadsheet should guide the treebanker to examples from languages that use that strategy, which illustrate how to make dependency trees for that strategy.

Future work. People who are interested in working on this topic after Dagstuhl:

- Lori and Joakim (group leaders)
- Alexandre, Amir, Archana, Jörg, Leonie, Nathan, Nurit, Sylvain

As a concrete first step the group agreed to do the following in the near future:

- add at least 10 languages and 10 meta-constructions to the Nivre list.
 - We will start with basic meta-constructions like clausal possession, comparison, and argument alignment (accusative or ergative).
 - We will write guidelines for each meta-construction and each strategy.
- when this has been finished, there will be an internal review within the group to decide about future steps.

⁹ [1] calls them “constructions”, but we refer to these as meta-constructions to differentiate them from language-specific constructions in the Construction Grammar sense.

Topic B. Standard for a construction annotation layer in UD

The general goal of this topic is to develop recommendations for how to annotate UD treebanks with constructions. They should be useful for several different use cases, workflows and granularities, such as:

- Use cases: we could be building a new construction from the ground up and want to come up with good definitions of constructions, or we might already have an existing construction which we want to use for annotation or extend with new constructions
- Different annotation granularities: from the coarsest level (to just annotate the head of a construction with its name), to the most fine-grained (to also annotate all the construction elements with their names and spans within the sentence)
- Different workflows: people might want to annotate one construction at a time, or several at once – or they might want to use an iterative approach where they start with coarse-level annotation and then refine them

Future work. People who are interested in working with this topic after Dagstuhl:

- Leonie (group leader)
- Alexandre, Amir, Archana, Francis, Lori, Manfred, Nathan, Nina, Nurit, Peter, Sylvain

As concrete first steps the group agreed to annotate a limited family of constructions in different languages, with the hope of writing a joint paper during autumn.¹⁰ Some initial ideas of constructions that could be interesting to annotate were:

- age constructions, rates (mph etc), comparatives, resultatives, ...
- idiosyncratic, lexicalised constructions, such as X-and-X (Swedish), N-über-N (German), N-after-N (English)
- cross-linguistically common constructions such as types of conditionals, possession, comparison etc. (i.e. exponents of meta-constructions, see above)

The group will continue discussing topics such as:

- naming convention for constructions
- integration with existing annotation tools
- annotating/marketing candidates that have been checked and are not a certain construction
- which token in the UD tree should be annotated with the construction?
 - the natural choice is to annotate the token that is highest in the UD tree – but it is unclear what to do if the construction covers disconnected parts of the UD tree

The group agreed to postpone more complicated questions, such as:

- how to handle cross-sentential constructions
- how to handle nesting and composition of constructions
- how to handle constructions on different levels of granularity (more specific vs. more general constructions)

Topic C. Searching for constructions in UD treebanks

The general goal of this topic is how to formulate search queries that can locate interesting examples of a given construction. This is very closely related to the previous two topics, as they all depend on being able to search for constructions in treebanks.

¹⁰ Possibly targeting LREC-COLING 2024, with submission deadline October, or ICCG 2024 with deadline in spring.

The group discussed some issues that arise when it comes to formulating search queries, such as:

- we want guidelines that help people with writing and refining queries
- we want (semi-)automatic techniques for extracting relevant search queries from an existing Constructicon entry
- precision/recall tradeoff: it is probably more important to have a good recall than good precision, but it is usually easier to improve the precision by modifying a query
- possible strategies to increase the recall can be to use approximate tree matching or to loosen some constraints in the query

Future work. People who are interested in working with this topic after Dagstuhl: the same as for topic B, with Leonie as group leader.

In the beginning this topic will be closely related with topic B. To be able to annotate the treebanks we will have to formulate search queries that can find potential candidates. While doing this iterative process for a diverse set of constructions in different languages we hope to come up with more general guidelines on how to write construction queries.

Related work/links

The following are the existing Constructicons that we are aware of:

- English: Berkeley FrameNet Constructicon: <http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/>
- English: Birmingham English Constructicon: <https://englishconstructicon.bham.ac.uk/>
- English: CASA (FAU Erlangen-Nürnberg): <https://constructicon.de/>
- German: FrameNet-Konstruktikon (HHU Düsseldorf): <http://framenet-constructicon.hhu.de/>
- Swedish: Svenskt konstruktikon (Univ. of Gothenburg): <https://spraakbanken.gu.se/karp/#?mode=konstruktikon>
- Brazilian Portuguese: FrameNet Brasil (FU Juiz de Fora): <https://www2.ufjf.br/framenetbr-en/>
- Japanese: Japanese FrameNet (Keio University): <https://jfn.st.hc.keio.ac.jp/>
- Russian: Russian Constructicon (UiT Arctic University of Norway): <https://constructicon.github.io/russian/>
- Most of the different constructions were presented at the Constructicon Alignment Workshop (CAW, December 2022), and video recordings are available here: <https://www.globalframenet.org/caw2022>

Croft’s “Morphosyntax: Constructions of the World’s Languages” [1] contains a glossary of different *comparative concepts* (meta-constructions, strategies, information packaging, etc.), and this glossary is available online here:

- Interactive interface: <https://spraakbanken.github.io/ComparativeConcepts/>
- GitHub repo: <https://github.com/spraakbanken/ComparativeConcepts>

Finally, here is a list of different search engines for corpora, tools and treebanks, that can be used to find constructions:

- Grew-match: <https://match.grew.fr/>
- SPIKE (query-by-example): <https://spike.apps.allenai.org/>
- DepEdit: <https://gucorpling.org/depedit/>
- UDAPI: <https://udapi.github.io>

- Korap (IDS-Mannheim: <https://korap.ids-mannheim.de/> and <https://github.com/KorAP/>)
- Corpus workbench (CWB, useful for larger corpora) – several sites use CWB, such as:
 - Språkbanken Korp (Univ. of Gothenburg): <https://spraakbanken.gu.se/korp/>
 - CQPWeb (Lancaster Univ.): <https://cqpweb.lancs.ac.uk/>
 - CWB source code can be found here: <https://cwb.sourceforge.io/>

Proposed standard for construction annotation in UD

We propose a new layer for selectively annotating constructions on top of UD trees. This is intended for constructions (in the sense of Construction Grammar) whose form and meaning/function is not already captured well by the UD tree. Construction instances receive a type name (possibly from a construction resource) and may contain relations to construction elements. The elements of the construction are not constrained by the UD tree: e.g., a construction element may cut across multiple UD subtrees. For now, we envision that they would be marked in the MISC column of .conllu files, though in principle they could be moved to a separate extension column.

The annotation layer does not have the goal of directly indicating the elements of form or meaning that are characteristic of or required by the construction, beyond indicating the construction evoker and spans of construction elements. Aspects of the UD analysis (tags, deprels, morphological features) that are characteristic of a construction's form should be described as such in a type-level construction entry. The precise contents of such an entry are not part of this proposal, but constructions incorporating UD information in some way already exist (e.g., the Russian Constructicon).

Full

Showing three overlapping constructions for completeness:

```

1 Sam    CxnEltOf=5:predicative-age.Individual,5:property-predication.Subj
2 is     CxnEltOf=property-predication.Cop
3 three  CxnEltOf=4:num-mod.Quantity,5:predicative-age.Value
4 years  Cxn=num-mod|CxnEltOf=4:num-mod.Counted,5:predicative-age.Units
5 old    Cxn=predicative-age,property-predication|CxnEltOf=5:property-predication.Pred

```

This effectively encodes construction-element relationships as dependencies (*offset:relation* notation echoes DEPS column), which would allow for straightforward graph querying. A common query might be to list the UD deprels associated with a construction element.

Note that i) a word may evoke multiple constructions, ii) a word may be both the evoker and an element of an evoked construction, iii) a word may participate in multiple elements of the same evoked construction.

Comma-separated lists should be sorted primarily by head node (where present), secondarily by construction name, thirdly by construction element name.

Full-consolidated

```

1 Sam    _
2 is     _
3 three  _
4 years  Cxn=num-mod(3:Quantity,4:Counted)
5 old    Cxn=predicative-age(1:Individual,3:Value,4:Units),\
        property-predication(1:Subj,2:Cop,3-5:Pred)

```

This is equivalent to the Full representation but consolidates all parts of an evoked construction on one line. It might be suitable for human annotation, to be automatically expanded to the Full representation with a script.

Comma-separated construction elements should be listed in node sort order. Constructions should be sorted alphabetically by name.

Simple

A partial representation may be useful in certain stages of an annotation workflow, e.g. before the full description of the construction is known, or before applying semiautomatic methods to identify construction elements.

The Simple notation includes the name of a construction, omitting any construction elements. A span may optionally be included for rendering purposes, but this span does not necessarily have any theoretical status.

```
1 Sam _
2 is _
3 three _
4 years Cxn=3-4:num-mod
5 old Cxn=1-5:predicative-age,property-predication
```

Exclusions

When manually reviewing forms that are candidate matches of a construction, it may be helpful to indicate that one of them is a non-match (a false positive). This can be done with the ExcludeCxn feature:

```
1 Sam _
2 is _
3 three _
4 years Cxn=3-4:num-mod
5 old Cxn=1-5:predicative-age,property-predication|ExcludeCxn=object-predication
```

Though we suggest the name ExcludeCxn in this standard, it should be regarded as a tool for development. Ideally, a corpus will be systematically reviewed for candidates of a construction, and excluded candidates discarded in the final version of the data.

Linking to a constructicon

If a constructicon resource exists, it should be declared in a metadata line in the file, and names of constructions from the resource should be prefixed with a namespace.

TBD issues

- Where are spans vs. heads used? Is a construction-evoking element allowed to be a span? Allow discontinuous spans (and change existing commas to semicolons)?
- A status field to indicate auto rather than gold matches?
- Allow question marks to indicate uncertainty during development?

Example annotations

(1) “The more you post the more money you make”. Let’s assume that the first comparative word (“more”) is the head. Then that word will be annotated like this in the simple format (the span 1–8 is optional):

```

1 the _
2 more Cxn=1-8:comparative-correlative
3 money _
...

```

And like this in the Full notation:

```

1 the _
2 more Cxn=1-8:comparative-correlative(1-4:Condition,6-10:Result,\
    2:ConditionDegree,7:ResultDegree)
3 money _
...

```

(2) “Sam is so glad that you are here that he baked a cake”. Advanced example (consolidated notation), showing two candidate matches of the same construction type on the same construction evoker, one of which is correct and one of which is incorrect (indicated by an excluded span):

```

...
3 so _
4 glad Cxn=causal-excess(1:Predicand,3:Degree,3-8:Cause,9-13:Result)\
    |ExcludeCxn=causal-excess(5-8:Result)
5 that _
...

```

Or in the simple form:

```

...
3 so _
4 glad Cxn=1-13:causal-excess|ExcludeCxn=1-8:causal-excess
5 that _
...

```

References

- 1 William Croft. 2022. *Morphosyntax: Constructions of the World’s Languages*. Cambridge: Cambridge University Press.

4.8 WG8: To Semantics and Beyond

Archna Bhatia (*Florida IHMC – Ocala, US*) and Kilian Evang (*Universität Düsseldorf, DE*)

License © Creative Commons BY 4.0 International license
© Archna Bhatia and Kilian Evang

Joint work of Timothy Baldwin, Emily B. Bender, Archna Bhatia, Francis Bond, Kilian Evang, Jan Hajič, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue, Amir Zeldes

Introduction

Somewhat continuing the discussion started in WG3, Working Group 8 (WG8) *To Semantics and Beyond* at the Dagstuhl Seminar 23191 on *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics* identified topics related to semantics that have presented to be problematic from various angles, such as semantic representations, inconsistencies or differences in representations/handling of notions across frameworks, cross-lingual handling of constructions or semantic notions, and coverage, e.g., both for arguments and modifiers. Specifically, the following four topics were identified, and we focused on the first two topics during the WG8 discussion sessions:

1. Serial verb constructions and light verb constructions from a cross-lingual perspective in terms of representation of their semantics, validity of statements such as “Predicates describe one event”, representations across different frameworks, leveraging other annotations or resources (e.g., the notion of scope tree in MRS), The topic was also discussed in WG3.
2. Inventory of semantic relations to describe states, processes and events cross-linguistically and for both arguments and adjuncts (also discussed in WG3 and WG5)
3. Inconsistent handling of entities and coreference across resources
4. Aspects of lexical semantics such as linking with wordnets, internal semantic structure, relation between internal semantic structure and compositionality, representation in existing resources/corpora

In the next two sections, we summarize the discussions corresponding to the two focused topics.¹¹

The Semantics of Serial Verb Constructions and Light Verb Constructions

Verbal constructions can present interesting phenomena involving a wide range of semantic considerations, e.g., in terms of determining whether a single event is involved or multiple events are involved, while at the same time each of these constructions themselves may also present a broad continuum to make it harder to demarcate the boundaries of the construction. This can present issues for describing such constructions or developing proper theoretical accounts of them. We discussed two verbal constructions, the serial verb constructions (SVCs) and light verb constructions (LVCs), that illustrate this issue.

SVCs, e.g., *persuade X to take a hike*, or *try to run*, are argued to involve control structures where one of the verbs is considered to be incorporated into the other verb. In such constructions, multiple independent lexical verbs are combined to indicate a single event (or two sub-events of a single event connected temporally indicating either simultaneous or consecutive temporality). LVCs, e.g., *take a bath*, *give a bath*, or *give a speech*, involve a verb which does not indicate the event itself but provides some aspectual or causal information about the event and the semantic core of the event is expressed by the nominal element.¹²

These constructions raise important questions about how one can determine semantic uniqueness of an event, whether the event is expressed by a verb, and also what an event is. We discussed tests for semantic uniqueness of an event and noted that neither of these tests were sufficient by themselves in indicating whether a single event was expressed or multiple events were expressed, but they might indicate tendencies that multiple tests together could help confirm to determine if a single event was involved. These included tests such as modification, argument sharing, and coordination.

In regards to modification, we identified examples in Thai SVCs indicating that negation (as modification) could be used but either in only one place (i.e., with one of the verbs) or it may mean the same irrespective of the position of negation. Cross-lingual examples such as these and the English, *I persuaded Francis not to take a hike* are interesting. Note that the use of negation in this position could be used to indicate persuading Francis to not take a particular hike or any hike, but in this position, it does not negate the act

¹¹ The document with more detailed notes for the discussions can be found at: <https://docs.google.com/document/d/13-J0kaCKAshDShRFE9NN81Yysc7jmLMlswmrfneRRo>

¹² In some languages, e.g., in Hindi, as the Hindi PARSEME annotated data indicate, the event may also be determined based on a combination of other elements such as an adjective with a verb.

of persuading Francis to take a particular/any hike. Quantification (as a modification strategy) can also be useful. In terms of semantic representation, “one event” may be represented semantically through, for example, the same position in the (Minimal Recursion Semantics/MRS representation) quantifier scope tree which can also entail a test of semantic uniqueness. However, quantifiers based test also does not always work. It is not always clear how to create the quantifier examples.

The argument sharing test is also not diagnostic. A vast amount of literature involving language specific approaches to SVCs discusses argument sharing as a criterion for a single event. While it may be a necessary criterion for some languages, it is not sufficient in determining semantic uniqueness of event. Also, while English SVC *I tried to run* involves subject sharing, a Thai construction equivalent to “*pound X flat*”, does not involve “subject sharing”¹³ but such resultative constructions in Thai are also considered to be SVCs. To take another example, the argument sharing test does not help resolve whether the Japanese construction equivalent to “*jump-rise the stairs*” involves one event with manner information or two events. In such verb-verb constructions, one verb (“jump”) is a hyponym of the other (“rise”).

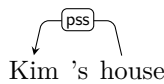
Coordination can also be considered as a test for semantic uniqueness but is again found to not be a strong diagnostic. For example, in the case of LVCs, since the N+V combination expresses an event, one can expect coordination of nouns in such constructions to not be possible. But PARSEME corpus shows there are quite a few of such cases. For example, it is possible to say *He took a bath and a shower*. Coordination also seems possible where the nominal part of the LVC is modified and is coordinated with a non-LVC noun, e.g., *the bath that I took and the bathtub were wonderful*.

The group found such examples with LVCs involving these diagnostics, and their flexibility when testing their boundaries particularly interesting and ended up focusing a large part of the WG8 sessions on discussing LVCs further. We are continuing our discussions involving LVCs, their characteristics and behavior across languages to arrive at a more general and typologically informed semantic representation of these constructions beyond Dagstuhl with many participants from WG8 and other groups (the current participants’ list continuing these discussions includes: Emily M. Bender, Archana Bhatia, Kilian Evang, Dan Flickinger, Jan Hajič, Dag Haug, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue). We plan to continue these discussions to also include other MWEs to develop their semantic representations while taking into account the observed cross-lingual patterns as well as the idiosyncratic behaviors they demonstrate.

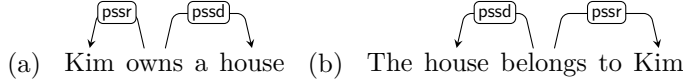
Towards a Unified Inventory of Semantic Dependency Labels

There is currently no commonly accepted inventory of semantic roles and relations that can be applied across languages, domains – something like Universal Dependencies, with similarly low barriers to use in applications, but for semantic roles and relations. Kilian Evang started the discussion by formulating some desiderata for such a scheme: it should be usable without reference to a lexicon, it should cover at minimum all dependencies between content words (thus, both arguments and modifiers) with a unified vocabulary, and handle states and events in a unified way. Semantic relations should not be too fine-grained, He illustrated what such a scheme might look like using a possession/control relation, labeled *pss*. This label could be used directly for modifier relations:

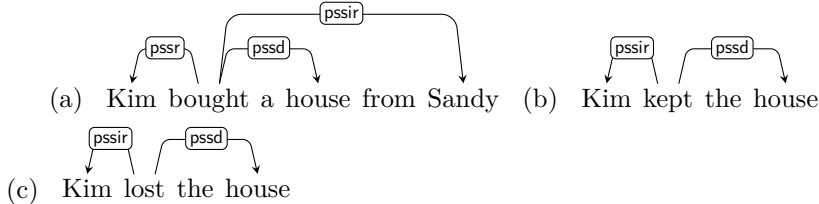
¹³ although an argument is shared



For arguments of predicates denoting states, the same relation is split over two dependency edges, dubbed the “domain” and the “range” of the (possession) relation:



For arguments of predicates denoting events, there is the additional possibility of having an argument denoting an *initial* range (ir):



Differences to existing schemes were discussed: MRS first assigns semantically unspecified argument labels (ARG0 = event, then ARG1, 2, 3 in order of decreasing obliqueness). Separate, language-specific lexicons à la FrameNet assign specific roles to these argument labels, per predicate (e.g., *own*, *belong* may be in the same frame, but for *belong* ARG2 is the owner, and for *own*, ARG1). In PropBank, all arguments are verb-specific. SynSemClass is an event type ontology under development that defines coarser, but still hierarchically ordered classes of events (and states), with e.g. the “ownership” class containing verbs like *own*, *possess*, *hold*. “Buying” would be a separate class, e.g., it is not clear that the seller is also the initial owner in all cases. This links to the general question of how much annotators should be allowed to infer that is not strictly entailed by the predicate, e.g. *I visited my mother at 4 o'clock yesterday* – should the annotator include the information that it was in the afternoon? Verbs in the Prague Dependency Treebank are linked to FrameNet, VerbNet, OntoNotes, PropBank, English WordNet, and SynSemClass. In the TRIPS ontology, there are events of change, events of state, and within those further classes.

Potential issues and next steps with the proposed schemes were discussed. Compared to UD, additional semantic edges need to be annotated compared to UD, e.g., in control constructions. It was therefore suggested that annotation start atop Enhanced UD, even if automatically predicted and imperfect. Similarly, existing PARSEME data could be used to automatically pre-annotate semantic edges in VMWEs, e.g. *kick* → *bucket* with special labels. The next steps should focus on annotating a bunch of data to evaluate the proposal, and also on lexical mapping to other resources like SynSemClass, PropBank, Universal Proposition Banks, VerbAtlas, VerbNet, FrameNet, SNACS, UMR to study synergies and enable comparisons.

4.9 WG9: Fostering Corpus-based Typology [“Big Hairy Problems”]

Laura Kallmeyer (*Universität Düsseldorf, DE*) and Christopher Manning (*Stanford University, US*)

License © Creative Commons BY 4.0 International license

© Laura Kallmeyer and Christopher Manning

Joint work of Laura Kallmeyer, Christopher Manning, Joakim Nivre, Reut Tsarfaty, Gosse Bouma, Agata Savary, Dag Haug

We started with two questions:

1. How can we make corpus-based typology easier, in particular with respect to investigating interesting phenomena beyond word order (also discussed in WG5)?
2. How can we bring some of the constituency information typologists might want to encode into an easy-to-use annotation format that connects UD and constituency (also discussed in WG4, WG5, WG7)?

First, we looked a little into RRG [Role & Reference Grammar] and its distinctions of nucleus, core, and clause for different levels of tightness of binding of elements and whether that can and should be captured in UD. We could represent this by adding something to the dependency label.

Then we dealt with polysynthetic languages with noun incorporation.

Question discussed after the break: Do we want to have constituency representations in UD or can we make do with phrase-level features?

Joakim says things to consider:

- Phrase-level features: We may need them.
- Do we want segmentation in the basic representation for incorporating or polysynthetic languages? Do we need complex feature structures, i.e., where features can take a content word as a feature value.
- One thing maybe missing is when there are multiple realizations of the same argument: *moi je pense que*: There are sort of 3 realizations of 1SG.

Chris:

- UG really has no treatment of complex predicates/light verbs/serial verbs/monoclausality.

Reut:

- Using a featural analysis we could represent something like a periphrastic passive the same as a single word morphological passive.

Gosse:

- Do we need more bracketing, starting to head in the direction of constituency?

Joakim:

- How can we do this with a manageable amount of complexity?

Agata:

- Field linguist thinks UD isn't linguistic enough for him. How? Is SUD better?

Chris:

- We don't need to provide a linguistic theory, if we're just providing enough that people can search for examples of interest, then we've won.

Reut:

- What's more important: Ease of annotation or ease of searching?

Note that the flat relation does not imply a head. We have to educate people that there are links in UD because everything is made into a tree, but not all of the arrows are head dependent relations.

Clause-level phrasal features are a good way to capture interesting types of constructions and simultaneously capture things that are above the level of a single token.

We compiled a list of phrasal features that one might want to have:

- Clause level features:
- Clause types:
 - Interrogative
 - * Presuppose: yes, no
 - * Wh-question: yes, no
 - * Tag-question: yes, no
 - Declarative
 - Imperative
- Voice/Valency changing constructions:
 - Active
 - Passive
 - Antipassive
 - Middle voice
 - Causative
 - Applicative
- Polarity:
 - Positive
 - Negative
- Information structure:
 - Cleft?
 - Pseudocleft?
 - Extraposition?
- Tense/Aspect/Mood
- Evidentiality
- Long-distance dependency?
- Nominal phrase level features:
- Nominal types:
 - Definiteness
 - Case
 - Construct state
 - Agreement
 - Deverbal noun

We then came back to the Nahuatl example. Our proposal: In the enhanced dependencies, have abstract nodes for those arguments that follow from the morphologically encoded information. Besides that, encode all other properties in the features and project these also to the clause level (on the same node).

Dag:

- We need to add in the edeps nodes for subj, obj shown by agreement (but pro-drop) or else we can't show control relationships that include them in the enhanced dependencies. UD is not available in tools used in fieldwork.
- Is UD a theoretical framework or a pre-theoretical annotation framework?

- But it does make some theoretical choices like being content-word head
We should test this with some concrete constructions!
- Serial verb constructions?

Friday: Phrase-level features

- Work more on the feature list: There seems to be a need for phrase-level features and we have a first cut at a list of them.
- Put a survey together on how to change things: Need to get more buy-in from a big community and understand more about more language families.
- Easiest place to put them immediately is in the FEATS column and just have new names.
- We are wanting to capture periphrastic tense and aspect, not just morphological form of individual verbs
- We also have a mechanism of layered features for agreement with multiple things: Hebrew: **אהבה** Gender[subj]=neut, number[subj]=sg, person[subj]=1, gender[obj]=fem, number[obj]=sg, person[obj]=3rd

Let's extend this to phrase level features: Definite[phrase]=Yes

- If we used something like Conx[Phrasal]=Periphrastic or Conx[Phrasal]=Light then we could capture this.

UD Governance

Can we have open zoom discussions of changes not just dictates in emails? [“This is an information session rather than an active discussion” – manage it.] There are going to be UD and PARSEME introductions in UniDive but they're not really community discussions. Can we go back to having real UD workshops where things are discussed and worked out? Perhaps together with the next combined Coling/LREC. For different language groupings, whether language families or subgroups such as user-generated content or ancient languages, can we explicitly have more organization and subcommunity organizers? Would that help? We decided to break early so that we could get our coffee ahead of time and be on time to the final session.

5 Open Problems

5.1 Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus

Alexandre Rademaker (IBM Research – Sao Paulo, BR), Francis Bond (Palacký University Olomouc, CZ), and Daniel Flickinger (North Newton, US)

License © Creative Commons BY 4.0 International license

© Alexandre Rademaker, Francis Bond, and Daniel Flickinger

Main reference Alexandre Rademaker, Abhishek Basu, Rajkiran Veluri: “Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus”, in Proc. of the Global Wordnet Conference, 2023.

In 2008, the Princeton team released the last version of the “Princeton Annotated Gloss Corpus”. In this corpus, the word forms from the definitions and examples (glosses) of Princeton WordNet are manually linked to the context-appropriate sense in WordNet. However, the annotation was incomplete, and the dataset was never officially released as part of WordNet 3.0, remaining as one of the standoff files available for download. Eleven years later, in 2019, one of the authors of this abstract restarted the project aiming to complete the sense annotation of the approximately 200 thousand word forms not yet annotated. Intending to provide an extra level of consistency in the sense annotation and a deep semantic representation of the definitions and examples promoting WordNet from a lexical resource to a lightweight ontology, we now employ the English Resource Grammar (ERG), a broad-coverage HPSG grammar of English to parse the sentences and project the sense annotations from the surface words to the ERG predicates.

The disambiguation of words in the glosses can also improve WordNet and provide completeness and consistency. For instance, the initial versions of WordNet do not contain relations that indicate how words like “racquet”, “ball”, and “net”, and the concepts behind them, are part of another concept that can be expressed by “court game” [12]. In WordNet 3.0 the “domain relations” between synsets were introduced to alleviate this so-called “tennis problem” of WordNet [21], but the disambiguated gloss of the synset *tennis, lawn_tennis* would already enrich the connections among the concepts. Another desired property is that all words used in the definitions are defined in this same resource. Hopefully, this completeness could also help us ensure quality in our long-term endeavor during the expansion of WordNet to highly technical domains. Once more concepts are added or redefined, the glosses would be refined and disambiguated, forcing us to use the newly added senses in a productive cycle of editing, testing, and correcting.

Regarding the multiword expressions such as “military formation”, “geological formation”, “reticular formation”, and “reaction formation”. The expression “military formation” stands out in many glosses. The expression exists as a MWE but a similar expression, “naval formation” does not, with both appearing in the gloss “the side of military or naval formation”. We discussed whether “naval formation” should be considered a MWE or whether “military formation” should not be considered one.

A familiarity with a particular domain also plays a role in the annotation process, affecting both the senses assigned and the decisions regarding which collocations should be considered MWEs. For instance, the expression “rock formation” is not part of PWN, but it appears many times in the corpus.

1. a national park in Utah having colorful *rock formations* and desert plants and wildlife (08603525-n)
2. the gradual movement and formation of continents (11434448-n)

Although some of us believe the expression should be added to PWN, it is not in the lexicon yet, and so, some annotators chose the sense “(geology) the geological features of the earth” for the word “formation” in all occurrences of the expression. This decision was understandable if we consider that the word “rock”, in one of its senses, naturally evokes the domain “geology”. The same can also be said for the word “continent” in Example 2. But one annotator, a geology expert, consistently took the sense “a particular spatial arrangement” for the word “formation” in this expression. His decision was based on the strict interpretation of “geological formation” as a domain-specific concept and reinforced by the fact that “geological formation” in PWN has “physical object” as its hyperonym, not “formation” (as a process).

Some cases of multiword expressions (MWE) seem to support our belief that sense annotation and PWN maintenance should be joint work. First, we need to define and enforce heuristics to determine when a given word sequence is a multiword expression (being sense annotated as a single entity), and when its component tokens should be annotated individually. The compositionality and conventionality criteria from [11] may help, however these criteria are not as clear-cut as we would like them to be. Take the case of “first degree” and the example “all of the terms in a linear equation are of the *first degree*” in its definition (synset 05861716-n); we can annotate it as “first degree” (this same sense being defined in the synset where the example is given); but there is no sense for “second degree”, or “third degree”, which are equally valid. This leads us to consider that it should be annotated individually, and that the “first degree” sense should be removed from PWN.

Another possible approach we are considering is to follow the criteria adopted in [31] and its implementation in the ERG grammar. The alignment of the sense annotation (in the tokens) with the predicates (from the MRS) reveals a lot of cases to consider going far beyond the tokenization mismatches. For some idiomatic expressions we can really on ERG lexicon information (e.g. “the one where digestion **takes place**”). We also have to consider the coordinations of expressions sharing tokens. In the definition “a semisolid mass of coagulated **red** and **white blood cells**”, our sense annotation explicit reuse the tokens “blood cells” and annotate two senses “red blood cell” and “white blood cell”. In the MRS representation, it is yet not clear in which predicate (or set of predicates) to attach the sense identifiers. The easier cases are the adjective-noun constructions, sharing handlers in the MRS representation (e.g. “big toe”), this is not too different from the case of multiple adjectives modifying the same noun (e.g. “uric acid”). But what about the noun-noun compounds? ERG has a special abstract predicate called “compound” to represent the underspecified relation between two nouns. For instance, in “a **blood disease** characterized by an abnormal multiplication of macrophages“, the **blood** noun is related to **disease** with this abstract predicate (consider it an underspecified preposition connecting this two nouns). The sense tagging annotate the two words as a glob associated to the sense “blooded disease” but not all compounds may represent a single sense (examples above). Verbal phrases such as “**coughing up** blood from the respiratory tract” is another frequent case where tokenization differ. In the manual sense annotation, the two tokens (“coughing” and “up”) forms a glob annotated with one sense, the ERG analyses produces a single predicate.

All of the cases we highlighted above are being considering in the on going work of expand the ERG lexicon with all WordNet lexical entries, an exploratory work that aims to support a more semantic driven parsing selection and raking model.

References

- 1 Eneko Agirre, Oier Lopez De Lacalle, Aitor Soroa, and Informatika Fakultatea. 2009. Knowledge-based wsd and specific domains: Performing better than generic supervised wsd. In *IJCAI*, pages 1501–1506.
- 2 Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd.
- 3 Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.
- 4 Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg. Springer Berlin Heidelberg.
- 5 Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile, Pasquale Lops, and Giovanni Semeraro. 2007. Uniba: Jigsaw algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic. Association for Computational Linguistics.
- 6 Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.
- 7 Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008a. Using and extending wordnet to support question-answering. In *Proceedings of the 4th Global Wordnet Conference*, pages 111–119, Hungary.
- 8 Peter Clark, Christiane Fellbaum, Jerry R Hobbs, Phil Harrison, William R Murray, and John Thompson. 2008b. Augmenting wordnet for deep understanding of text. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 45–57.
- 9 Ann Copestake. 2002. *Implementing typed feature structure grammars*, volume 110. CSLI publications Stanford.
- 10 Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.
- 11 Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.
- 12 Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- 13 Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1), pages 15–28.
- 14 Dan Flickinger. 2011. Accuracy v. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.
- 15 Dan Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank. a dynamically annotated treebank of the wall street journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.
- 16 Michael Wayne Goodman. 2019. A python library for deep linguistic resources. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–7. IEEE.
- 17 Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2: a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*, pages 1–8.

- 18 Michael C McCord. 2004. Word sense disambiguation in a slot grammar framework. Technical Report RC23397, IBM.
- 19 John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).
- 20 Rada Mihalcea and Dan I. Moldovan. 2001. extended wordnet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.
- 21 George A Miller. 1993. The association of ideas. *The General Psychologist*, 29:69–74.
- 22 George A Miller and Christiane Fellbaum. 2007. WordNet then and now. *Language Resources and Evaluation*, 41(2):209–214.
- 23 George A Miller, Claudia Leacock, Randee Teng, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.
- 24 Dan Moldovan and Adrian Novischi. 2004. Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18(3):301–317.
- 25 Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Ike*, pages 412–416.
- 26 Stephan Oepen. 2001. [incr tsdb()] – competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany. In preparation.
- 27 Stephan Oepen, Kristina Toutanova, Stuart M Shieber, Christopher D Manning, Dan Flickinger, and Thorsten Brants. 2002. The lingo redwoods treebank: Motivation and preliminary applications. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.
- 28 Woodley Packard. 2015. *Full forest treebanking*. Ph.D. thesis, University of Washington.
- 29 Adam Pease and Andrew Cheung. 2018. Toward a semantic concordancer. In *Proceedings of the 9th Global Wordnet Conference*, pages 97–104, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.
- 30 Alexandre Rademaker, Bruno Cuconato, Alessandra Cid, Alexandre Tesseracto, and Henrique Andrade. 2019. Completing the Princeton annotated gloss corpus project. In *Proceedings of the 10th Global Wordnet Conference*, pages 378–386, Wrocław, Poland. Global Wordnet Association.
- 31 Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.
- 32 Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

5.2 NLP-based Study of Universals of Linguistic Idiosyncrasy

Agata Savary (University Paris-Saclay, CNRS – Orsay, FR)

License  Creative Commons BY 4.0 International license
 © Agata Savary

According to William Croft, the universals of linguistic idiosyncrasy established so far include the following:¹⁴

1. Meaning becomes idiosyncratic before morphosyntactic form does (which is why most MWEs are syntactically mostly regular)
2. Idiosyncrasy leading to lexicalization is most likely to develop in:
 - a. (typifying) modifier-referent constructions (*women’s magazine*, *table leg*)
 - b. verb-object constructions (*to pull one’s legs*, *to pay a visit*)
3. Idiosyncrasy leading to grammaticalization is most common in:
 - a. adposition path (*in light of N*)
 - b. tense-aspect-modality-polarity path (*be go-ing to V*)
 - c. discourse marker → sentence connective path (*all the same*, *oh by the way*).

The PARSEME framework, with its cross-linguistically unified guidelines for verbal multiword expressions (VMWEs), and its corpus annotated for VMWEs in 26 languages, could help corroborate these hypotheses. In particular, it could provide some evidence about the frequency of lexicalization in verb-object constructions (2b above).

Namely, in the latest 1.3 release of the corpus [5], the global statistics for all 26 languages show that among the 127,500 annotated VMWEs:¹⁵

- 44,171 are labeled as light verb constructions (LVCs)
- 29,062 are inherently reflexive verbs (IRVS)
- 26,214 are verbal idioms (VIDs)

These 3 categories contain large percentages of verb-object combinations.

More precisely, according to the PARSEME guidelines:

- LVCs are combinations of semantically light verbs and predicative nouns expressing the semantics of the action or state. Two subcategories are defined. In LVC.full the verb’s subject is the noun’s semantic argument as in (sl) *imeti predavanje* ‘give a lecture’. In LVC.cause the verb’s subject is the cause or source of the noun, as in (en) *to grant right*. Most LVCs in their so-called *canonical forms* (the least syntactically marked syntactic variants which preserve the idiomatic reading), consist of verbs heading direct objects.¹⁶
- IRVs are combinations of a verb *V* and a reflexive clitic *R* such that (i) *V* never occurs without *R*, as in (sv) *gifta sig* (lit. *get-married oneself*) ‘get married’, or (ii) *R* distinctly changes the meaning or valency of *V*, as in (es) *recogerse* (lit. *to gather oneself*) ‘to go home’.
- VIDs gather cases not covered by other categories. The verb’s dependents are unrestricted, including subjects, as in (en) *a little bird told me*, direct objects, as in (ro) *a întoarce foaia* (lit. *to turn the sheet*) ‘to become harsher’, etc. The verb can have several

¹⁴ <https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023/-/wikis/Universals-of-linguistic-idiosyncrasy-established-so-far>

¹⁵ <https://parseme.grew.fr/tables/?data=parseme/labels@1.3>

¹⁶ They can also be verbs with oblique complements, whether introduced by a preposition, as in (pl) *występować w obronie uciekinierów* (lit. *to stand out in the defense of refugees*) ‘to defend refugees’, or by a non-accusative case, as in (pl) *obdarzać kogoś zaufaniem* ‘to endow sb trust.INS’. These verb-oblique combinations are more rare than verb-object combinations.

dependents, as in (en) *cut a long story short*, or combine features from other VMWE categories, as in (sv) *sätta sig upp mot någon* (lit. *sit oneself up against someone*) ‘defy someone’.

The precise frequency and distribution of verb-object combinations in VMWEs in various languages can be estimated e.g. with the Grew-match corpus browser [2],¹⁷ however, care must be taken with the design of the queries. For instance query (1), which seems to straightforwardly correspond to hypothesis 2b, results in only 1,296 matches out of all 7,313 VMWEs annotated in Polish,¹⁸ i.e. only 17%.

```
(1) pattern {
      MWE [label]; %A MWE is a new node with a "label" feature
      MWE -> V;    %The MWE has at least two nodes, marked V
      MWE -> O;    %...and O
      V[upos=VERB]; %The universal POS of V is VERB
      V -[obj]-> O; %The dependency between V and O is obj
    }
```

At first sight, this seems to invalidate hypothesis 2b. However, a finer study of encoding and variants of VMWEs provides more insight into verb-object combinations they contain.

Firstly, the syntactic dependencies underlying the PARSEME VMWE annotations rely on the Universal Dependencies standards, corpora and tools [1, 3, 6]. In UD IRVs would often contain the *expl* relation (rather than *obj*) to indicate that the reflexive pronoun cannot be mapped on any semantic argument of the verb, even if, strictly syntactically speaking, the reflexive clitic is most often truly a direct object. On the other hand, all IRVs in Polish cannot be considered verb-object combinations since in some of them the reflexive is in dative or instrumental case, indicating an oblique rather than a direct object, as in (pl) *wyobrażać sobie* (lit. *to imagine oneself.dat*) ‘to imagine’. Therefore, a more precise query finding IRVs with the reflexive clitic really playing the role of a direct object might be as in (2).

```
(2) pattern {
      MWE [label="IRV"]; %the MWE is an IRV
      MWE -> R;          %It has a node...
      R[form="się"];     %...whose surface form is "się"
    }
```

Secondly, VMWEs frequently occur in syntactic variants which violate the prototypical verb-object structure with the verb dominating the noun via the *obj* dependency. Namely, in Polish, and likely in many other Slavic languages, objects often require *structural case*, i.e. their case depends on the presence of negation and on the form of the head [4]. When the verb is not negated or nominalized, as in (pl) *bić pianę* (lit. *to whip foam.acc*) ‘to speak a lot without adding much to the discussion’, the object is in accusative case. Otherwise it is in genitive, as in (pl) *nie bić piany* (lit. *not to whip foam.gen*) and (pl) *bicie piany* (lit. *whipping foam.gen*). The object also takes genitive case when preceded by some quantifiers, as in (pl) *bić dużo piany* (lit. *to whip a lot of foam.acc*). These case fluctuations may provoke the assignment of the *obl* rather than *obj* dependency between the verb and the noun, notably when the syntactic data in the PARSEME corpus stem from automatic parsing.

¹⁷<https://parseme.grew.fr/>

¹⁸<https://parseme.grew.fr/?corpus=PARSEME-PL@1.3>

Note also that when a verb in a VMWE is nominalised, it can be tagged with the NOUN part-of-speech rather than VERB. Other syntactic variants e.g. extractions, as in (pl) *kara, którą wymierzili* ‘the penalty which they imposed’ invert the direction of the dependency and change its label (here `acl:recl` from the noun *kara* ‘penalty’ to the verb *wymierzili* ‘imposed’). These subtleties suggest that query (1) should be more relaxed, as in (3), and even then it might not have enough coverage.

```

pattern {
  MWE [label="LVC.full"|"LVC.cause"|"VID"];
(3)  MWE -> V;
     MWE -> O;
     V -[obj]-> O;
}

```

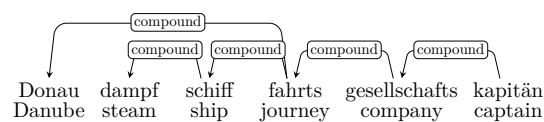
When these finer queries (2) and (3) are run on the Polish PARSEME corpus they indicate that VMWEs with direct objects are contained in at least:

- 1564 out of 3625 LVC and VID occurrences (43%)
- 3642 out of 3688 IRV occurrences (99%)
- 5206 out of all 7313 annotated VMWEs occurrences, i.e. **71%**

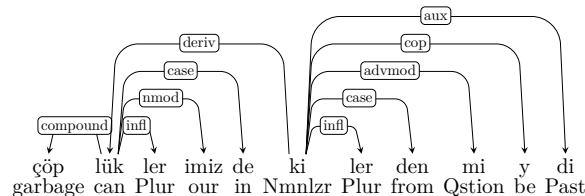
This is a much more encouraging quantification of idiosyncrasy in verb-object pairs, in relation to hypothesis (2b). Of course for this statistic tendency to be considered universal, similar sets of queries should be designed in order to appropriately cover the verb-object occurrences in other languages. It is also worth noting that hypothesis 2b and the others from the beginning of this section do not mention if the likelihood of particular types of idiosyncrasies is suggested with respect to occurrences or types (unique constructions). Here, we only dealt with the former but studying the latter is equally interesting.

References

- 1 Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, **47**(2), pages 255–308.
- 2 Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 168–175, Online. Association for Computational Linguistics.
- 3 Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- 4 Agnieszka Patejuk, Adam Przepiórkowski. 2014. Structural case assignment to objects in Polish. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'14 Conference*, pages 429–447, Stanford, CA. CSLI Publications.
- 5 Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.
- 6 Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.



■ **Figure 1** *Donaudampfschiffahrtsgesellschaftskapitän* “Danube steamship company captain”.



■ **Figure 2** *çöplüklerimizdekilerdenmiydi* “was it from those that were in our garbage cans?”.

5.3 Subword Relations, Superword Features

Daniel Zeman (Charles University – Prague, CZ)

License © Creative Commons BY 4.0 International license
© Daniel Zeman

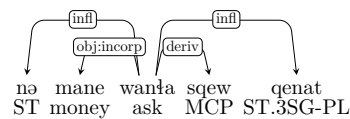
Introduction

Universal Dependencies (UD) subscribes to the lexicalist principle, claiming that dependency relations connect *words*, while the process of constructing words by combining smaller units (*morphemes*) is substantially different. Consequently, word-internal structure is normally not shown in UD.¹⁹ Nevertheless, there seems to be some demand [1] for a UD extension that would allow for showing word-internal structure in a way similar to how inter-word relations are represented. Here are some motivational examples:

- German compounds are written as one word and represented by one tree node in UD. English compounds may be perfectly parallel to the German ones, yet they are typically written as multiple orthographic words. In UD, they are multiple nodes connected by **compound** relations. The parallel structure is not visible in German UD but it could if the compounds were split into multiple tree nodes (Fig. 1).²⁰ Moreover, other annotation may pertain just to one part of a compound: We may want to annotate the MWE *Rolle spielen* “to play a role” in the compound *Hauptrolle spielen* “to play the main role”.
- Turkish words may combine several derivational and inflectional steps. Traditional analysis would break them up to *inflection groups* but in UD they are mostly kept together and the internal structure is not visible (unlike Fig. 2). See also [4].
- Chukchi transitive verbs may incorporate their objects and switch to intransitive inflection (Fig. 3) [5].
- Fieldworkers may prefer morpheme-based analysis when documenting a language; a UD example is the treebank of Beja [3].

¹⁹ Except for the optional **MSeg** and **MGloss** attributes in the MISC column of some treebanks, which can at least hint at the morphemic composition of a word.

²⁰ In fact, compounds are a gray zone. While most UD languages do not split them, they are split in Sanskrit UD, as such analysis is traditional in Sanskrit linguistics.



■ **Figure 3** *nəmanewan¹asqewqena* “they constantly asked for money”.

```
# text = Er spielt die Hauptrolle im Haus.
# text_en = He plays the main role in the house.
1 Er er PRON _ Case=Nom|PronType=Prs 2 nsubj _ _
2 spielt spielen VERB _ Mood=Ind|VerbForm=Fin 0 root _ _
3 die der DET _ Case=Acc|PronType=Art 4 det _ _
4-6 Hauptrolle _ _ _ _ _ _
4 Hauptrolle Hauptrolle NOUN _ Case=Acc|Number=Sing 2 obj _ _
5 haupt haupt ADJ _ Degree=Pos 6 amod _ _
6 Rolle Rolle NOUN _ Case=Acc|Number=Sing 4 wroot _ _
7-8 im _ _ _ _ _ _
7 in in ADP _ _ 9 case _ _
8 dem der DET _ Case=Dat|PronType=Art 9 det _ _
9 Haus Haus NOUN _ Case=Dat|Number=Sing 2 obl _ SpaceAfter=No
10 . . PUNCT _ _ 2 punct _ _
```

■ **Figure 4** CoNLL-U with subword relations.

Precisely defining a *word* (even a *syntactic word*) cross-linguistically is a difficult task [2]. However, it matters less if we can annotate inter-word and intra-word relations in a similar manner. We propose to work within WG1 (and partially WG2) on an extension of UD that would support such annotation.

Subword Relations

As relations between subword units violate the lexicalist principle, they cannot be part of a regular UD treebank under the current guidelines; they have to be an extension that stands outside UD proper. Nevertheless, the file format should retain low-level compatibility with CoNLL-U so that existing tools can still process it. So, while new relation labels are conceivable, there should be no new line types beyond the existing 5 (comment, multiword token, node, empty node, empty line). There may be extra columns for readability (CoNLL-U Plus²¹) but it should be possible to collapse them into MISC attributes if needed.

Ideally, the format should accommodate normal UD treebank plus additional subword annotation and there should be a script that throws away the extra relations and extracts the regular UD treebank. If a word is decomposed, the relations between its parts should probably form a tree (\Rightarrow single root). The annotation of the root morpheme will differ from the annotation of the whole word, so we need nodes for both.²² Multiword token lines must be used to indicate the mapping of the nodes to surface tokens (Fig. 4).

²¹ <https://universaldependencies.org/ext-format.html>

²² As one of the reviewers noted, this has drawbacks, too. Parallelism between languages will be somewhat spoiled, as German *Hauptrolle* will now have three nodes, while English *main role* will have only two. Alternatively, the word-level morphological annotation could be stored for the morphemes spanning the word in a similar manner to what we propose for superword features in the next section.

Superword Features

Conversely, we may want to assign word-like annotation to a multiword expression. For example, a MWE functions like an adverb although its member words are not adverbs. Some treebanks already mark this with `MWEPOS=ADV` (or `ExtPos`) in MISC. Similarly, for German verbs with separable prefixes (e.g. *ein/steigen* “get on”), we may want to indicate the lemma that describes the two parts together. We may also want to add morphological features to sets of words, e.g., `Tense=Fut` for periphrastic future (composed of words that are not future themselves).

The MWE does not have to be linearly contiguous, so we cannot abuse multiword token lines for this purpose. MWEs tend to be catenas,²³ suggesting that the MISC column of the head node could hold such annotations. They are not complete subtrees though: in *I have come home*, the head of the periphrastic verb form *have come* is *come*, but we want to exclude the other dependents (*I* and *home*) from the annotation of the verbal features. We thus need a MISC attribute with the IDs of the nodes that are included in the MWE, e.g., `MWSpan=1-3,5`.

Multiple MWEs could have their annotation placed at the same head node, meaning that we have to use numeric ids to mark MISC attributes that pertain to the same MWE. For example, in *He has played the main role in the process*, we could annotate `MWSpan[1]=2-3 | MWLemma[1]=play | MWUPOS[1]=VERB | MWAspect[1]=Perf` and `MWSpan[2]=2-3,6 | MWLemma[2]=play role | MWUPOS[2]=VERB | MWAspect[2]=Perf`. Essentially, what we are looking at is a constituent-oriented analysis combined with dependencies, although ‘constituents’ in this sense are not linearly contiguous spans of words.

Acknowledgements

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation; and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7):89 – 138.
- 2 Martin Haspelmath. 2022 Draft. Defining the Word.
- 3 Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48 – 60, Sofia, Bulgaria. Association for Computational Linguistics.
- 4 Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the BOUN treebank reflecting the agglutinative nature of Turkish. *arXiv preprint*, arXiv:2207.11782.
- 5 Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195 – 204, Barcelona, Spain (Online). Association for Computational Linguistics.

²³ Even catena is probably not always granted. Grouping auxiliaries without the main verb would be a problem, although one may argue that this can be left for SUD to deal with. But coordination may complicate things. In *The food has been cooked and eaten*, one may want to combine the auxiliaries not only with *cooked* but also with *eaten*. Maybe we can say that this would be a catena in the enhanced dependency graph.

Participants

- Timothy Baldwin
MBZUAI – Abu Dhabi, AE
- Emily M. Bender
University of Washington –
Seattle, US
- Archana Bhatia
Florida IHMC – Ocala, US
- Nina Böbel
Universität Düsseldorf, DE
- Francis Bond
Palacký University Olomouc, CZ
- Gosse Bouma
University of Groningen, NL
- Jörg Bücker
Universität Düsseldorf, DE
- Mathieu Constant
ATILF – Nancy, FR
- Marie-Catherine de Marneffe
FNRS – UC Louvain, BE & Ohio
State University – Columbus, US
- Kilian Evang
Universität Düsseldorf, DE
- Daniel Flickinger
North Newton, US
- Omer Goldman
Bar-Ilan University –
Ramat Gan, IL
- Jan Hajic
Charles University – Prague, CZ
- Dag Haug
University of Oslo, NO
- Sylvain Kahane
University Paris Nanterre, FR
- Laura Kallmeyer
Universität Düsseldorf, DE
- Maria Koptjevskaja-Tamm
Stockholm University, SE
- Lori Levin
Carnegie Mellon University –
Pittsburgh, US
- Peter Ljunglöf
University of Gothenburg, SE
- Teresa Lynn
MBZUAI – Abu Dhabi, AE
- Christopher Manning
Stanford University, US
- Nurit Melnik
The Open University of Israel –
Raanana, IL
- Joakim Nivre
Uppsala University, SE
- Alexandre Rademaker
IBM Research – Sao Paulo, BR
- Carlos Ramisch
Aix-Marseille University, FR
- Manfred Sailer
Goethe-Universität Frankfurt am
Main, DE
- Agata Savary
University Paris-Saclay, CNRS –
Orsay, FR
- Nathan Schneider
Georgetown University –
Washington, DC, US
- Sara Stymne
Uppsala University, SE
- Reut Tsarfaty
Bar-Ilan University –
Ramat Gan, IL
- Francis M. Tyers
Indiana University –
Bloomington, US
- Ekaterina Vylomova
The University of Melbourne, AU
- Leonie Weissweiler
LMU München, DE
- Nianwen Xue
Brandeis University –
Waltham, US
- David Yarowsky
Johns Hopkins University –
Baltimore, US
- Amir Zeldes
Georgetown University –
Washington, DC, US
- Daniel Zeman
Charles University – Prague, CZ



Topological Data Analysis and Applications

Ulrich Bauer*¹, Vijay Natarajan*², and Bei Wang*³

1 TU München, DE. mail@ulrich-bauer.org

2 Indian Institute of Science – Bangalore, IN. vijayn@iisc.ac.in

3 University of Utah – Salt Lake City, US. beiwang@sci.utah.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23192 ‘Topological Data Analysis and Applications’. The seminar brought together researchers with backgrounds in mathematics, computer science, and different application domains with the aim of identifying and exploring emerging directions within computational topology for data analysis. This seminar was designed to be a followup event to two successful Dagstuhl Seminars (17292, July 2017; 19212, May 2019). The list of topics and participants were updated to reflect recent developments and to engage wider participation. Close interaction between the participants during the seminar accelerated the convergence between mathematical and computational thinking in the development of theories and scalable algorithms for data analysis, and the identification of different applications of topological analysis.

Seminar May 7–12, 2023 – <https://www.dagstuhl.de/23192>

2012 ACM Subject Classification Human-centered computing → Visualization; Information systems → Data analytics; Mathematics of computing → Algebraic topology; Theory of computation → Computational geometry

Keywords and phrases algorithms, applications, computational topology, topological data analysis, visualization

Digital Object Identifier 10.4230/DagRep.13.5.71

1 Executive Summary

Vijay Natarajan (Indian Institute of Science – Bangalore, IN)

Ulrich Bauer (TU München, DE)

Bei Wang (University of Utah – Salt Lake City, US)

License © Creative Commons BY 4.0 International license
© Vijay Natarajan, Ulrich Bauer, and Bei Wang

This Dagstuhl Seminar titled “Topology, Computation, and Data Analysis” brought together researchers in mathematics, computer science, and visualization to engage in active discussions on theoretical, computational, practical, and application aspects of topology for data analysis.

Context

Topology is considered one of the most prominent research fields in mathematics. It is concerned with the properties of a space that are preserved under continuous deformations and provides abstract representations of the space and functions defined on the space. The modern field of topological data analysis (TDA) plays an essential role in connecting mathematical theories to practice. It uses stable topological descriptors as summaries

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Topological Data Analysis and Applications, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 71–95

Editors: Ulrich Bauer, Vijay Natarajan, and Bei Wang



DAGSTUHL
REPORTS

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

of data, separating features from noise in a robust way. The seminar brought together researchers from mathematics, computer science, and application domains (e.g., materials science, neuroscience, and biology) to accelerate emerging research directions and inspire new ones in the field of TDA.

Goals

The Dagstuhl Seminars 17292 (July 2017) and 19212 (May 2019) were successful in enabling close interaction between researchers from diverse backgrounds. The attendees consistently remarked about the benefits of building bridges between the two communities. The goals from the previous seminars were to strengthen existing ties, establish new ones, identify challenges that require the two communities to work together, and establish mechanisms for increased communication and transfer of results from one to the other. A key goal of the current seminar was to additionally bring in experts from a few application domains to provide the necessary context for identifying research problems in topological data analysis and visualization. Furthermore, we also encouraged interaction between researchers who worked within the same community to identify challenging problems in the area and to establish new collaborations.

Topics

The research topics, listed below, reflect highly active and emerging areas in TDA. They were chosen to span topics in theory, algorithms, and applications.

Multivariate data analysis. Topics include theoretical studies of multivariate topological descriptors (including multiparameter persistence), efficient algorithms for computing and comparing them, formal guarantees for data analysis based on such comparisons, and the development of practical tools based on such analysis. Combining topological analysis together with statistical learning-based methods were also of interest.

Geometry and topology of metric spaces. A cornerstone of TDA is the study of metric and geometric data sets by means of filtrations of geometric complexes, formed by connecting subsets of the data points according to some proximity parameter. The study of such filtrations using homology leads to a multi-scale descriptor of the data that combines geometric and topological aspects of its shape. Besides their use in TDA, geometric complexes also play an important role in geometric group theory and metric geometry. The results and insights from both areas carry great promise for mutual interactions, leading to a unified view on computational and theoretical aspects.

Applications. TDA is an emerging area in exploratory data analysis and has received growing interest and notable successes with an expanding research community. The application of topological techniques to large and complex data has opened new opportunities in science, engineering, and business intelligence. This seminar focused on a few key application areas, including material sciences, neuroscience, and biology.

Parallel and distributed computation. The computational challenges in TDA call for the use of advanced techniques of high-performance computing, including parallel, distributed, and GPU-based software. Many of the core methods of TDA, including persistent homology,

mapper, merge trees, and contour trees, have received implementations beyond serial computing, and the interest in utilizing modern state-of-the-art techniques continues unabated. The task of optimizing algorithms in TDA is not only a question of engineering. Many of the key insights leading to breakthrough improvements are based on a careful utilization of theoretical properties and insights.

Participants, schedule, and organization

The invitees were chosen based on their background in mathematics, computer science, and application domains. We also ensured diversity in terms of gender, country or region of workplace, and experience.

While welcoming theoretical talks, the attendees were strongly encouraged to prepare a talk that is rooted in applications. The aim was to foster discussions on topics and projects related to practical applications of topological analysis and visualization. The program for the week consisted of talks of different lengths, breakout sessions, and summary / discussion sessions with all participants. We scheduled a total of six long talks (35 minutes + 10 minutes Q&A) on Day-1 and the morning session of Day-2, each providing an introduction either to one of the four chosen topics of the seminar or to a specific application domain. The talks were given by Yasu Hiraoka (Curse of dimensionality in persistence diagrams), Manish Saggarr (Precision dynamical mapping to anchor psychiatric diagnosis into biology), Andreas Ott (Topological data analysis and coronavirus evolution), Kelin Xia (Mathematical AI for molecular data analysis), Gunther Weber (Topological analysis for exascale computing: challenges & approaches), and Facundo Mémoli (Some recent results about Vietoris-Rips persistence). Short research talks (16 total, 15 minutes + 10 minutes Q&A) were scheduled during the morning sessions of Day-2 through Day-5.

The afternoon sessions were devoted to discussions, working groups, and interactions. On Monday, we led an open problem session where participants identified different open problems and future directions for research. This initial discussion helped identify working groups and topics for discussion during the week. We organized breakout sessions on Tuesday and Thursday. On Tuesday, after a quick discussion regarding discussion topics, we identified four topics of interest. Participants chose one of the four groups: the curse of dimensionality, distances on Morse and Morse-Smale complexes, computation of generalized persistence diagrams, Codistortion and Gromov–Hausdorff distance. After a quick discussion, we decided to continue discussions on the four topics on Thursday, and some participants chose to join a different group.

We organized an excursion to Trier on Wednesday afternoon followed by dinner at a restaurant. Many participants attended the guided tour and the dinner.

All working groups summarized the discussion during their breakout sessions and presented it to all participants on Thursday evening and Friday morning. These summary sessions were also interactive and resulted in follow-up discussions between smaller groups of participants. We organized a final discussion and feedback session on Friday morning to close the seminar and to make future plans.

Results and reflection

The schedule for the first day helped initiate interaction between participants and continue the discussions during the week. While the introductory talks provided sufficient details on interesting application domains, the open problem session allowed many participants to quickly pose topics of interest. In particular, the format of this session extended beyond proposing specific stated open problems, asking also for contributions, discussion points, and thoughts that would not typically be brought up in such a session. This resulted in a very lively and engaging discussion that encouraged participants to share their perspectives on important current and future research directions.

In summary, we think that the seminar was successful in achieving the objective of encouraging discussions and interaction between researchers with backgrounds in mathematics, computer science, and application domains who are interested in the areas of topological data analysis and visualization. It helped identify new directions for research and has hopefully sparked the engagement of researchers from one community into the activities and research workshops and venues of the other. We strongly believe that the seminar provided a highly valuable contribution to bridging the gap between theory and applications in TDA.

The participants were highly appreciative of the balance between theoretical and applied topics and between the participants and those who presented during the week. They highlighted that the diverse group of participants sharing a strong interest in novel perspectives and exchange of ideas made the workshop an exceptional experience. Several felt that the discussions helped them identify topics for future research or introduced them to new collaboration possibilities.

2 Table of Contents

Executive Summary

<i>Vijay Natarajan, Ulrich Bauer, and Bei Wang</i>	71
--	----

Overview of Talks

Quantifying and tracking inter-feature separation <i>Talha Bin Masood</i>	77
Density-based Riemannian metrics and persistent homology <i>Ximena Fernández</i>	77
Modified Finsler metrics for vector field visualization <i>Hans Hagen</i>	77
Persistent homology of a periodic filtration <i>Teresa Heiss</i>	78
Curse of dimensionality in persistence diagrams <i>Yasuaki Hiraoka</i>	78
Topological feature tracking in visualization applications <i>Ingrid Hotz</i>	78
The Density-Delaunay-Cech bifiltration <i>Michael Kerber</i>	79
Towards a theory of persistence for gradient-like Morse-Smale vector fields <i>Claudia Landi</i>	79
The (not so) mysterious rhomboid bifiltration <i>Michael Lesnick</i>	80
A spontaneous demo of the Topology ToolKit (TTK) <i>Joshua A. Levine</i>	80
Some recent results about Vietoris-Rips persistence <i>Facundo Mémoli</i>	81
Topological optimization with big steps <i>Dmitriy Morozov</i>	81
Topological data analysis and coronavirus evolution <i>Andreas Ott</i>	82
Precision dynamical mapping to anchor psychiatric diagnosis into biology <i>Manish Saggat</i>	82
Persistent homology of the multiscale clustering filtration <i>Dominik Schindler</i>	83
Persistence diagrams and Mobius inversion <i>Primoz Skraba</i>	83
Betti matching <i>Nico Stucki</i>	84
TGDA for graph learning? <i>Yusu Wang</i>	84

Topological analysis for exascale computing: challenges and approaches <i>Gunther Weber</i>	84
A distance for geometric graphs via labeled merge tree interleavings <i>Erin Moriarty Wolf Chambers</i>	85
Mathematical AI for molecular data analysis <i>Kelin Xia</i>	85
Minimal cycle representatives in persistent homology using linear programming <i>Lori Ziegelmeier</i>	86
Working groups	
Codistortion and Gromov–Hausdorff distance <i>Ulrich Bauer and Facundo Mémoli</i>	86
Curse of Dimensionality <i>Teresa Heiss, Ximena Fernández, Yasuaki Hiraoka, Claudia Landi, Andreas Ott, Manish Saggat, and Dominik Schindler</i>	88
Computation of Generalized Persistence Diagrams <i>Michael Lesnick, Teresa Heiss, Michael Kerber, Dmitriy Morozov, Primoz Skraba, and Nico Stucki</i>	90
Distances on Morse and Morse-Smale complexes <i>Erin Moriarty Wolf Chambers, Ulrich Bauer, Talha Bin Masood, Ximena Fernández, Hans Hagen, Ingrid Hotz, Claudia Landi, Joshua A. Levine, Vijay Natarajan, Dominik Schindler, Bei Wang, Yusu Wang, Gunther Weber, Kelin Xia, and Lori Ziegelmeier</i>	92
Participants	95

3 Overview of Talks

3.1 Quantifying and tracking inter-feature separation

Talha Bin Masood (Linköping University, SE)

License © Creative Commons BY 4.0 International license
© Talha Bin Masood

Topological descriptors such as merge trees and extremum graphs have proven to be very useful for multiscale feature-based analysis of scalar field data. However, in some applications extraction of features is not enough, understanding the separation/topological distance between the extracted features is also important. Intrinsic tree distance can be used to quantify this topological distance. I will talk about two different applications where quantifying feature separation is useful and has physically interpretable meaning. One of the challenges that arise in the context of time-varying or ensemble scalar field data is tracking and visualization of the change in inter-feature separation with the change in time or input parameters. I will present some preliminary ideas and results in this direction.

3.2 Density-based Riemannian metrics and persistent homology

Ximena Fernández (Durham University, GB)

License © Creative Commons BY 4.0 International license
© Ximena Fernández
Joint work of Ximena Fernández, Eugenio Borghini, Gabriel Mindlin, Pablo Groisman
URL https://ximenafernandez.github.io/reveal.js-presentations/slides/FermatDistance_Dagstuhl.html#/

Several methods for geometric inference relies in the choice of an appropriate metric in the sample point cloud. Consider a scenario where the data is a noisy sample of a manifold embedded in Euclidean space, drawn according to a positive density over the manifold. I propose to learn a metric directly from the data (called *Fermat distance*) that turns out to be an estimator of an intrinsic density-based metric over the underlying manifold. I will show some convergence results, robustness properties of the use of this metric in the computation of persistent homology and some applications in real data. I will also discuss a couple of open questions.

3.3 Modified Finsler metrics for vector field visualization

Hans Hagen (RPTU – Kaiserslautern, DE)

License © Creative Commons BY 4.0 International license
© Hans Hagen

Visualizing vector fields and their impact on free-form surface modelling like car hoods or airplane wings is a hot topic. We can “use” these vector fields to “deform” the metric of these surfaces, generating a Finsler metric. Can such a Finsler metric be useful for vector field visualization?

3.4 Persistent homology of a periodic filtration

Teresa Heiss (IST Austria – Klosterneuburg, AT)

License  Creative Commons BY 4.0 International license
© Teresa Heiss

Joint work of Teresa Heiss, Herbert Edelsbrunner, Chiara Martyka, Dmitriy Morozov

Persistent homology is well-defined and well studied for tame filtrations, for example various ones arising from finite point sets. However, periodic filtrations – for example used to study periodic point sets, like the atom positions of a crystal – are not tame, because there are infinitely many periodic copies of a homology class appearing at the same filtration value. We therefore extend the definition of persistent homology to periodic filtrations, which is a surprisingly difficult endeavor. In contrast to related work, we quantify how fast the multiplicities of persistence pairs tend to infinity with increasing window size, in a way that is stable under perturbations and invariant under different finite representations of the infinite periodic filtration. This project is still ongoing research, but I'll explain what we already know and what we don't know yet.

3.5 Curse of dimensionality in persistence diagrams

Yasuaki Hiraoka (Kyoto University, JP)

License  Creative Commons BY 4.0 International license
© Yasuaki Hiraoka

It is well known that persistence diagrams stably behave under small perturbations to the input data. This is the consequence of stability theorems, firstly proved by Cohen-Steiner, Edelsbrunner, and Harer (2007), and then extended by several researchers. On the other hand, if the input data is realized in a high-dimensional space with a small noise, the curse of dimensionality (CoD) causes serious adverse effects on data analysis, especially leading to inconsistency of distances. In this talk, I will show several examples of CoD appearing in persistence diagrams (e.g., from single-cell RNA sequencing data in biology). Those examples demonstrate that the classical stability theorems are not sufficient to guarantee stable behaviors of persistence diagrams for high-dimensional data. Then I will show several mathematical results about the existence and the (partial) resolution of CoD in persistence diagrams. This is a joint work with Liu Enhao, Yusuke Imoto and Shu Kanazawa.

3.6 Topological feature tracking in visualization applications

Ingrid Hotz (Linköping University, SE)

License  Creative Commons BY 4.0 International license
© Ingrid Hotz

Topology in visualization – balance between beautiful concepts and practical needs

Tracking of features is a fundamental task in visual data analysis. In our work, we use topological descriptors as an abstraction for tracking. An essential step thereby is the choice of appropriate similarity measures to detect structural changes and establish a correspondence between individual features respecting their spatial embedding. One way to approach both demands is to consider labeled merge trees as the feature descriptor. In this talk, some examples of such approaches for tracking features are discussed.

3.7 The Density-Delaunay-Cech bifiltration



Michael Kerber (TU Graz, AT)

License  Creative Commons BY 4.0 International license
 Michael Kerber

The density-Rips bifiltration is a standard construction in multi-parameter persistence, but suffers from the size explosion, as its single-parameter counterpart. On the other hand, it is well-known that at least in low Euclidean dimensions, alpha filtrations are much faster to compute and also geometrically more accurate. There are two major challenges to define and compute alpha-filtrations for two parameters. I will propose a way how to handle them. This is (very) ongoing work with Angel Alonso (TU Graz).

3.8 Towards a theory of persistence for gradient-like Morse-Smale vector fields

Claudia Landi (University of Modena, IT)

License  Creative Commons BY 4.0 International license
 Claudia Landi
 Joint work of Claudia Landi, Clemens Luc Bannwart

In topological data analysis, a function $f : M \rightarrow R$ is often studied through the homology of its sublevel sets. One can obtain a topological summary of f in the form of a persistence barcode [3]. By a result of Morse theory, if M is a closed manifold and f is nice enough, then M is homotopy equivalent to a CW-complex with one k -cell for each critical point of index k [5]. Persistent homology and Morse theory are closely related, since the values of f at the critical points are equal to the start- and endpoints of the bars in the barcode. The gradient of f induces a chain complex, where the boundary operator is defined by counting the flow lines between critical points (see e.g. [1]). This process works more generally for gradient-like Morse-Smale vector fields and also for combinatorial vector fields in the sense of Forman [4]. However, for gradient-like Morse-Smale vector fields, there does not yet exist a persistence barcode such as for functions.

We present a pipeline that takes as an input a gradient-like Morse-Smale vector field on a surface, produces a parameterized epimorphic chain complex, and encodes it as a barcode. More precisely, we produce a sequence of chain complexes, such that the first one is the chain complex induced by the vector field and after that, each one is a quotient of the previous one. These quotients correspond to topological simplifications of the vector field by certain moves (introduced in [2]), and the times of taking the quotients depend on the value of a parameter measuring the local robustness of the vector field. In the end we are left with a vector field that has a very simple topological structure. Geometrically, for each move that is applied, we extract a topological feature. Algebraically, for each quotient, we split off an indecomposable contractible summand from the initial chain complex. Remembering the times when the moves were applied then yields a barcode. Similarly to the usual persistent homology construction for real valued functions, this pipeline paves the way for the development of a theory of persistence for vector fields.

References

- 1 A. Banyaga and D. Hurtubise, Lectures on Morse homology, Texts in the Mathematical Sciences, Springer Netherlands, 2004.
- 2 M. J. Catanzaro, J. Curry, B. T. Fasy, J. Lazovskis, G. Malen, H. Riess, B. Wang, and M. J. Zabka, Moduli spaces of Morse functions for persistence, *Journal of Applied and Computational Topology* (2019), 1–33.
- 3 H. Edelsbrunner and J. Harer, Persistent homology—a survey, *Discrete & Computational Geometry – DCG 453* (2008).
- 4 R. Forman, Combinatorial vector fields and dynamical systems, *Mathematische Zeitschrift* 228 (1998), 629–681.
- 5 J. W. Milnor, Morse theory, *Annals of mathematics studies*, Princeton University Press, 1963.

3.9 The (not so) mysterious rhomboid bifiltration

Michael Lesnick (University at Albany, US)

License  Creative Commons BY 4.0 International license
© Michael Lesnick

The multicover bifiltration is a density-sensitive extension of the union-of-balls bifiltration commonly considered in TDA. It is robust to outliers, in a strong sense, and doesn't depend on any extra parameters. These properties make the multicover bifiltration a natural candidate for applications, if it can be computed. With this in mind, Edelsbrunner and Osang introduced a polyhedral bifiltration called the rhomboid bifiltration and gave a polynomial time algorithm for computing it. Corbet et al. showed that this bifiltration is topologically equivalent to the multicover bifiltration. In this talk, I'll give a poset-theoretic definition of the rhomboid tiling which is different from (but equivalent to) the one given by Edelsbrunner and Osang. With this as inspiration, I'll sketch a new proof of topological equivalence of the multicover and rhomboid bifiltrations.

3.10 A spontaneous demo of the Topology ToolKit (TTK)

Joshua A. Levine (University of Arizona – Tucson, US)

License  Creative Commons BY 4.0 International license
© Joshua A. Levine

In this short talk, I'll give a brief of overview of some of the features of the Topology ToolKit, a software package for topological data analysis of scalar fields. Rather than diving into the implementation details, this presentation will focus on ease of use and applications. To demonstrate, I'll walk through a surprise demo.

TTK comes shipped with Kitware's ParaView, and it can also be built from source. Many more examples are available at <https://topology-tool-kit.github.io/examples/index.html>.

3.11 Some recent results about Vietoris-Rips persistence

Facundo Mémoli (Ohio State University – Columbus, US)

License © Creative Commons BY 4.0 International license
© Facundo Mémoli

Persistence barcodes provide computable signatures for datasets (metric spaces). These signatures absorb both geometric and topological information from metric spaces in a stable manner.

One question that motivated our work is: how strong are these signatures? A related question is that of ascertaining their relationship to other more classical invariants such as curvature.

In this talk I will describe some results about characterizing metric spaces via persistence barcodes arising from Vietoris-Rips filtrations. Of particular interest is a relationship which we established linking persistence barcodes to Gromov's filling radius.

Another aspect I will mention is the determination of the Gromov-Hausdorff distance between spheres (when endowed with their geodesic distance). In this case, VR-barcodes do permit telling spheres apart, but $1/2$ of the bottleneck distance does not match the exact value of the GH-distance.

This work is joint with Sunhyuk Lim, Osman Okutan, and Zane Smith.

3.12 Topological optimization with big steps

Dmitriy Morozov (Lawrence Berkeley National Laboratory, US)

License © Creative Commons BY 4.0 International license
© Dmitriy Morozov
Joint work of Dmitriy Morozov, Arnur Nigmatov

Using persistent homology to guide optimization has emerged as a novel application of topological data analysis. Existing methods treat persistence calculation as a black box and backpropagate gradients only onto the simplices involved in particular pairs. We show how the cycles and chains used in the persistence calculation can be used to prescribe gradients to larger subsets of the domain. In particular, we show that in a special case, which serves as a building block for general losses, the problem can be solved exactly in linear time. We present empirical experiments that show the practical benefits of our algorithm: the number of steps required for the optimization is reduced by an order of magnitude.

3.13 Topological data analysis and coronavirus evolution

Andreas Ott (KIT – Karlsruher Institut für Technologie, DE)

License © Creative Commons BY 4.0 International license
© Andreas Ott

Joint work of Michael Bleher, Lukas Hahn, Maximilian Neumann, Juan Ángel Patiño-Galindo, Mathieu Carrière, Ulrich Bauer, Raúl Rabadán, Andreas Ott, KIT Steinbuch Centre for Computing

Main reference Michael Bleher, Lukas Hahn, Maximilian Neumann, Juan Ángel Patiño-Galindo, Mathieu Carrière, Ulrich Bauer, Raúl Rabadán, Andreas Ott: “Topological data analysis identifies emerging adaptive mutations in SARS-CoV-2”, medRxiv, Cold Spring Harbor Laboratory Press, 2023.

URL <https://doi.org/10.1101/2021.06.10.21258550>

Topological methods have in recent years found applications in the life sciences. In this talk, I will present an application of persistent homology to the surveillance of critical mutations in the evolution of the coronavirus SARS-CoV-2. I will explain the underlying geometric idea, how it connects with biology, its implementation in the CoVtRec pipeline, and some concrete results from the analysis of current pandemic data.

3.14 Precision dynamical mapping to anchor psychiatric diagnosis into biology

Manish Saggar (Stanford University, US)

License © Creative Commons BY 4.0 International license
© Manish Saggar

URL <https://braindynamicslab.github.io/projects/dp2/>

Understanding the neurobiological underpinnings of psychiatric disorders has long been a challenge in the field of neuroscience. This talk aims to address this issue by exploring how noninvasive neuroimaging, despite its inherent limitations, can be leveraged to anchor psychiatric disorders into neurobiology. Two main challenges in this endeavor are identified: (a) the inherent noise in noninvasive neuroimaging devices, and (b) the limited utilization of biophysical models.

To tackle the first challenge, we propose the application of Topological Data Analysis (TDA), specifically Mapper, as a novel approach. I present some promising results on how Mapper can capture evoked transitions during tasks, intrinsic transitions during resting states, and changes in the landscape or shape associated with psychiatric disorders such as Major Depressive Disorder (MDD), Attention Deficit Hyperactivity Disorder (ADHD), as well as various pharmacological interventions (e.g., Methylphenidate, Psilocybin) and neuromodulation techniques (e.g., sp-TMS, rTMS).

I will also highlight methodological advances in TDA that enhance its applicability in the context of noninvasive neuroimaging studies. By harnessing the power of TDA, we can gain deeper insights into the complex dynamics of brain activity and its relation to psychiatric disorders.

Finally, the talk concludes by posing open questions that warrant further investigation. These questions touch upon the potential integration of TDA with other analytical approaches, the optimization of experimental protocols, and the translation of findings into clinical practice. By addressing these open questions, we can foster a greater understanding of the neurobiological basis of psychiatric disorders and pave the way for innovative therapeutic strategies.

3.15 Persistent homology of the multiscale clustering filtration

Dominik Schindler (Imperial College London, GB)

License © Creative Commons BY 4.0 International license
© Dominik Schindler

Joint work of Dominik Schindler, Mauricio Barahona

Main reference Dominik J. Schindler, Mauricio Barahona: “Persistent Homology of the Multiscale Clustering Filtration”, *CoRR*, Vol. abs/2305.04281, 2023.

URL <https://doi.org/10.48550/arXiv.2305.04281>

In many applications in data clustering, it is desirable to find not just a single partition but a sequence of partitions that describes the data at different scales, or levels of coarseness, leading naturally to Sankey diagrams as descriptors of the data. The problem of multiscale clustering then becomes how to select robust intrinsic scales, and how to analyse and compare the (not necessarily hierarchical) sequences of partitions. Here, we define a novel filtration, the Multiscale Clustering Filtration (MCF), which encodes arbitrary patterns of cluster assignments across scales. We prove that the MCF is a proper filtration, give an equivalent construction via nerves, and show that in the hierarchical case the MCF reduces to the Vietoris-Rips filtration of an ultrametric space. We also show that the zero-dimensional persistent homology of the MCF provides a measure of the level of hierarchy in the sequence of partitions, whereas the higher-dimensional persistent homology tracks the emergence and resolution of conflicts between cluster assignments across scales. We briefly illustrate numerically how the structure of the persistence diagram can serve to characterise multiscale data clusterings.

3.16 Persistence diagrams and Mobius inversion

Primoz Skraba (Queen Mary University of London, GB & Jožef Stefan Institute – Ljubljana, SI)

License © Creative Commons BY 4.0 International license
© Primoz Skraba

Joint work of Primoz Skraba, Amit Patel

There are many ways of defining persistence diagrams. In this talk I will discuss the definition based on the Mobius inversion function which was introduced by Amit Patel under the name Generalized Persistence Diagrams. I will cover how this approach has appeared implicitly and explicitly in various results on persistence as well as various implications of this approach and (very) new developments. In particular, I will cover a surprising connection between Euler characteristics and persistence diagrams and discuss the many questions and directions which arise.

3.17 Betti matching

Nico Stucki (TU München, DE)

License  Creative Commons BY 4.0 International license
© Nico Stucki

Main reference Nico Stucki, Johannes C. Paetzold, Suprosanna Shit, Bjoern H. Menze, Ulrich Bauer: “Topologically faithful image segmentation via induced matching of persistence barcodes”, CoRR, Vol. abs/2211.15272, 2022.

URL <https://doi.org/10.48550/arXiv.2211.15272>

Segmentation models predominantly optimize pixel-overlap-based loss, an objective that is actually inadequate for many segmentation tasks. In recent years, their limitations fueled a growing interest in topology-aware methods, which aim to recover the topology of the segmented structures. However, so far, existing methods only consider global topological properties, ignoring the need to preserve topological features spatially, which is crucial for accurate segmentation. We introduce the concept of induced matchings from persistent homology to achieve a spatially correct matching between persistence barcodes in a segmentation setting. Based on this concept, we define the Betti matching error as an interpretable, topologically and feature-wise accurate metric for image segmentation, which resolves the limitations of the Betti number error. The Betti matching error is differentiable and efficient to use as a loss function. We demonstrate that it improves the topological performance of segmentation networks significantly across six diverse datasets while preserving the performance with respect to traditional scores.

3.18 TGDA for graph learning?


Yusu Wang (University of California, San Diego – La Jolla, US)

License  Creative Commons BY 4.0 International license
© Yusu Wang

In recent years, graph neural networks have emerged as a power family of ML architectures for graph learning and optimization. Nevertheless, various limitations and challenges remain. In this talk, I will briefly introduce the message passing graph neural networks (MPNN), and describe a few results in aiming to provide better understanding of GNNs or to enhance their power using geometric and topological ideas. My goal is to stimulate further discussions / interests / new perspectives in this interesting direction of TGDA + GNN.

3.19 Topological analysis for exascale computing: challenges and approaches

Gunther Weber (Lawrence Berkeley National Laboratory, US)

License  Creative Commons BY 4.0 International license
© Gunther Weber

Simulation has quickly evolved to become the “third pillar of science” and supercomputing centers provide the computational power needed for accurate simulations. The Exascale Computing Project (ECP) is a concentrated effort to cross the next barrier and build a supercomputer that can run simulations at quintillion calculations per second. Exascale computing exacerbates the already existing I/O-bottleneck that makes it impossible to write

all simulation results to disk. To mitigate this problem, in situ approaches perform data analysis and visualization while the simulation is running. This talk provides an overview over how topological data analysis enables automated choice of visualization parameters like isovalue for isosurface extraction. It furthermore outlines the challenges that current developments in supercomputer architecture pose to efficient algorithm design for topological data analysis and presents solution approaches.

3.20 A distance for geometric graphs via labeled merge tree interleavings

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

License © Creative Commons BY 4.0 International license
© Erin Moriarty Wolf Chambers

Geometric graphs appear in many real world datasets, such as road networks, sensor networks, and molecules. We investigate the notion of distance between graphs and present a semi-metric to measure the distance between two geometric graphs via the directional transform combined with the labeled merge tree distance. Our distance is not only reflective of the information from the input graphs, but also can be computed in polynomial time. We illustrate its utility by implementation on a Passiflora leaf dataset.

3.21 Mathematical AI for molecular data analysis

Kelin Xia (Nanyang TU – Singapore, SG)

License © Creative Commons BY 4.0 International license
© Kelin Xia

Artificial intelligence (AI) based molecular data analysis has begun to gain momentum due to the great advancement in experimental data, computational power and learning models. However, a major issue that remains for all AI-based learning models is the efficient molecular representations and featurization. Here we propose advanced mathematics-based molecular representations and featurization (or feature engineering). Molecular structures and their interactions are represented as various simplicial complexes (Rips complex, Neighborhood complex, Dowker complex, and Hom-complex), hypergraphs, and Tor-algebra-based models. Molecular descriptors are systematically generated from various persistent invariants, including persistent homology, persistent Ricci curvature, persistent spectral, and persistent Tor-algebra. These features are combined with machine learning and deep learning models, including random forest, CNN, RNN, GNN, Transformer, BERT, and others. They have demonstrated great advantage over traditional models in drug design and material informatics.

3.22 Minimal cycle representatives in persistent homology using linear programming

Lori Ziegelmeier (Macalester College – St. Paul, US)

License  Creative Commons BY 4.0 International license
© Lori Ziegelmeier

Main reference Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, Lori Ziegelmeier: “Minimal Cycle Representatives in Persistent Homology Using Linear Programming: An Empirical Study With User’s Guide”, *Frontiers Artif. Intell.*, Vol. 4, p. 681117, 2021.

URL <https://doi.org/10.3389/frai.2021.681117>

Cycle representatives of persistent homology classes can be used to provide descriptions of topological features in data. However, the non-uniqueness of these representatives creates ambiguity and can lead to many different interpretations of the same set of classes. One approach to solving this problem is to optimize the choice of representative against some measure that is meaningful in the context of the data. In this work, we provide a study of the effectiveness and computational cost of several ℓ_1 -minimization optimization procedures for constructing homological cycle bases for persistent homology with rational coefficients in dimension one, including uniform-weighted and length-weighted edge-loss algorithms as well as uniform-weighted and area-weighted triangle-loss algorithms. We conduct these optimizations via standard linear programming methods, applying general-purpose solvers to optimize over column bases of simplicial boundary matrices.

Our key findings are: (i) optimization is effective in reducing the size of cycle representatives, (ii) the computational cost of optimizing a basis of cycle representatives exceeds the cost of computing such a basis in most data sets we consider, (iii) the choice of linear solvers matters a lot to the computation time of optimizing cycles, (iv) the computation time of solving an integer program is not significantly longer than the computation time of solving a linear program for most of the cycle representatives, using the Gurobi linear solver, (v) strikingly, whether requiring integer solutions or not, we almost always obtain a solution with the same cost and almost all solutions found have entries in $-1, 0, 1$ and therefore, are also solutions to a restricted ℓ_0 optimization problem, and (vi) we obtain qualitatively different results for generators in Erdős-Rényi random clique complexes.

4 Working groups

4.1 Codistortion and Gromov–Hausdorff distance

Ulrich Bauer (TU München, DE) and Facundo Mémoli (Ohio State University – Columbus, US)

License  Creative Commons BY 4.0 International license
© Ulrich Bauer and Facundo Mémoli

Let \mathcal{M} be the collection of compact metric spaces. The *Gromov–Hausdorff distance* between (X, d_X) and (Y, d_Y) in \mathcal{M} is defined as

$$d_{GH}(X, Y) = \frac{1}{2} \inf_{\phi: X \leftrightarrow Y: \psi} \max(\text{dis}(\phi), \text{dis}(\psi), \text{codis}(\phi, \psi)),$$

where

$$\text{dis}(\phi) = \sup_{x, x' \in X} |d_X(x, x') - d_Y(\phi(x), \phi(x'))|,$$

$$\text{codis}(\phi, \psi) = \sup_{x \in X, y \in Y} |d_X(x, \psi(y)) - d_Y(\phi(x), y)|$$

are the *distortion* of a map and the *codistortion* of a pair of maps between metric spaces, respectively. Separating the distortion and codistortion terms in this formula for the Gromov–Hausdorff distance, we obtain the variants

$$\hat{d}_{GH}(X, Y) = \frac{1}{2} \inf_{\phi: X \leftrightarrow Y: \psi} \max(\text{dis}(\phi), \text{dis}(\psi)),$$

$$\check{d}_{GH}(X, Y) = \frac{1}{2} \inf_{\phi: X \leftrightarrow Y: \psi} \text{codis}(\phi, \psi).$$

► **Example 1.** If $*$ denotes the one point metric space, then we have $\check{d}_{GH}(X, *) = \frac{1}{2}\text{rad}(X)$.

Clearly $\hat{d}_{GH}, \check{d}_{GH} \leq d_{GH}$. The following facts about the *distortion distance* \hat{d}_{GH} are known. Below \cong denotes the equivalence relation of isometry on \mathcal{M} .

1. \hat{d}_{GH} is a legit distance on the set of isometry classes of compact metric spaces \mathcal{M}/\cong .
2. \hat{d}_{GH} and d_{GH} generate the same topology.
3. $\hat{d}_{GH} \neq d_{GH}$.
4. \hat{d}_{GH} can be computed via curvature sets.

Less is known about the *codistortion distance* \check{d}_{GH} . We state a few interesting questions.

1. Is \check{d}_{GH} a distance on \mathcal{M}/\cong ?
2. Is \check{d}_{GH} bi-Lipschitz equivalent to d_{GH} ?

In our discussion group, we answered these questions to the affirmative.

► **Proposition 2.**

$$\check{d}_{GH} \leq d_{GH} \leq 2\check{d}_{GH}.$$

► **Remark.** The inequality $d_{GH} \leq 2\check{d}_{GH}$ is tight. To see this, consider the finite metric spaces X consisting of two points at distance 4 and Y consisting of the three points $\{0, 2, 3\}$ on the real line (with the usual metric).

► **Theorem 3.** \check{d}_{GH} a legitimate distance on \mathcal{M}/\cong .

The following lemma is key to relating distortion and codistortion.

► **Lemma 4.** Consider a pair of maps $\phi : X \leftrightarrow Y : \psi$ between metric spaces. Then

- $\text{codis}(\phi, \psi) \geq \sup_{x \in X} d_X(x, \psi \circ \phi(x))$.
- $2 \text{codis}(\phi, \psi) \geq \max(\text{dis}(\phi), \text{dis}(\psi))$,

Further insights and questions

1. \hat{d}_{GH} is not bi-Lipschitz equivalent to d_{GH} . There is a family of pairs of finite ultrametric spaces $(X_k, Y_k)_k$ such that $d_{GH}(X_k, Y_k) \geq \frac{k}{2}$ but $\hat{d}_{GH}(X_k, Y_k) \leq 1$.
2. For all compact metric spaces X and Y we have

$$\check{d}_{GH}(X, Y) \geq \frac{1}{4} d_B(\text{dgm}(X), \text{dgm}(Y)),$$

where $\text{dgm}(X)$ denotes the usual persistence diagram of the Vietoris-Rips filtration of X . Can this be improved to

$$\check{d}_{GH}(X, Y) \geq \frac{1}{2} d_B(\text{dgm}(X), \text{dgm}(Y))?$$

The stronger bound, when combined with the fact that $\check{d}_{GH} \leq d_{GH}$, would imply an improvement upon the usual Gromov–Hausdorff stability theorem for persistence diagrams arising from Vietoris-Rips filtrations.

3. Is it true that $\check{d}_{GH} \geq \check{d}_{GH} \circ H^{sl}$, where H^{sl} is single-linkage clustering, taking a finite metric space to a finite ultrametric space?
4. Is there a case where $\check{d}_{GH} < d_{GH}$?
5. What are natural lower bounds for \check{d}_{GH} ? One of them is half the the difference of the respective radii of the spaces:

$$\check{d}_{GH}(X, Y) \geq \frac{1}{2} |\text{rad}(X) - \text{rad}(Y)|.$$

6. If $\text{codis}(\phi, \psi) < \delta$, then $\text{codis}(\phi \circ \psi \circ \phi, \psi \circ \phi \circ \psi) < 2\delta$.
7. If X and Y are ultrametric, do we have $\check{d}_{GH}(X, Y) = d_{GH}(X, Y)$?
8. Is there a constant $C > 0$ such that $d_H^Y(\phi(X), Y), d_H^X(\psi(Y), X) \leq C \cdot \text{codis}(\phi, \psi)$?

4.2 Curse of Dimensionality

Teresa Heiss (IST Austria – Klosterneuburg, AT), Ximena Fernández (Durham University, GB), Yasuaki Hiraoka (Kyoto University, JP), Claudia Landi (University of Modena, IT), Andreas Ott (KIT – Karlsruher Institut für Technologie, DE), Manish Saggat (Stanford University, US), and Dominik Schindler (Imperial College London, GB)

License © Creative Commons BY 4.0 International license
 © Teresa Heiss, Ximena Fernández, Yasuaki Hiraoka, Claudia Landi, Andreas Ott, Manish Saggat, and Dominik Schindler

The stability result of Persistent Homology is not guaranteeing much in very high dimensions, when noise of at most ε is added in each dimension to the data. Indeed, the ℓ_2 -distance between a point $x \in \mathbb{R}^d$ and its perturbed point $x + p$ with $\|p\|_{\ell_\infty} < \varepsilon$ is $\|p\|_{\ell_2} \leq \sqrt{d}\varepsilon$. Hence, the stability bound, namely the Hausdorff distance between the original and the perturbed point set, is $O(\sqrt{d}\varepsilon)$ as well.

We prove that this effect, the curse of dimensionality, cannot be circumvented in full generality, i.e., when an adversary is allowed to make the choices. This shows that we need some assumption on the data. We list some ideas for possible assumptions, and approaches that seem promising within these different assumptions.

Setting

The setting is for example motivated by gene expression data, with few (s) essential genes, many more ($d - s$) housekeeping genes, and a small measuring error for each gene. The persistence diagram of such data will, due to the curse of dimensionality, be very different than the desired persistence diagram of only the essential genes.

Given $\varepsilon > 0$, $s \in \mathbb{N}$, $A \subseteq \mathbb{R}^s$, an integer $d > s$, an affine linear map $L : \mathbb{R}^s \rightarrow \mathbb{R}^d$ with determinant 1, and for every $a \in A$, a vector p_a with $\|p_a\|_{\ell_\infty} < \varepsilon$. We denote the embedded point set $L(A)$ by X and the perturbed set $\{L(a) + p_a \mid a \in A\}$ by Y .

The Gromov-Hausdorff distance is $d_{GH}(A, Y) = d_{GH}(X, Y) \leq \max_{a \in A} \|p_a\|_{\ell_2} \leq \sqrt{d}\varepsilon$ and thus for $d \gg s$ the stability theorem does not give a good bound for the Vietoris-Rips persistence diagrams:

$$d_B(PD(VR(A)), PD(VR(Y))) \leq 2d_{GH}(A, Y) = O(\sqrt{d}\varepsilon). \quad (1)$$

Note that since we consider getting $d_B(PD(VR(A)), PD(VR(Y))) = O(1)$ by matching everything to the diagonal as “cheating”, one might want to consider a distance between persistence diagrams that does not allow matchings with the diagonal, like the Hausdorff distance. Another approach is to keep using the bottleneck distance and not be satisfied with $d_B(PD(VR(A)), PD(VR(Y))) = O(1)$, but insisting on wanting $d_B(PD(VR(A)), PD(VR(Y))) = O(1)\varepsilon$ or $o(1)$ as ε goes to 0.

We are searching for a modification $Z \subseteq \mathbb{R}^d$ of the observed data Y , such that $d_B(PD(VR(A)), PD(VR(Z))) = o(1)$. For example by dimensionality reduction.

When the Adversary Makes the Choices

There cannot be any fix to the curse of dimensionality in full generality, as the following argument shows. For every ε , and every $s \geq 1$, an adversary can choose

- the point set A as two points on the x -axis with distance 1 from each other,
- the embedding dimension $d > \frac{1}{\varepsilon^2}$,
- the affine linear map L with determinant 1 to map the x -axis to the direction spanned by the vector $(1, 1, \dots, 1)$,
- and the two vectors p_1 and p_2 such that $L(a_1) + p_{a_1} = L(a_2) + p_{a_2}$, e.g. $p_{a_1} = \frac{1}{\sqrt{d}}(1, 1, \dots, 1)$ and $p_{a_2} = 0$. As $\frac{1}{\sqrt{d}} < \varepsilon$, such a perturbation is allowed.

Then, the observed set Y consists of two points in the same spot, and thus does not have any non-essential homology, whereas A has a persistence pair with persistence 1. Hence, the distance $d_B(PD(VR(A)), PD(VR(Y))) = 1$ does not converge to zero when ε goes to zero. Furthermore, since all structure of A has been destroyed in Y , there is no hope to reconstruct an adequate modification Z to reconstruct the persistence of A , since when only given Y , we cannot know whether it has been created from a set A consisting of two points in the same position that have not been perturbed at all or from the above set A .

This shows that in order to have a chance against the curse of dimensionality, we need some assumptions on our data, instead of letting the adversary choose the data. Note that in the proof above, it was essential to let the adversary choose the embedding dimension d , the map L , and the perturbation vectors. Choosing A does not seem to be essential, it just makes the proof more convenient.

Possible Assumptions

Since we want to talk about the curse of dimensionality, we do not want to bound the embedding dimension d but instead keep letting the adversary choose d , or in other words imagine d as very large. Instead we can make assumptions on the affine linear map L or on the perturbation vectors:

1. A weak assumption would be assuming that the perturbation vectors have the form $p_a = w_a v_a$ with $w_a > 0$ an unknown constant depending on a , and v_a i.i.d. with an unknown distribution.
2. One can strengthen this by assuming a fixed known distribution for the v_a .
3. Or assuming that $w_a = 1$.
4. Another approach is assuming that the map L is axis-parallel or not too far from axis parallel.
5. Maybe assuming that the data is sampled very densely (for example $\frac{d}{n}$ converging to a constant).


Possible Solutions

Possible ideas how to pass from Y to Z :

- Assuming Assumption 4 above, there are $d - s$ coordinates that are pure noise. Hence, choose the s coordinates with the most variance and set the other coordinates to zero. In application where the noise p_a might be approximately linearly depending on the length $\|a\|_{\ell_2}$, one could for example replace the variance by the variance divided by the mean.
- Assuming at least Assumption 1 above: Use neural network auto-encoder (and afterwards possibly UMAP?) for dimensionality reduction from Y to Z .
- Assuming assumption 2: Mimic what RECODE does, namely, if I understand correctly, using a PCA technique that is designed by statisticians for weakening the curse of dimensionality for that particular distribution.
- Assuming at least Assumption 1 above: Dominik's idea: apply some variant of hierarchical clustering to observed data and obtain dendrogram -> this leads to an ultrametric -> we can analyze the new ultrametric space with Vietoris Rips Persistent Homology. Alternatively: clustering and then MCF.
- Assume at least Assumptions 1 and 5 above: Hope that something like the law of large numbers would yield that the effects of the many strong (up to $O(\sqrt{d}\epsilon)$) perturbations average each other out, such that a degree-Rips / density-Rips approach or something similar to Ximena's work might filter out the noise.
- Assuming maybe Assumption 1 or 4 above: Some kind of bootstrap idea would be to subsample, say s (or a bit more) out of d , dimensions many times, knowing that most of the time one would mostly just get the noise, but maybe there is some way to distinguish the non-noise persistence diagrams from the purely-noise ones. The advantage would be that the diagram where the correct s dimensions are selected, would not have the curse of dimensionality. But it seems difficult to extract this useful information from the huge bag of persistence diagrams. Furthermore, $\binom{d}{s}$ is very large, so it does not seem feasible from a computational perspective.
- Additional to the other ideas, Primoz Skraba said that it might help to look at persistence rather as death divided by birth, rather than death – birth. However, that alone would not be enough of course.

4.3 Computation of Generalized Persistence Diagrams

Michael Lesnick (University at Albany, US), Teresa Heiss (IST Austria – Klosterneuburg, AT), Michael Kerber (TU Graz, AT), Dmitriy Morozov (Lawrence Berkeley National Laboratory, US), Primoz Skraba (Queen Mary University of London, GB & Jožef Stefan Institute – Ljubljana, SI), and Nico Stucki (TU München, DE)

License  Creative Commons BY 4.0 International license

© Michael Lesnick, Teresa Heiss, Michael Kerber, Dmitriy Morozov, Primoz Skraba, and Nico Stucki

One breakout session was focused loosely on understanding the problem of computing generalized persistence diagrams (GPDs), as defined by Kim and Mémoli.

Given a poset P , persistence module $M : P \rightarrow \text{Vec}$, and a *generalized interval* (a.k.a. *spread*) $I \subset P$, the generalized rank of M over I is the rank of the map

$$\lim_I M \rightarrow \text{colim}_I M.$$

The map sending each such I to its generalized rank is the *generalized rank invariant* (GRI) of M . Taking the Möbius inversion of the GRI yields the *generalized persistence diagram* (GPD), a kind of signed barcode for generalized persistence.

Signed barcodes have become a hot topic in TDA in the last few years. There are multiple ways to define a signed barcode, namely, by taking Möbius inversions of different functions, or by relative homological algebra with respect to different exact structures.

Among the various options, the GPD studied here is an appealing choice because it is a relatively rich invariant and also has a very simple interpretation in the case of *spread-decomposable modules*: On such modules, the GPD simply counts the number of copies of each spread in the decomposition. In contrast, other types of signed barcodes can be rather complicated on such modules. This makes the problem of computing the GPD interesting. This problem is mostly open, in spite of some interesting recent work by Dey, Kim, Mémoli on the related problem of computing the GRI at fixed indices.

Our group explored (in a very preliminary way), the following related questions:

- What does the GPD look like on specific examples of non-spread decomposable modules? How quickly does its size grow as the support of the module grows.
- In the special case that M has a small *encoding* in the sense of Ezra Miller’s work (i.e., there exists a surjection of posets $f : P \rightarrow Q$ and a functor $N : Q \rightarrow \text{Vec}$ with $f = N \circ g$ and $|Q|$ small), is efficient computation of M possible? How does the complexity of computing the GPD depend on $|Q|$? What bounds on $|Q|$ can be expected in practical 2-parameter persistence computations? How does one compute f ?
- Can the ideas underlying recent work by Morozov and Patel on the output-sensitive computation of signed barcodes for 2-parameter persistence also be useful for computing GPDs?

There was some progress made. The first example we looked at was

$$\begin{array}{ccccc}
 \mathbb{k} & \longrightarrow & \mathbb{k}^2 & \longrightarrow & \mathbb{k}^2 \\
 & & \uparrow & & \uparrow \\
 & & \mathbb{k} & \longrightarrow & \mathbb{k}^2 \\
 & & & & \uparrow \\
 & & & & \mathbb{k}
 \end{array}$$

We made an attempt to understand if a module was nearly interval indecomposable, how complex could the generalized rank invariant be. Under the appropriate choice of morphisms, the above decomposes into many non-trivial pieces in the generalized rank invariant.

Following up on this, there was also a discussion on whether such non-interval indecomposables occur in practice/arises in random settings, e.g. from random point clouds. Michael Kerber suggested a bifiltration example which was finite and the point positions were generic, i.e. a positive but small perturbation does not affect the decomposition. We then showed that this will occur in a uniform Poisson point process with probability 1 as the number of points goes to ∞ . The idea behind the proof is that one can define a random variable of the event of such a configuration occurring, i.e. using the small neighborhood of the example. As the configuration is finite and generic, the probability of the event is strictly positive. As the expected number of such neighborhoods goes to ∞ , the probability must go to 1. Michael Kerber also reported his experimental results which indeed show that non-interval indecomposables arise in many experiments.

While we did not make decisive progress, the discussions were illuminating and left us with a better understanding of invariants and their computation.

4.4 Distances on Morse and Morse-Smale complexes

Erin Moriarty Wolf Chambers (St. Louis University, US), Ulrich Bauer (TU München, DE), Talha Bin Masood (Linköping University, SE), Ximena Fernández (Durham University, GB), Hans Hagen (RPTU – Kaiserslautern, DE), Ingrid Hotz (Linköping University, SE), Claudia Landi (University of Modena, IT), Joshua A. Levine (University of Arizona – Tucson, US), Vijay Natarajan (Indian Institute of Science – Bangalore, IN), Dominik Schindler (Imperial College London, GB), Bei Wang (University of Utah – Salt Lake City, US), Yusu Wang (University of California, San Diego – La Jolla, US), Gunther Weber (Lawrence Berkeley National Laboratory, US), Kelin Xia (Nanyang TU – Singapore, SG), and Lori Ziegelmeier (Macalester College – St. Paul, US)

License © Creative Commons BY 4.0 International license

© Erin Moriarty Wolf Chambers, Ulrich Bauer, Talha Bin Masood, Ximena Fernández, Hans Hagen, Ingrid Hotz, Claudia Landi, Joshua A. Levine, Vijay Natarajan, Dominik Schindler, Bei Wang, Yusu Wang, Gunther Weber, Kelin Xia, and Lori Ziegelmeier

Given $f: \mathcal{M} \rightarrow \mathbb{R}$, the Morse complexes partition \mathcal{M} into ascending/descending manifolds of minima/maxima. The Morse-Smale complex is the intersection of the two Morse complexes. Given f_1, f_2 , and their Morse-Smale complexes MSC_1, MSC_2 , how to define distances or metrics $d(MSC_1, MSC_2)$? This working group met on two days and focused on brainstorming likely lists of ideas worth pursuing, hopefully sparking ideas for future work in the participants when tackling this difficult and surprisingly open problem.

Summary from Day 1

We first discussed how the Morse and Morse-Smale complexes are defined, and how they can differ. Note that we often require that f has some ‘niceness’ assumptions about how the ascending/descending manifolds intersect, requires transversality, etc. These assumptions are fairly common, and ensure controlled behavior such as degree constraints on the graph and genericity of the resulting curves.

We then brainstormed a list of possible “objects” we could compute distances on (i.e. which piece of the complex) and what sorts of distances we could compute on that object. All of these objects are some structure given on the Morse complex, but some retain more structure (i.e. just keeping the graph versus using information about the 2-dimensional pieces of the complex).

1. First, we discussed just keeping a 1-dimensional skeleton (specifically, the Morse Graph of separatrices), rather than the full complex. With this information, we could consider any of the following distances:
 - Graph-based distances, e.g. interleaving and edit distances. There is a wealth of these in the literature, but it is unclear if they utilize the real structure of the Morse complex.
 - Distances on geometrically-embedded graphs / metric graphs (e.g. edge lengths + function values). These are well studied, but often computationally intractable.
 - Distances based on computational geometric measures (e.g. Frechet distances): These are well studied in computational geometry, but again unclear how they match with Morse graphs.
 - Distances based on optimal transport
 - Graph / graph kernel / spectral methods

2. Using a 2D complex: This retains more information from the Morse complex, but higher dimensional comparisons are plausibly more difficult. We considered the following options for this structure:
 - Information-theoretic distances (KL divergence)
 - Partition-based distances (Rand index)
 - Interleaving distance on the Reeb space: While interleavings on Reeb graphs are more well known, the basic idea should extend up a dimension to Reeb spaces as well.
 - Optimal transport: There is preliminary work on these in the viz community, so they may be more tractable.
 - Hausdorff distance / CG metrics / Frechet: Many of these become NP-Hard on two dimensional surfaces or even terrains or polygons with holes, but they are not well-studied on Morse complexes.
 - Distance on the Hasse graph: This yields a much different graph, which perhaps would be amenable to different types of computations but still captures much of the topology and adjacency information.
3. Finally, we also mentioned a few alternatives and other objects we could use which are based on the Morse complex as well:
 - One option was the extremal graph, which is a subset of the 1-skeleton, rather than the full skeleton. This perhaps is a simpler object than the full skeleton retaining the most ‘interesting’ information, although it is not clear what would be lost.
 - Dual graph of face adjacency: Again, this retains something interesting but flips the graph to the dual, which in some cases in computational geometry will allow different operations than the primal gives and/or can have nicer properties.

After discussing options of what objects we could to study, we then considered what desirable properties of interest would exist for such metrics, both theoretical and practical. These include:

1. Stability wrt small changes in scalar field, i.e. a bound on $d(MSC_1, MSC_2)$ vs $\|f_1 - f_2\|$
2. Stability wrt topological simplification of the field
3. Metric properties (i.e. triangle inequality, symmetric, etc)
4. Universality: This is an idea from topological data analysis, which looks for the most descriptive option amongst stable metrics. There has been recent work on Reeb graphs, which perhaps may be extended to the slightly more general Morse complex.
5. Discriminativity: Again drawing from Reeb graph metrics literature, there are many times when one distance is strictly more powerful (often at the cost of complexity). This may be a useful notion in order to compare the relative power of metrics on Morse complexes as well.
6. Computability (and/or heuristics to reduce computation time)
7. Interpretability / Locality (i.e., edit distance can tell us which edits cost what, so the cost has a discrete mapping which can be considered on its own)
8. Practicality, as opposed to worst case computational complexity: Which distances are actually feasible to implement and/or approximate?

Summary from Day 2

Thursday began with a brief review, but then we focused on a couple of specific possible directions, discussing how best to proceed in computing and/or using a distance computed in that manner. We outlined several promising approaches, which we would like to propose as likely directions for developing distances.

First, we began with a discussion of what could be done when focusing on the full complex. In this case, we considered the optimal transport approach, which seems most likely to succeed in practice, although there are some interesting challenges.

- We began by discussing how to compute optimal transport metrics, in terms of optimizing the matching to connectivity while also preserving associated properties stored on vertices (position, function value), edges (edge length, edge geometry), etc.
- We then determined several strategies towards extending this notion to Morse complexes. In particular, we see some complexity with extending the adjacency portion of these to a cell-based metric. The basic construction we considered most likely creates a bipartite graph of the adjacencies between cells of dimensions differing by 1. The challenge will then be managing this across all dimensions in a consistent way.

While there are significant challenges with optimal transport, it nonetheless seems a major alternative worth future study, given its success in other practical domains.

Our next portion of the discussion was based on the recent success of the study of Reeb graph metrics. While perhaps less practical, these have desirable theoretical properties, and so it seems worth investigating which might generalize to more general Morse or Morse-Smale complexes.

- Interleaving distances are well studied in topological data analysis, and in Reeb graphs have a nice combinatorial characterization via the thickening functor. In addition, interleavings are computable on general persistence modules and are fixed parameter tractable on simple classes of Reeb graphs. To the best of our group's knowledge, there is no notion of interleavings formally defined on Morse complexes, but the theoretical definitions would likely generalize.
- Edit distances are well studied on graphs and combinatorial objects, and appeal to computer scientists given their utility in other domains. On Reeb graphs, they have been generalized in an unusual way in order to prove stability and universality in quite recent work. We again are unaware of any work generalizing these edit distances to Morse complexes.
- Two more recently defined Reeb metrics are the functional distortion and contortion distances, which draw inspiration from Gromov-Hausdorff notions of metrics. One possible approach to generalize this to Morse complexes is to 'thicken' the space along the normal direction of separatrices, rather than along the function value space.

Finally, the group discussed complexity. Unfortunately, we suspect many if not all of these notions will be difficult to compute. There is perhaps hope of approximation or heuristics, but work remains even on Reeb graphs.

We concluded with a general discussion of other computational geometry and topology notions which have been used in simpler settings, such as Fréchet-based distances, local homology, and persistence distortion. Unfortunately, none seemed obvious candidates for study on Morse complexes.

Participants

- Ulrich Bauer
TU München, DE
- Talha Bin Masood
Linköping University, SE
- Ximena Fernández
Durham University, GB
- Hans Hagen
RPTU – Kaiserslautern, DE
- Teresa Heiss
IST Austria –
Klosterneuburg, AT
- Yasuaki Hiraoka
Kyoto University, JP
- Ingrid Hotz
Linköping University, SE
- Michael Kerber
TU Graz, AT
- Claudia Landi
University of Modena, IT
- Michael Lesnick
University at Albany, US
- Joshua A. Levine
University of Arizona –
Tucson, US
- Facundo Memoli
Ohio State University –
Columbus, US
- Dmitriy Morozov
Lawrence Berkeley National
Laboratory, US
- Vijay Natarajan
Indian Institute of Science –
Bangalore, IN
- Andreas Ott
KIT – Karlsruher Institut für
Technologie, DE
- Manish Saggat
Stanford University, US
- Dominik Schindler
Imperial College London, GB
- Primoz Skraba
Queen Mary University of
London, GB & Jožef Stefan
Institute – Ljubljana, SI
- Nico Stucki
TU München, DE
- Yusu Wang
University of California,
San Diego – La Jolla, US
- Bei Wang Phillips
University of Utah – Salt Lake
City, US
- Gunther Weber
Lawrence Berkeley National
Laboratory, US
- Erin Moriarty Wolf Chambers
St. Louis University, US
- Kelin Xia
Nanyang TU – Singapore, SG
- Lori Ziegelmeier
Macalester College –
St. Paul, US



Regular Transformations

Rajeev Alur^{*1}, Mikołaj Bojańczyk^{*2}, Emmanuel Filiot^{*3},
Anca Muscholl^{*4}, and Sarah Winter^{†5}

1 University of Pennsylvania – Philadelphia, US. alur@cis.upenn.edu

2 University of Warsaw, PL. bojan@mimuw.edu.pl

3 UL – Brussels, BE. efiliot@gmail.com

4 University of Bordeaux, FR. anca@labri.fr

5 UL – Brussels, BE. Sarah.Winter@ulb.be

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 23202 “Regular Transformations”. The goal of this seminar was to advance on a list of topics about transducers that have gathered much interest recently, and to explore new connections between the theory of regular transformations and its applications in linguistics.

Seminar May 14–17, 2023 – <https://www.dagstuhl.de/23202>

2012 ACM Subject Classification Theory of computation → Formal languages and automata theory

Keywords and phrases transducers; (poly-)regular functions; linguistic transformations

Digital Object Identifier 10.4230/DagRep.13.5.96

1 Executive Summary

Rajeev Alur

Mikołaj Bojańczyk

Emmanuel Filiot

Anca Muscholl

License  Creative Commons BY 4.0 International license
© Rajeev Alur, Mikołaj Bojańczyk, Emmanuel Filiot, and Anca Muscholl

Transducers, i.e. automata with outputs, are one of the oldest computational models in theoretical computer science. They even predate the usual boolean automata, going back to Church, Shannon, Moore, and Mealy. In spite of being considered too complex¹, transducers remained an active research topic ever since. Also connections to practical applications in efficient processing of streaming data have been established recently. The purpose of this Dagstuhl Seminar was to gather researchers to discuss recent developments on transducers and their applications.

The goal was twofold, to advance on a list of topics about transducers that have gathered much interest recently, and to explore new connections with researchers from linguistics. We enjoyed very interesting talks about:

- polyregular functions over trees and growth of regular functions
- data transducer synthesis and transducers over data words
- decomposition of finite-valued streaming string transducers

* Editor / Organizer

† Editorial Assistant / Collector

¹ Rabin and Scott argued in their 1959 paper that they are better off by “doing away with a complicated output function and having our machines simply give ‘yes’ or ‘no’ answers.”



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Regular Transformations, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 96–113

Editors: Rajeev Alur, Mikołaj Bojańczyk, Emmanuel Filiot, Anca Muscholl, and Sarah Winter



DAGSTUHL
REPORTS

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

- automata over series-parallel graphs and transducers for data management
- learning linguistic transformations and large language models as transducers
- tree transducers: class characterisations, macro tree transducers with semantic constraints
- transducers and complexity

The scientific programme was quite dense, given that we had only 3 days and almost all participants proposed to give a talk. Exchanges were very lively, and we are confident that new research directions and new collaborations emerged from this seminar.

Rajeev Alur, Mikołaj Bojańczyk, Emmanuel Filiot, Anca Muscholl

2 Table of Contents

Executive Summary

Rajeev Alur, Mikołaj Bojańczyk, Emmanuel Filiot, and Anca Muscholl 96

Overview of Talks

Notions of Symmetry in Transducers and Games <i>Shaul Almagor</i>	100
Regular languages of series-parallel graphs <i>Rajeev Alur</i>	100
Polyregular functions on unordered trees of bounded height <i>Mikołaj Bojańczyk</i>	101
From Finite-Valued Nondeterministic Transducers to Deterministic Two-Tape Automata <i>Elisabet Burjons</i>	101
A Circuit Complexity Approach to Transductions <i>Michaël Cadilhac</i>	102
Determinization of transducers over real numbers <i>Olivier Carton</i>	102
Recurrent Neural Network LMs as weighted formal language recognizers <i>Ryan Cotterell</i>	102
From Regular Transducer Expressions to Reversible Transducers <i>Luc Dartois</i>	103
A Generic Solution to Register-bounded Synthesis for Systems over Data Words <i>Léo Exibard</i>	103
Regular functions of infinite words <i>Emmanuel Filiot</i>	104
Macro Tree Transducers with semantic constraints <i>Paul Gallot</i>	104
Solving String Constraints with References using Streaming String Transducers <i>Matthew Hague</i>	105
Regular Transformations in Linguistics <i>Jeffrey Heinz</i>	105
The Decomposition Theorem for streaming string transducers <i>Ismaël Jecker</i>	106
Towards regular functions with exponential growth <i>Nathan Lhote</i>	106
On finite-valuedness of streaming string transducers <i>Anca Muscholl</i>	106
Higher-order phenomena in transducer theory <i>Lê Thành Dũng Nguyễn</i>	107
Learning (sub)regular transformations <i>Jon Raski</i>	108

A transducer model for streaming enumeration problems <i>Cristian Riveros</i>	108
Definability Results for Tree Transducers <i>Sebastian Maneth</i>	109
When is a Bottom-up Deterministic Tree Transducer Top-down Deterministic? <i>Helmut Seidl</i>	110
An Algebraic Theory for Single-use Transducers Over Data Words <i>Rafal Stefanski</i>	110
How to decide Functionality of Compositions of Top-Down Tree Transducers <i>Martin Vu</i>	111
A Regular and Complete Notion of Delay for Streaming String Transducers <i>Sarah Winter</i>	111
Open problems	
Unambiguous Single-use Register Automata <i>Rafal Stefanski</i>	112
Participants	113

3 Overview of Talks

3.1 Notions of Symmetry in Transducers and Games

Shaull Almagor (Technion – Haifa, IL)

License © Creative Commons BY 4.0 International license
© Shaull Almagor

Joint work of Shaull Almagor, Antonio Abu-Nassar, Shai Guendelman

Main reference Antonio Abu Nassar, Shaull Almagor: “Simulation by Rounds of Letter-To-Letter Transducers”, in Proc. of the 30th EACSL Annual Conference on Computer Science Logic, CSL 2022, February 14-19, 2022, Göttingen, Germany (Virtual Conference), LIPIcs, Vol. 216, pp. 3:1–3:17, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <https://doi.org/10.4230/LIPIcs.CSL.2022.3>

We consider various notions of “symmetry” for transducers and for games. The main focus is on process symmetry, whereby several processes interact, but may be unaware of the identification given to them by the system. They therefore must act interchangeably, in order to satisfy some specifications.

We examine definitions of process-symmetry, and the algorithms and decision problems pertaining to them. Specifically, we define symmetry in probabilistic transducers, symmetry-by-rounds for deterministic transducers, and symmetry in games.

3.2 Regular languages of series-parallel graphs

Rajeev Alur (University of Pennsylvania – Philadelphia, US)

License © Creative Commons BY 4.0 International license
© Rajeev Alur

Joint work of Rajeev Alur, Caleb Stanford, Christopher Watson

Main reference Rajeev Alur, Caleb Stanford, Christopher Watson: “A Robust Theory of Series Parallel Graphs”, Proc. ACM Program. Lang., Vol. 7(POPL), pp. 1058–1088, 2023.

URL <https://doi.org/10.1145/3571230>

Motivated by distributed data processing applications, we introduce a class of labeled directed acyclic graphs constructed using sequential and parallel composition operations, and study automata and logics over them. We show that deterministic and non-deterministic acceptors over such graphs have the same expressive power, which can be equivalently characterized by Monadic Second-Order logic and the graded mu-calculus. We establish closure under composition operations and decision procedures for membership, emptiness, and inclusion. A key feature of our graphs, called “synchronized series-parallel graphs” (SSPG), is that parallel composition introduces a synchronization edge from the newly introduced source vertex to the sink. The transfer of information enabled by such edges is crucial to the determinization construction, which would not be possible for the traditional definition of series-parallel graphs.

SSPGs allow both ordered ranked parallelism and unordered unranked parallelism. The latter feature means that in the corresponding automata, the transition function needs to account for an arbitrary number of predecessors by counting each type of state only up to a specified constant, thus leading to a notion of “counting complexity” that is distinct from the classical notion of state complexity. The determinization construction translates a nondeterministic automaton with n states and k counting complexity to a deterministic automaton with 2^{n^2} states and kn counting complexity, and both these bounds are shown to be tight. Furthermore, for nondeterministic automata a bound of 2 on counting complexity suffices without loss of expressiveness.

3.3 Polyregular functions on unordered trees of bounded height

Mikołaj Bojańczyk (University of Warsaw, PL)

License © Creative Commons BY 4.0 International license
© Mikołaj Bojańczyk

Joint work of Mikołaj Bojańczyk, Bartek Klin

We consider first-order interpretations that input and output trees of bounded height. The corresponding functions have polynomial output size, since a first-order interpretation can use a t -tuple of input nodes to represent a single output node. We prove that the equivalence problem for such functions is decidable, i.e. given two such interpretations, one can decide whether for every input tree, the two output trees are isomorphic. This result is incomparable to the open problem about deciding equivalence for polyregular string-to-string functions. We also give a calculus, based on prime functions and combinators, which derives all first-order interpretations for unordered trees of bounded height. The calculus is based on a type system, where the type constructors are products, co-products and multisets. Thanks to the results about tree-to-tree interpretations, the equivalence problem is decidable for this calculus. As an application of our results, we show that the equivalence problem is decidable for first-order interpretations between classes of graphs that have bounded tree depth. In all cases studied in this paper, first-order logic and MSO have the same expressive power, and hence all results apply also to MSO interpretations.

3.4 From Finite-Valued Nondeterministic Transducers to Deterministic Two-Tape Automata

Elisabet Burjons (York University – Toronto, CA)

License © Creative Commons BY 4.0 International license
© Elisabet Burjons

Joint work of Elisabet Burjons, Fabian Frei, Martin Raszyk


Main reference Elisabet Burjons, Fabian Frei, Martin Raszyk: “From Finite-Valued Nondeterministic Transducers to Deterministic Two-Tape Automata”, in Proc. of the 36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2021, Rome, Italy, June 29 – July 2, 2021, pp. 1–13, IEEE, 2021.

URL <https://doi.org/10.1109/LICS52264.2021.9470688>

Every nondeterministic finite transducer defines a binary relation associating each input word with all output words that the transducer can successfully produce on the given input. Finite-valued transducers are those for which there is a finite upper bound on the number of output words that the relation associates with every input word. We characterize finite-valued, functional, and unambiguous nondeterministic transducers whose relations can be verified by a deterministic two-tape automaton and show how to construct such an automaton if one exists. The necessary and sufficient property for a finite-valued transduction to be verifiable by a deterministic two-tape transducer is called bounded trailing, which is a more refined condition than bounded variation. We prove the undecidability of the criterion, i.e., it is undecidable to know whether a transduction has bounded trailing or not.

3.5 A Circuit Complexity Approach to Transductions

Michaël Cadilhac (DePaul University – Chicago, US)


License  Creative Commons BY 4.0 International license
© Michaël Cadilhac

Joint work of Michaël Cadilhac, Andreas Krebs, Michael Ludwig, Charles Paperman
Main reference Michaël Cadilhac, Andreas Krebs, Michael Ludwig, Charles Paperman: “A Circuit Complexity Approach to Transductions”, in Proc. of the Mathematical Foundations of Computer Science 2015 – 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 9234, pp. 141–153, Springer, 2015.
URL https://doi.org/10.1007/978-3-662-48057-1_11

Low circuit complexity classes and regular languages exhibit very tight interactions that shade light on their respective expressiveness. We propose to study these interactions at a functional level, by investigating the deterministic rational transductions computable by constant-depth, polysize circuits. To this end, a circuit framework of independent interest that allows variable output length is introduced. Relying on it, there is a general characterization of the set of transductions realizable by circuits. It is then decidable whether a transduction is definable in AC^0 and, assuming a well-established conjecture, the same for ACC^0 .

3.6 Determinization of transducers over real numbers

Olivier Carton (Université Paris Cité, FR)

License  Creative Commons BY 4.0 International license
© Olivier Carton

Joint work of Olivier Carton, Alexis Bès, Christian Choffrut

We consider one-way transducers that realize real number functions. We show that if the real function is continuous, the transducer can be made input-deterministic up to a change of the digit set used in the output. This is very similar to Avizienis numeration system used to perform sequences of additions.

3.7 Recurrent Neural Network LMs as weighted formal language recognizers

Ryan Cotterell (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
© Ryan Cotterell

Joint work of Ryan Cotterell, Anej Svete

Studying language models (LMs) in terms of well-understood formalisms allows us to precisely characterize their abilities and limitations. Previous work has investigated the expressive power of recurrent neural network (RNN) LMs in terms of their capacity to recognize unweighted formal languages. However, LMs do not describe unweighted formal languages—rather, they define *probability distributions* over strings. In this work, we study what classes of such probability distributions RNN LMs can represent, which allows us to make more direct statements about their capabilities. We show that simple RNNs are equivalent to a subclass of probabilistic finite-state automata, and can thus model a strict subset of probability distributions expressible by finite-state models. Furthermore, we study the space complexity of representing finite-state LMs with RNNs. We show that, to represent an

arbitrary deterministic finite-state LM with N states over an alphabet Σ , an RNN requires $\Omega(N|\Sigma|)$ neurons. These results present a first step towards characterizing the classes of distributions RNN LMs can represent and thus help us understand their capabilities and limitations.

3.8 From Regular Transducer Expressions to Reversible Transducers

Luc Dartois (University Paris-Est – Créteil, FR)

License © Creative Commons BY 4.0 International license
© Luc Dartois

Joint work of Luc Dartois, Paul Gastin, R. Govind, Shankaranarayanan Krishna

Main reference Luc Dartois, Paul Gastin, R. Govind, Shankara Narayanan Krishna: “Efficient Construction of Reversible Transducers from Regular Transducer Expressions”, CoRR, Vol. abs/2202.04340, 2022.

URL <https://arxiv.org/abs/2202.04340>

The class of regular transformation enjoy several equivalent characterizations (MSO Transductions, Deterministic two-way transducers, Streaming String Transducers). Regular Transducer Expressions (RTE) can be seen as a lift of Rational Expressions to string functions. They offer a combinator system to describe regular transformations. In this talk, we present a way to construct, given an RTE, an equivalent operational machine, in our case a reversible, i.e. deterministic and co-deterministic, two-way transducer. The procedure is doubly exponential for the class of regular functions, and simply exponential for the rational functions.

3.9 A Generic Solution to Register-bounded Synthesis for Systems over Data Words

Léo Exibard (Gustave Eiffel University – Marne-la-Vallée, FR)

License © Creative Commons BY 4.0 International license
© Léo Exibard

Joint work of Léo Exibard, Emmanuel Filiot, Ayrat Khalimov

Main reference Léo Exibard, Emmanuel Filiot, Ayrat Khalimov: “A Generic Solution to Register-bounded Synthesis with an Application to Discrete Orders”, CoRR, Vol. abs/2205.01952, 2022.

URL <https://doi.org/10.48550/arXiv.2205.01952>

In this talk, we consider synthesis of reactive systems interacting with environments using an infinite data domain. A popular formalism for specifying and modelling those systems is register automata and transducers. They extend finite-state automata by adding registers to store data values and to compare the incoming data values against stored ones. Synthesis from nondeterministic or universal register automata is undecidable in general. However, its register-bounded variant, where additionally a bound on the number of registers in a sought transducer is given, is known to be decidable for universal register automata which can compare data for equality, i.e., for data domain $(\mathbb{N}, =)$.

After briefly reviewing this result, we extend it to the domain $(\mathbb{N}, <)$ of natural numbers with linear order. Our solution is generic: we define a sufficient condition on data domains (regular approximability) for decidability of register-bounded synthesis. It allows one to use simple language-theoretic arguments and avoid technical game-theoretic reasoning. Further, by defining a generic notion of reducibility between data domains, we show the decidability of synthesis in the domain $(\mathbb{N}^d, <^d)$ of tuples of numbers equipped with the component-wise partial order and in the domain $(\Sigma^*, <)$ of finite strings with the prefix relation.

3.10 Regular functions of infinite words

Emmanuel Filiot (UL – Brussels, BE)

License © Creative Commons BY 4.0 International license
© Emmanuel Filiot

Joint work of Rajeev Alur, Olivier Carton, Gaëtan Douéneau-Tabot, Emmanuel Filiot, Sarah Winter
Main reference Olivier Carton, Gaëtan Douéneau-Tabot, Emmanuel Filiot, Sarah Winter: “Deterministic Regular Functions of Infinite Words”, in Proc. of the 50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10-14, 2023, Paderborn, Germany, LIPIcs, Vol. 261, pp. 121:1–121:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.
URL <https://doi.org/10.4230/LIPIcs.ICALP.2023.121>

Regular functions of infinite words are (partial) functions realized by deterministic two-way transducers with infinite look-ahead. Equivalently, Alur et. al. have shown that they correspond to functions realized by deterministic Muller streaming string transducers, and to functions defined by MSO-transductions [1]. Regular functions are however not computable in general (for a classical extension of Turing computability to infinite inputs), and we consider in this talk the class of deterministic regular functions of infinite words, realized by deterministic two-way transducers without look-ahead. We prove in [2] that it is a well-behaved class of functions: they are computable, closed under composition, characterized by the guarded fragment of MSO-transductions, by deterministic Büchi streaming string transducers, by deterministic two-way transducers with finite look-ahead, and by finite compositions of sequential functions and one fixed basic function called map-copy-reverse.

References

- 1 Rajeev Alur, Emmanuel Filiot, Ashutosh Trivedi. *Regular transformations of infinite strings*. LICS 2012: 65-74.
- 2 Olivier Carton, Gaëtan Douéneau-Tabot, Emmanuel Filiot, Sarah Winter. *Deterministic regular functions of infinite words*. ICALP 2023: 121:1-121:18.

3.11 Macro Tree Transducers with semantic constraints

Paul Gallot (Universität Bremen, DE)

License © Creative Commons BY 4.0 International license
© Paul Gallot

Joint work of Charles Peyrat, Paul Gallot, Sebastian Maneth

Macro Tree Transducers (MTT) are a very expressive model of tree transducers, but its restriction to tree-to-tree functions of Linear Size Increase (LSI) has been shown to be as expressive as MSO tree transductions. In this talk we study variations of this LSI semantic constraint, namely Linear Height Increase (LHI) and Linear input Size to output Height Increase (LSHI), and provide a new normal form for MTTs called Depth-normal form. This new normal form allows an elegant characterization of the LHI and LSHI constraints, which in turn gives an algorithm for deciding if a MTT respects these constraints.

3.12 Solving String Constraints with References using Streaming String Transducers

Matthew Hague (Royal Holloway, University of London, GB)

License © Creative Commons BY 4.0 International license

© Matthew Hague

Main reference Taolue Chen, Alejandro Flores-Lamas, Matthew Hague, Zhilei Han, Denghang Hu, Shuanglong Kan, Anthony W. Lin, Philipp Rümmer, Zhilin Wu: “Solving string constraints with Regex-dependent functions through transducers with priorities and variables”, Proc. ACM Program. Lang., Vol. 6(POPL), pp. 1–31, 2022.

URL <https://doi.org/10.1145/3498707>

String constraints arise naturally when analysing string-manipulating programs. For example, during symbolic execution, determining whether an identified program path is feasible means checking if a set of constraints (from program branches, assignments, and loop conditions) has some satisfying assignment. That, there is some input that will lead to the given path being executed. This then requires constraint satisfaction tools to be able to handle complex string operations, as well as basics such as concatenation and containment in a regular language.

One such operation is the “replaceall” operation, where a matched pattern is replaced in one string to obtain another. The replacement may itself use “references” to the matched text. For example, an author list in the format “first-name last-name” may be converted to “last-name first-name”.

Streaming String transducers provide an elegant model for such string operations. Moreover, they have an important property that the pre-image of a regular language under a replaceall remains regular. Since we additionally need to ensure that matches are “left-most longest”, we introduce “prioritised streaming string transducers” along the lines of Berglund and van der Merwe. Using these techniques, we extend the string constraint solver OSTRICH to support this replaceall operation.

In this talk, we discuss the application of streaming string transducers in constraint satisfaction, and briefly summarise some theoretical and experimental results.

3.13 Regular Transformations in Linguistics

Jeffrey Heinz (Stony Brook University, US)

License © Creative Commons BY 4.0 International license

© Jeffrey Heinz

Joint work of Jeffrey Heinz, Dakotah Lambert

Main reference Dakotah Lambert: “Unifying Classification Schemes for Languages and Processes with Attention to Locality and Relativizations Thereof”. Ph.D. thesis, Stony Brook University. 2022.

URL <https://vvulpes0.github.io/PDF/dissertation.pdf>

Recent work in theoretical computational linguistics highlights the importance of regular transformations over strings and trees in characterizing linguistic generalizations. This talk reviews some of the main results of this research program. Regular transformations provide an upper bound on morpho-phonological processes in natural languages. Particular subclasses appear to be sufficient, and these subclasses are generally (much) less than FO(<). Additionally, these subclasses are learnable with relatively low time and data complexity in ways the full class of regular transformations is not.

3.14 The Decomposition Theorem for streaming string transducers

Ismaël Jecker (University of Warsaw, PL)

License  Creative Commons BY 4.0 International license
© Ismaël Jecker

Joint work of Ismaël Jecker, Emmanuel Filiot, Christof Löding, Anca Muscholl, Gabriele Puppis, Sarah Winter

The recently introduced notion of delay for streaming string transducers (SST) has various applications, one of which is demonstrated through the Decomposition Theorem: This theorem states that every k -valued SST (where each input is mapped to at most k outputs) can be decomposed into a finite union of k single-valued SST. While similar results have been established in other settings such as finite state transducers, tree transducers and weighted automata, the specific case of SST remained open up to this day. We present a solution that relies on the notion of delay to extract, one after the other, k single-valued SST from a given k -valued SST, thus creating the desired decomposition. The Decomposition Theorem holds significant interest due to its implications: it reveals that k -valued SST recognise a robust class of functions, for instance equivalent to those recognised by finite-valued transducers. Moreover, it shows that this class enjoys good algorithmic properties. For instance, equivalence between finite-valued SST can be decided efficiently.

3.15 Towards regular functions with exponential growth

Nathan Lhote (Aix-Marseille University, FR)

License  Creative Commons BY 4.0 International license
© Nathan Lhote

Joint work of Nathan Lhote, Emmanuel Filiot, P.-A. Reynier

We present a class of transductions with exponential growth based on logical interpretations, but with monadic second-order parameters instead of first-order ones. We study some of the closure properties of this class as well as potential alternative characterizations.

3.16 On finite-valuedness of streaming string transducers

Anca Muscholl (University of Bordeaux, FR)

License  Creative Commons BY 4.0 International license
© Anca Muscholl

Joint work of Anca Muscholl, Emmanuel Filiot, Ismaël Jecker, Christof Löding, Gabriele Puppis, Sarah Winter

While one-valued streaming string transducers define a very robust class of transductions, finite-valued ones still enjoy some good properties. Recent results (see talk by Ismaël Jecker) show that they are equivalent to finite-valued two-way transducers. Moreover, jointly with G. Puppis, we have shown their equivalence problem to be decidable (ICALP 2019).

In this talk I presented a conjecture towards deciding whether a streaming string transducer is finitely valued. The characterisation conjectured reminds of the one for one-way transducers obtained by A. Weber, and exploits ideas about state invariants involved in the equivalence decision procedure for finitely-valued streaming string transducers.

3.17 Higher-order phenomena in transducer theory

Lê Thành Dũng Nguyễn (*ENS – Lyon, FR*)

License © Creative Commons BY 4.0 International license

© Lê Thành Dũng Nguyễn

Joint work of Lê Thành Dũng Nguyễn, Cécilia Pradic

This talk was about the connections between transducer theory and functional programming. An important role is played by type systems inspired from linear logic – linearity in programming language theory provides counterparts to various “copyless” or “single-use” restrictions on automata.

1. In our “implicit automata in typed λ -calculi” research programme [1, 2], Cécilia Pradic and I show that natural questions concerning the expressive power of theoretical functional programming languages turn out to have automata-theoretic answers. In particular, the study of a linear λ -calculus led us to discover a robust subclass of polyregular functions [3].
2. Meanwhile, machine models manipulating higher-order data (functions as first-class values, that may themselves take functions as arguments), taking inspiration from the higher-order grammars of the 1970s, arise from the study of composition hierarchies of transducers [4]. In this context, transducers whose memory stores linear λ -terms have been used to characterize MSO transductions of trees [5] – which is very close to one of the results of [2].

An example of phenomenon (2) is the composition of streaming string transducers (SSTs, the topic of many other talks): their registers store concatenable strings, which should be seen as order-1 functions on appendable strings, so composing n SSTs results in a machine using order- n registers.

With this point of view, the natural counterpart of SSTs on trees is a bottom-up transducer whose registers store tree *contexts*, as noted in [6, 7]. But in fact, this model is none other than the much older *macro tree transducer* (MTT) [8]! A recursive equation such as $q_1 \langle a(t, u) \rangle = b \langle q_2 \langle u \rangle, q_3 \langle t \rangle \rangle$, read as a top-down propagation of states in the old-style presentation of MTTs (= becomes \rightarrow), can also be understood as a bottom-up computation of registers (= becomes $:=$). This remark was the subject of its own lightning talk.

References

- 1 Lê Thành Dũng Nguyễn and Cécilia Pradic. Implicit automata in typed λ -calculi I: aperiodicity in a non-commutative logic. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPICs*, pages 135:1–135:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.ICALP.2020.135.
- 2 Lê Thành Dũng Nguyễn. *Implicit automata in linear logic and categorical transducer theory*. PhD thesis, Université Paris XIII (Sorbonne Paris Nord), December 2021. URL: <https://nguyentito.eu/thesis.pdf>.
- 3 Lê Thành Dũng Nguyễn, Camille Noûs, and Cécilia Pradic. Comparison-Free Polyregular Functions. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming (ICALP 2021)*, volume 198 of *Leibniz International Proceedings in Informatics (LIPICs)*, pages 139:1–139:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.ICALP.2021.139.
- 4 Joost Engelfriet and Heiko Vogler. High level tree transducers and iterated pushdown tree transducers. *Acta Informatica*, 26(1/2):131–192, 1988. doi:10.1007/BF02915449.

- 5 Paul Gallot, Aurélien Lemay, and Sylvain Salvati. Linear high-order deterministic tree transducers with regular look-ahead. In Javier Esparza and Daniel Král', editors, *45th International Symposium on Mathematical Foundations of Computer Science, MFCS 2020, August 24-28, 2020, Prague, Czech Republic*, volume 170 of *LIPICs*, pages 38:1–38:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.MFCS.2020.38.
- 6 Rajeev Alur and Loris D'Antoni. Streaming Tree Transducers. *Journal of the ACM*, 64(5):1–55, August 2017. doi:10.1145/3092842.
- 7 Mikołaj Bojańczyk and Amina Doumane. First-order tree-to-tree functions. In Holger Hermanns, Lijun Zhang, Naoki Kobayashi, and Dale Miller, editors, *LICS '20: 35th Annual ACM/IEEE Symposium on Logic in Computer Science, Saarbrücken, Germany (online conference), July 8-11, 2020*, pages 252–265. ACM, 2020. doi:10.1145/3373718.3394785.
- 8 Joost Engelfriet and Heiko Vogler. Macro tree transducers. *Journal of Computer and System Sciences*, 31(1):71–146, 1985. doi:10.1016/0022-0000(85)90066-2.

3.18 Learning (sub)regular transformations

Jon Rawski (San José State University, US)

License  Creative Commons BY 4.0 International license
© Jon Rawski

This talk highlights foundational and recent results on learning regular transformations. Restrictions on the class of grammars, combined with various ingredients in a learning paradigm, show positive results. The talk overviews positive learning results on subsequential string functions, as well as the subclass of input and output strictly local string functions. On the empirical side, using classes of regular string transformations allows us to study the inference capabilities of black box modes such as neural networks. The talk showcases several negative results in this area.

3.19 A transducer model for streaming enumeration problems

Cristian Riveros (PUC – Santiago de Chile, CL)

License  Creative Commons BY 4.0 International license
© Cristian Riveros

Joint work of Cristian Riveros, Antoine Amarilli, Marco Bucci, Alejandro Grez, Louis Jachiet, Martin Muñoz, Stijn Vansumeren

In this talk, we present Annotated automata (AnnA), a non-deterministic transducer model that outputs a subsequence of increasing positions annotated from a finite alphabet. We show that this model is suitable for encoding computational models for complex event recognition (CER) and information extraction. We show that AnnA allows for streaming enumeration algorithms, which implies the evaluation problems in the previously mentioned settings. Further, one can extend the annotation approach to visibly pushdown automata, context-free grammars, or compressed words, with similar efficient algorithmic results. Finally, we present some implementations of AnnA and the algorithms in the context of CER and information extraction.

References

- 1 Martin Muñoz, Cristian Riveros. *Streaming Enumeration on Nested Documents*. ICDT 2022.
- 2 Antoine Amarilli, Louis Jachiet, Martin Muñoz, Cristian Riveros. *Efficient Enumeration for Annotated Grammars*. PODS 2022.
- 3 Martin Muñoz, Cristian Riveros. *Constant-Delay Enumeration for SLP-Compressed Documents*. ICDT 2023.
- 4 Marco Bucci, Alejandro Grez, Andrés Quintana, Cristian Riveros. *CORE: a COmplex event Recognition Engine*. VLDB 2022.

3.20 Definability Results for Tree Transducers

Sebastian Maneth

License © Creative Commons BY 4.0 International license
© Sebastian Maneth

Joint work of Sebastian Maneth, Helmut Seidl, Martin Vu

We give a summary of three definability results that were established recently. First, we discuss how to decide whether a given bottom-up tree translation can be realised by a top-down tree transducer. This is somewhat similar to deciding whether a left-to-right finite state string transducer can be realised by a right-to-left one (simply by checking the “twinning” property), but turns out to be far more complex in our tree case. This result was presented with Helmut Seidl at ICALP’2020. In fact, we cannot yet prove this result in its entirety, but we can decide whether for a given “uniform-copying” top-down tree transducer with look-ahead an equivalent transducer exists that has no look-ahead (but is uniform copying). It is well-known that every bottom-up tree transducer can be realised by a top-down tree transducer with look-ahead. Next, we show that for a given top-down tree transducer with look-ahead it is decidable whether or not an equivalent linear top-down transducer (without look-ahead) exists. This result will appear in IJFCS with Helmut Seidl and Martin Vu. Lastly, we present preliminary results concerning the conjecture (of Joost Engelfriet), that a macro tree translation can be realised by an attributed transducer (with look-around) if and only if the translation has linear increase of the number of distinct output subtrees with respect to the input tree size.

References

- 1 Sebastian Maneth, Helmut Seidl, Martin Vu. *Definability Results for Top-Down Tree Transducers*. Int. J. Found. Comput. Sci. 34(2&3): 253-287 (2023).
- 2 Sebastian Maneth, Helmut Seidl. *When Is a Bottom-Up Deterministic Tree Translation Top-Down Deterministic?*. ICALP 2020: 134:1-134:18.

3.21 When is a Bottom-up Deterministic Tree Transducer Top-down Deterministic?

Helmut Seidl (TU München – Garching, DE)

License © Creative Commons BY 4.0 International license
© Helmut Seidl

Joint work of Sebastian Maneth, Helmut Seidl

Main reference Sebastian Maneth, Helmut Seidl: “When Is a Bottom-Up Deterministic Tree Translation Top-Down Deterministic?”, in Proc. of the 47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference), LIPIcs, Vol. 168, pp. 134:1–134:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.

URL <https://doi.org/10.4230/LIPIcs.ICALP.2020.134>

We consider two natural subclasses of deterministic top-down tree-to-tree transducers, namely, linear and uniform-copying transducers. For both classes we show that it is decidable whether the translation of a transducer with look-ahead can be realized by a transducer from the same class without look-ahead.

The transducers constructed in this way, may still make use of inspection, i.e., have an additional top-down deterministic tree automaton restricting the domain. We provide a second procedure which decides whether inspection can be removed and if so, constructs an equivalent transducer without inspection.

The construction relies on a precise abstract interpretation of inspection requirements and a dedicated earliest-normal form for linear as well as uniform-copying transducers which can be constructed in polynomial time. As a consequence, equivalence of these transducers can be decided in polynomial time.

Applying these results to deterministic bottom-up transducers, we obtain that it is decidable whether or not their translations can be realized by deterministic uniform-copying top-down transducers without look-ahead (but with inspection) - or without both look-ahead and inspection.

3.22 An Algebraic Theory for Single-use Transducers Over Data Words

Rafal Stefanski (University College London, GB)

License © Creative Commons BY 4.0 International license
© Rafal Stefanski

Joint work of Mikołaj Bojańczyk, Rafal Stefanski

Main reference Mikołaj Bojańczyk, Rafal Stefanski: “Single-Use Automata and Transducers for Infinite Alphabets”, in Proc. of the 47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference), LIPIcs, Vol. 168, pp. 113:1–113:14, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.

URL <https://doi.org/10.4230/LIPIcs.ICALP.2020.113>

In a recent paper (ICALP, 2020) Bojańczyk and I have shown that single-use register automata are equivalent to orbit-finite monoids. In this talk I am going to extend this result to transducers. For this purpose, I introduce a new algebraic model called local semigroups transductions, and I show that it is equivalent to single-use register Mealy machines.

3.23 How to decide Functionality of Compositions of Top-Down Tree Transducers

Martin Vu (Universität Bremen, DE)

License © Creative Commons BY 4.0 International license

© Martin Vu

Joint work of Sebastian Maneth, Helmut Seidl, Martin Vu

Main reference Sebastian Maneth, Helmut Seidl, Martin Vu: “How to Decide Functionality of Compositions of Top-Down Tree Transducers”, in Proc. of the Algebraic Informatics – 9th International Conference, CAI 2022, Virtual Event, October 27-29, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13706, pp. 175–191, Springer, 2022.

URL https://doi.org/10.1007/978-3-031-19685-0_13

Given a nondeterministic tree transducer, it is a natural question to ask whether or not it is functional. i.e., whether its induced tree transformation is a (partial) function? Ésik has shown that this problem is decidable even in the presence of look-ahead. We extend this result by showing that even for compositions of tree transducers the question of functionality is decidable. We prove this result by reducing the problem to the functionality of a single top-down tree transducer with look-ahead.

3.24 A Regular and Complete Notion of Delay for Streaming String Transducers

Sarah Winter (UL – Brussels, BE)

License © Creative Commons BY 4.0 International license

© Sarah Winter

Joint work of Emmanuel Filiot, Ismaël Jecker, Christof Löding, Sarah Winter

Main reference Emmanuel Filiot, Ismaël Jecker, Christof Löding, Sarah Winter: “A Regular and Complete Notion of Delay for Streaming String Transducers”, in Proc. of the 40th International Symposium on Theoretical Aspects of Computer Science, STACS 2023, March 7-9, 2023, Hamburg, Germany, LIPIcs, Vol. 254, pp. 32:1–32:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPIcs.STACS.2023.32>

The notion of delay between finite transducers is a core element of numerous fundamental results of transducer theory. In this talk we present a new notion of delay tailored to measure the similarity between streaming string transducers (SSTs). We show that our notion is regular: we design a finite automaton that can check whether the delay between any two SSTs executions is smaller than some given bound. Moreover, we show that our notion has good completeness properties: we prove that two SSTs are equivalent if and only if they are equivalent up to some (computable) bounded delay. Together with the regularity of our delay notion, it provides an alternative proof that SST equivalence is decidable. Finally, the definition of our delay notion is machine-independent, as it only depends on the origin semantics of SSTs. As a corollary, the completeness result also holds for equivalent machine models such as deterministic two-way transducers, or MSO transducers.

4 Open problems

4.1 Unambiguous Single-use Register Automata

Rafal Stefanski (*University College London, GB*)

License  Creative Commons BY 4.0 International license
 Rafal Stefanski

Joint work of Mikołaj Bojańczyk, Rafal Stefanski

Main reference Mikołaj Bojańczyk, Rafal Stefanski: “Single-Use Automata and Transducers for Infinite Alphabets”, in Proc. of the 47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference), LIPIcs, Vol. 168, pp. 113:1–113:14, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.

URL <https://doi.org/10.4230/LIPIcs.ICALP.2020.113>

A recent result by Bojańczyk and me (ICALP 2020) shows that deterministic one-way single-use register automata (over data words) are equivalent to deterministic two-way single-use register automata. On the other hand, it is known that nondeterministic one-way register automata are stronger than their deterministic version. In this open question, we ask if unambiguously nondeterministic one-way single-use register automata are equivalent to the deterministic one-way single-use register automata.

Participants

- Shaul Almagor
Technion – Haifa, IL
- Rajeev Alur
University of Pennsylvania –
Philadelphia, US
- Mikołaj Bojańczyk
University of Warsaw, PL
- Elisabet Burjons
York University – Toronto, CA
- Michaël Cadilhac
DePaul University – Chicago, US
- Olivier Carton
Université Paris Cité, FR
- Ryan Cotterell
ETH Zürich, CH
- Luc Dartois
University Paris-Est –
Créteil, FR
- Gaëtan Douéneau-Tabot
Université Paris Cité, FR
- Léo Exibard
Gustave Eiffel University –
Marne-la-Vallée, FR
- Emmanuel Filiot
UL – Brussels, BE
- Paul Gallot
Universität Bremen, DE
- Matthew Hague
Royal Holloway, University of
London, GB
- Jeffrey Heinz
Stony Brook University, US
- Ismaël Jecker
University of Warsaw, PL
- Sandra Kiefer
University of Oxford, GB
- Nathan Lhote
Aix-Marseille University, FR
- Sebastian Maneth
Universität Bremen, DE
- Anca Muscholl
University of Bordeaux, FR
- Le Thanh Dũng Nguyen
ENS – Lyon, FR
- Cécilia Pradic
Swansea University, GB
- Gabriele Puppis
University of Udine, IT
- Jon Rawski
San José State University, US
- Cristian Riveros
PUC – Santiago de Chile, CL
- Helmut Seidl
TU München – Garching, DE
- Rafal Stefanski
University College London, GB
- Martin Vu
Universität Bremen, DE
- Sarah Winter
UL – Brussels, BE



Scalable Data Structures

Gerth Stølting Brodal^{*1}, John Iacono^{*2}, László Kozma^{*3},
Vijaya Ramachandran^{*4}, and Justin Dallant^{†5}

- 1 Aarhus University, DK. gerth@cs.au.dk
- 2 UL – Brussels, BE. john.iacono@ulb.be
- 3 FU Berlin, DE. laszlo.kozma@fu-berlin.de
- 4 University of Texas – Austin, US. vlr@cs.utexas.edu
- 5 UL – Brussels, BE. Justin.Dallant@ulb.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23211 “Scalable Data Structures”. Data structures enable the organization, storage and retrieval of data across a variety of applications. As they are deployed at unprecedented scales, data structures can profoundly affect the efficiency of almost all computational tasks. The study of data structures thus continues to be an important and active area of research with an interplay between data structure design and analysis, developments in computer hardware, and the needs of modern applications. The extended abstracts included in this report give a snapshot of the current state of research on scalable data structures and establish directions for future developments in the field.

Seminar May 21–26, 2023 – <https://www.dagstuhl.de/23211>

2012 ACM Subject Classification Theory of computation → Data structures design and analysis; Theory of computation → Design and analysis of algorithms; Theory of computation → Parallel algorithms

Keywords and phrases algorithms, big data, computational models, data structures, GPU computing, parallel computation

Digital Object Identifier 10.4230/DagRep.13.5.114

1 Executive summary

Gerth Stølting Brodal (Aarhus University, DK)

John Iacono (UL – Brussels, BE)

László Kozma (FU Berlin, DE)

Vijaya Ramachandran (University of Texas – Austin, US)

License  Creative Commons BY 4.0 International license
© Gerth Stølting Brodal, John Iacono, László Kozma, Vijaya Ramachandran

About the seminar

The scale at which data is generated and processed is increasing unabated and novel applications continue to arise, posing new challenges for data structure design and analysis. The performance of data structures can dramatically affect the overall efficiency of computing systems, motivating research on scalable data structures along the entire spectrum from purely theoretical to purely empirical.

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Scalable Data Structures, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 114–135

Editors: Gerth Stølting Brodal, John Iacono, László Kozma, Vijaya Ramachandran, and Justin Dallant



Dagstuhl Reports

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

The focus of data structure research has continuously shifted to better align with the changing realities of the underlying hardware (e.g. by refining computational models to capture memory hierarchies and parallelism), and the requirements of applications (e.g. by finding the right input models, novel modes of operation, and special requirements such as data privacy or adapting to side-information). Suitable data structuring abstractions have often been crucial components of algorithmic breakthroughs, for instance in static or dynamic graph algorithms, e.g. for maximum flow or minimum spanning trees. Research on classical problems and long-standing open questions continues, with surprising recent improvements, e.g. for list labeling.

Data structure research has been a core part of computer science from the beginnings, and the field appears as vibrant as ever, with research continuing on deep old questions, as well as on new directions reflecting the changing realities of the computational landscape.

This seminar was the 15th in a series of loosely related Dagstuhl Seminars on data structures, bringing together researchers from several research directions to illuminate various aspects and solutions to the problem of data structure scalability. Following the previous, purely virtual meeting, the seminar was organized as a fully on-site event.

Topics

The presentations covered both advances in classic data structure fields, as well as insights that addressed the scalability of computing in different models of computation and across a diverse range of applications.

Parallelism was an important theme of the seminar. Blelloch (Section 4.1) discussed possible ways of making parallelism a core part of a computer science curriculum, Agrawal (Section 4.12) talked about incorporating data structures in parallel algorithms, and Sun (Section 4.5) presented algorithms for Longest Increasing Subsequence, building on parallel data structures.

Classic questions of data structure design and analysis in the comparison- or pointer-based models were the topic of multiple talks. Tarjan (Section 4.17) discussed various self-adjusting heaps and new results on their amortized efficiency. Munro (Section 4.24) talked about the ordered majority problem. Sorting was the subject of a series of talks: Wild (Section 4.3) talked about new, practically efficient sorting algorithms used in Python. Nebel (Section 4.6) talked about the practical efficiency of Lomuto's QuickSort variant. Pettie (Section 4.21) studied efficiently sorting inputs with pattern-avoiding properties, and Jacob (Section 4.27) studied variants of the sorting problem with priced comparisons.

Several talks reflected the central importance of hashing in scalable data structures, giving a broad picture of modern developments in this area. Conway (Section 4.2) presented Iceberg Hashing and Mosaic Pages. Sanders (Section 4.7) presented Sliding Block Hashing. Farach-Colton (Section 4.9) considered a simplified hash table design with strong guarantees. Bercea (Section 4.18) presented a data structure for incremental stable perfect hashing, Johnson (Section 4.22) presented Maplets, and Even (Section 4.29) talked about dynamic filters with one memory access.

The very recent $O(\log^{3/2} n)$ -time online list labeling result was presented by Wein (Section 4.28) and results for the online list labeling problem with machine learning advice were presented by Singh (Section 4.8).

Static and dynamic graph algorithms were the topic of several presentations. Kyng (Section 4.12) talked about dynamic spanners and data structuring problems arising in the context of minimum cost flow. King (Section 4.13) presented algorithms for dynamic

connectivity and Rotenberg (Section 4.15) presented dynamic graph algorithms that are adaptive to sparsity of the input. For data structures in string problems, Gørtz (Section 4.20) discussed regular expression matching and Kopelowitz (Section 4.16) presented speedups of the dynamic program for the Dyck edit distance problem.

Multiple talks reflected the interplay between data structures and computational geometry: Dallant (Section 4.30) spoke about conditional lower bounds for dynamic geometric problems, Oh (Section 4.25) presented an algorithm for the planar disjoint paths problem, and Arseneva (Section 4.23) talked about morphing graph drawings, including results obtained jointly with Oh during the seminar.

Possible computational models for designing algorithms for GPUs were addressed by Sitchinava (Section 4.19) and models of in-memory processing were presented by Silvestri (Section 4.26).

Phillips (Section 4.10) presented the design and analysis of a large-scale stream monitoring system. Liu (Section 4.4) discussed the role of scalable data structures in the context of differential privacy for graph data. Xu (Section 4.11) presented a search-optimized layout for packed memory arrays.

Final Thoughts

The organizers would like to thank the Dagstuhl team for their continuous support and also thank all participants for their contributions to this seminar. Following the earlier virtual seminar, the current (15th) seminar was fully on-site. The opportunity to personally meet and interact was highly appreciated by the community, as reflected by a very strong response to the first round of invitations and subsequent positive feedback. The seminar fills a unique need in bringing together data-structures-researchers from around the world and facilitating collaboration and exchange of ideas between them.

Earlier seminars in the series had few female participants. An important focus of the previous and the current seminar was to significantly increase female attendance. In the current seminar, 48% of the invited participants were female, resulting in a 37% female attendance. Another important focus of the seminar is to encourage the interaction between senior and early career researchers, the latter comprising 27% of the invited participants and 32% of the eventual attendees.

In the post-seminar survey the diversity of junior/senior and female/male participants were both appreciated, respondents also drawing attention to the importance of tactful and clear communication on these matters. The survey respondents also praised the coverage of a diverse range of research topics, as well as the mix between theoreticians and more applied researchers.

2 Table of Contents

Executive summary

Gerth Stølting Brodal, John Iacono, László Kozma, Vijaya Ramachandran 114

Seminar program 119

Overview of Talks 121

Should We Teach Parallelism throughout Undergraduate Algorithm Courses?
Guy E. Blelloch 121

Iceberg Hashing and Mosaic Pages
Alexander Conway 121

Quicksort, Timsort, Powersort – Algorithmic ideas, engineering tricks, and trivia
behind CPython’s new sorting algorithm
Sebastian Wild 122

Scalable Data Structures for Privacy on Graphs
Quanquan C. Liu 122

Parallel Longest Increasing Subsequence and Van Emde Boas Trees
Yihan Sun 123

Lomuto’s comeback or the unpredictability of program efficiency
Markus E. Nebel 123

Sliding Block Hashing (Slick)
Peter Sanders 124

Online List Labeling with Predictions
Shikha Singh 124

Simple Hash Tables
Martin Farach-Colton 125

Write-Optimized Algorithms for Stream Monitoring
Cynthia A. Phillips 125

Optimizing Search Layouts in Packed Memory Arrays
Helen Xu 125

Using Data Structures within Parallel Algorithms
Kunal Agrawal 126

Dynamic Spanners
Rasmus Kyng 126

Batch Parallel fast Worst Case Dynamic Connectivity
Valerie King 126

Sparsity-adaptive dynamic graph algorithms
Eva Rotenberg 126

The k-Dyck Edit Distance Problem
Tsvi Kopelowitz 127

Heaps
Robert E. Tarjan 128

An Extendable Data Structure for Incremental Stable Perfect Hashing <i>Ioana Oriana Bercea</i>	128
How to design algorithms for GPUs <i>Nodari Sitchinava</i>	128
Sparse Regular Expression Matching <i>Inge Li Gørtz</i>	129
Sorting Pattern-avoiding Permutations and Forbidden 0-1 Matrices <i>Seth Pettie</i>	129
Maplets and their Application <i>Rob Johnson</i>	130
Morphing planar drawings of graphs: main results, morphing queries and a progress report <i>Elena Arseneva</i>	130
The Ordered Majority Problem <i>Ian Munro</i>	131
Parameterized algorithm for the planar disjoint paths problem <i>Eunjin Oh</i>	131
Algorithms for Processing-In-Memory <i>Francesco Silvestri</i>	132
Sorting with Priced Comparisons: The General, the Bichromatic, and the Universal <i>Riko Jacob</i>	132
Online List Labeling: Breaking the $\log^2 n$ Barrier <i>Nicole Wein</i>	133
Dynamic Filter and Retrieval with One Memory Access <i>Guy Even</i>	133
Conditional Lower Bounds for Dynamic Geometric Measure Problems <i>Justin Dallant</i>	134
Participants	135

3 Seminar program

Sunday May 21, 2023

18:00 *Dinner buffet*

Monday May 22, 2023

07:30 *Breakfast*

09:00 *Opening & Introductions*

10:30 *Coffee break*

11:00 *Should We Teach Parallelism throughout Undergraduate Algorithm Courses?*

Guy E. Blelloch

12:15 *Lunch*

15:30 *Coffee & Cake*

16:00 *Iceberg Hashing and Mosaic Pages: A Data-Structural Approach to Faster Virtual Address Translation*

Alexander Conway

16:35 *Quicksort, Timsort, Powersort – Python’s new Sorting Algorithm*

Sebastian Wild

17:10 *Scalable Data Structures for Privacy on Graphs*

Quanquan C. Liu

18:00 *Dinner*

Tuesday May 23, 2023

07:30 *Breakfast*

09:00 *Parallel Longest Increasing Subsequence and van Emde Boas Trees*

Yihan Sun

09:30 *Lomuto’s Comeback or the Unpredictability of Program Efficiency*

Markus E. Nebel

10:00 *Sliding Block Hashing*

Peter Sanders

10:30 *Coffee break*

11:00 *Open problem session*

12:15 *Lunch*

15:30 *Coffee & Cake*

16:00 *Online List Labeling with Predictions*

Shikha Singh

16:24 *Simple Hash Tables*

Martin Farach-Colton

16:48 *Write-Optimized Algorithms for Stream Monitoring*

Cynthia A. Phillips

17:12 *Optimizing Search Layouts in Packed Memory Arrays*

Helen Xu

17:38 *Using Data Structures within Parallel Algorithms*

Kunal Agrawal

18:00 *Dinner*

Wednesday May 24, 2023

- 07:30 *Breakfast*
- 09:00 *Dynamic Spanners*
Rasmus Kyng
- 09:30 *Batch Parallel fast Worst Case Dynamic Connectivity*
Valerie King
- 10:00 *Sparsity-Adaptive Dynamic Graph Algorithms*
Eva Rotenberg
- 10:30 *Coffee break*
- 11:00 *The k -Dyck Edit Distance Problem*
Tsvi Kopelowitz
- 11:30 *Heaps*
Robert Tarjan
- 12:00 *Group picture*
- 12:15 *Lunch*
- 14:00 *Hike*
- 15:30 *Coffee & Cake*
- 18:00 *Dinner*

Thursday May 25, 2023

- 07:30 *Breakfast*
- 09:00 *An Extendable Data Structure for Incremental Stable Perfect Hashing*
Ioana-Oriana Bercea
- 09:30 *How to design algorithms for GPUs*
Nodari Sitchinava
- 10:00 *Sparse Regular Expressions*
Inge Li Gørtz
- 10:30 *Coffee break*
- 11:00 *Sorting pattern-avoiding permutations and forbidden 0-1 matrices*
Seth Pettie
- 11:30 *Maplets and their Application*
Rob Johnson
- 12:15 *Lunch*
- 15:30 *Coffee & Cake*
- 16:00 *Morphing Graph Drawings*
Elena Arseneva
- 16:30 *Ordered Majority*
Ian Munro
- 17:00 *Parameterized algorithm for the planar disjoint paths problem*
Eunjin Oh
- 17:30 *Algorithms for Processing in Memory*
Francesco Silvestri
- 18:00 *Dinner*

Friday May 26, 202307:30 *Breakfast*09:00 *Sorting with Priced Comparisons: The General, the Bichromatic, and the Universal*
Riko Jacob09:30 *Online List Labeling: Breaking the $\log^2 n$ Barrier*
Nicole Wein10:00 *Dynamic Filters and Retrieval with one Memory Access*
Guy Even10:30 *Conditional Lower Bounds for Dynamic Geometric Problems*
Justin Dallant11:00 *Coffee break*12:15 *Lunch***4 Overview of Talks****4.1 Should We Teach Parallelism throughout Undergraduate Algorithm Courses?***Guy E. Blelloch (Carnegie Mellon University – Pittsburgh, US)***License** © Creative Commons BY 4.0 International license
© Guy E. Blelloch**4.2 Iceberg Hashing and Mosaic Pages***Alexander Conway (VMware Research – Palo Alto, US)***License** © Creative Commons BY 4.0 International license
© Alexander Conway**Joint work of** Alexander Conway, Krishnan Gosakan, Jaehyun Han, William Kuszmaul, Ibrahim N. Mubarek, Nirjhar Mukherjee, Karthik Sriram, Guido Tagliavini, Evan West, Michael A. Bender, Abhishek Bhattacharjee, Martin Farach-Colton, Jayneel Gandhi, Rob Johnson, Sudarsun Kannan, Donald E. Porter**Main reference** Krishnan Gosakan, Jaehyun Han, William Kuszmaul, Ibrahim N. Mubarek, Nirjhar Mukherjee, Karthik Sriram, Guido Tagliavini, Evan West, Michael A. Bender, Abhishek Bhattacharjee, Alex Conway, Martin Farach-Colton, Jayneel Gandhi, Rob Johnson, Sudarsun Kannan, Donald E. Porter: “Mosaic Pages: Big TLB Reach with Small Pages”, in Proc. of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2023, Vancouver, BC, Canada, March 25-29, 2023, pp. 433–448, ACM, 2023.**URL** <https://doi.org/10.1145/3582016.3582021>

In this talk, I present an algorithmic approach to co-designing TLB hardware and the paging mechanism to increase TLB reach without the fragmentation issues incurred by huge pages. Along the way, I’ll introduce a new hash-table design that overcomes existing tradeoffs, and achieves better performance than state-of-the-art hash tables both in theory and in practice. Key to these results are “tiny pointers,” an algorithmic technique for compressing pointers.

4.3 Quicksort, Timsort, Powersort – Algorithmic ideas, engineering tricks, and trivia behind CPython’s new sorting algorithm

Sebastian Wild (*University of Liverpool, GB*)

License © Creative Commons BY 4.0 International license

© Sebastian Wild

Main reference J. Ian Munro, Sebastian Wild: “Nearly-Optimal Mergesorts: Fast, Practical Sorting Methods That Optimally Adapt to Existing Runs”, in Proc. of the 26th Annual European Symposium on Algorithms, ESA 2018, August 20-22, 2018, Helsinki, Finland, LIPIcs, Vol. 112, pp. 63:1–63:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

URL <https://doi.org/10.4230/LIPIcs.ESA.2018.63>

Writing a sorting function is easy – coding a fast and reliable reference implementation less so. In this talk, I tell the story behind CPython’s latest updates of the list sort function.

After using Quicksort for a long while, Tim Peters invented Timsort, a clever Mergesort variant, for the CPython reference implementation of Python. Timsort is both effective in Python and a popular export product: it is used in many languages and frameworks, notably OpenJDK, the Android runtime, and the V8 JavaScript engine.

Despite this success, algorithms researchers eventually pinpointed two flaws in Timsort’s underlying algorithm: The first could lead to a stack overflow in CPython (and Java); although it has meanwhile been fixed, it is curious that 10 years of widespread use didn’t bring it to surface. The second flaw is related to performance: the order in which detected sorted segments, the “runs” in the input, are merged, can be 50% more costly than necessary. Based on ideas from the little known puzzle of optimal alphabetic trees, the Powersort merge policy finds nearly optimal merging orders with negligible overhead, and is now (Python 3.11.0) part of the CPython implementation.

References

- 1 J. Ian Munro and Sebastian Wild. *Nearly-Optimal Mergesorts: Fast, Practical Sorting Methods That Optimally Adapt to Existing Runs*. ESA 2018
- 2 William Cawley Gelling, Markus E. Nebel, Benjamin Smith, and Sebastian Wild. *Multiway Powersort*. ALENEX 2023

4.4 Scalable Data Structures for Privacy on Graphs

Quanquan C. Liu (*Northwestern University – Evanston, US*)

License © Creative Commons BY 4.0 International license

© Quanquan C. Liu

Joint work of Laxman Dhulipala, Quanquan C. Liu, Sofya Raskhodnikova, Jessica Shi, Julian Shun, Shangdi Yu
Main reference Laxman Dhulipala, Quanquan C. Liu, Sofya Raskhodnikova, Jessica Shi, Julian Shun, Shangdi Yu: “Differential Privacy from Locally Adjustable Graph Algorithms: k-Core Decomposition, Low Out-Degree Ordering, and Densest Subgraphs”, in Proc. of the 63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 – November 3, 2022, pp. 754–765, IEEE, 2022.

URL <https://doi.org/10.1109/FOCS54457.2022.00077>

Main reference Talya Eden, Quanquan C. Liu, Sofya Raskhodnikova, Adam D. Smith: “Triangle Counting with Local Edge Differential Privacy”, in Proc. of the 50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10-14, 2023, Paderborn, Germany, LIPIcs, Vol. 261, pp. 52:1–52:21, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPIcs.ICALP.2023.52>

Differential privacy is the gold standard for rigorous data privacy guarantees where the traditional central model assumes a trusted curator who takes private input and outputs privatized answers to user queries. However, one major assumption in this model is that the trusted curator always keeps the input private. Unfortunately, such a strong notion of

trust is too ideal in today’s world where massive data leaks occur. Thus, in this talk, I’ll discuss private graph algorithms in the local model, where nodes trust no one with their private information. One major issue in the local privacy model for graphs is scalability. Scalability is often an issue for graph algorithms in the local privacy model because many techniques for making graph algorithms private change the density of the input graph (i.e. making the graph much denser than it was previously). I’ll present the first local edge differentially private (LEDP) algorithms for k -core decomposition, low out-degree ordering, and densest subgraph. Furthermore, our algorithms are scalable and can be implemented in distributed and parallel models without the issue of changing graph density. Our algorithm’s approximation factor matches that of the currently best non-private distributed algorithm for k -core decomposition with only an $\text{poly}(\log n)/\epsilon$ additive error. I’ll conclude with a discussion of some open questions and potential future work.

4.5 Parallel Longest Increasing Subsequence and Van Emde Boas Trees

Yihan Sun (University of California – Riverside, US)

License © Creative Commons BY 4.0 International license
© Yihan Sun

Joint work of Yan Gu, Ziyang Men, Zheqi Shen, Yihan Sun, Zijin Wan

Main reference Yan Gu, Ziyang Men, Zheqi Shen, Yihan Sun, Zijin Wan: “Parallel Longest Increasing Subsequence and van Emde Boas Trees”, in Proc. of the 35th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA 2023, Orlando, FL, USA, June 17-19, 2023, pp. 327–340, ACM, 2023.

URL <https://doi.org/10.1145/3558481.3591069>

This paper studies parallel algorithms for the longest increasing subsequence (LIS) problem. Let n be the input size and k be the LIS length of the input. Sequentially, LIS is a simple problem that can be solved using dynamic programming (DP) in $O(n \log n)$ work. However, parallelizing LIS is a long-standing challenge. We are unaware of any parallel LIS algorithm that has optimal $O(n \log n)$ work and non-trivial parallelism (i.e., $\tilde{O}(k)$ or $o(n)$ span). This paper proposes a parallel LIS algorithm that costs $O(n \log k)$ work, $\tilde{O}(k)$ span, and $O(n)$ space, and is much simpler than the previous parallel LIS algorithms. We also generalize the algorithm to a weighted version of LIS, which maximizes the weighted sum for all objects in an increasing subsequence. To achieve a better work bound for the weighted LIS algorithm, we designed parallel algorithms for the van Emde Boas (vEB) tree, which has the same structure as the sequential vEB tree, and supports work-efficient parallel batch insertion, deletion, and range queries.

We also implemented our parallel LIS algorithms. Our implementation is light-weighted, efficient, and scalable.

4.6 Lomuto’s comeback or the unpredictability of program efficiency

Markus E. Nebel (Universität Bielefeld, DE)

License © Creative Commons BY 4.0 International license
© Markus E. Nebel

Joint work of Markus E. Nebel, David Komnick

Main reference <https://dlang.org/blog/2020/05/14/lomutos-comeback/>

We report experimental results for a Quicksort variant based on Lomuto’s partitioning suggested by Andrei Alexandrescu. This variant eliminates branches inside the main loop of the partitioning process for the price of an increased number of overall executed instructions.

However, the resulting reduction of mispredicted branches gives rise to a heavily used pipeline. As a consequence, the resulting Quicksort implementation runs faster than one using Hoare/Sedgwick partitioning. Our adaptations of the latter to different branch free versions imply a similar overhead of instructions and a comparable reduction of branch misses. However, the speedup observed due to improved pipelining is way smaller than for the Lomuto variant and results in an overall worse runtime. So far we have no explanation for this.

4.7 Sliding Block Hashing (Slick)

Peter Sanders (KIT – Karlsruhe Institut für Technologie, DE)

License © Creative Commons BY 4.0 International license
© Peter Sanders

Joint work of Hans-Peter Lehmann, Peter Sanders, Stefan Walzer

Main reference Hans-Peter Lehmann, Peter Sanders, Stefan Walzer: “Sliding Block Hashing (Slick) – Basic Algorithmic Ideas”, CoRR, Vol. abs/2304.09283, 2023.

URL <https://doi.org/10.48550/arXiv.2304.09283>

We present **Sliding Block Hashing (Slick)** a simple hash table data structure that combines high performance with very good space efficiency.

4.8 Online List Labeling with Predictions

Shikha Singh (Williams College – Williamstown, US)

License © Creative Commons BY 4.0 International license
© Shikha Singh

Joint work of Samuel McCauley, Benjamin Moseley, Aidin Niaparast, Shikha Singh

Main reference Samuel McCauley, Benjamin Moseley, Aidin Niaparast, Shikha Singh: “Online List Labeling with Predictions”, CoRR, Vol. abs/2305.10536, 2023.

URL <https://doi.org/10.48550/arXiv.2305.10536>

A growing line of work shows how learned predictions can be used to break through worst-case barriers to improve the running time of an algorithm. However, incorporating predictions into data structures with strong theoretical guarantees remains underdeveloped. This work takes a step in this direction by showing that predictions can be leveraged in the fundamental online list labeling problem. In the problem, n items arrive over time and must be stored in sorted order in an array of size $\Theta(n)$. The array slot of an element is its label and the goal is to maintain sorted order while minimizing the total number of elements moved (i.e., relabeled). We present a new list labeling data structure and bound its performance in two models. In the worst-case learning-augmented model, we give guarantees in terms of the error in the predictions. Our data structure provides strong theoretical guarantees— it is optimal for any prediction error and guarantees the best-known worst-case bound even when the predictions are entirely erroneous. We also consider a stochastic error model and bound the performance in terms of the expectation and variance of the error. Finally, the theoretical results are demonstrated empirically. In particular, we show that our data structure performs well on numerous real datasets, including temporal data sets where predictions are constructed from elements that arrived in the past (as is typically done in a practical use case).

4.9 Simple Hash Tables

Martin Farach-Colton (Rutgers University – Piscataway, US)

License © Creative Commons BY 4.0 International license
© Martin Farach-Colton

We present a hash table that achieves nearly the state of the art but requires only basic analytical techniques. The aim is to present an algorithm that can be taught to graduate students (or perhaps advanced undergrads).

4.10 Write-Optimized Algorithms for Stream Monitoring

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

License © Creative Commons BY 4.0 International license
© Cynthia A. Phillips

Joint work of Shikha Singh, Prashant Pandey, Michael Bender, Jonathan Berry, Daniel Delayo, Martin Farach-Colton, Rob Johnson, Thomas Kroeger, Cynthia Phillips, David Tench, Eric Thomas
Main reference Shikha Singh, Prashant Pandey, Michael A. Bender, Jonathan W. Berry, Martin Farach-Colton, Rob Johnson, Thomas M. Kroeger, Cynthia A. Phillips: “Timely Reporting of Heavy Hitters Using External Memory”, *ACM Trans. Database Syst.*, Vol. 46(4), pp. 14:1–14:35, 2021.
URL <https://doi.org/10.1145/3472392>

We describe data structures and data-management algorithms for monitoring cyber streams. We wish to identify specific patterns that arrive slowly over time, hidden among high-speed streams of normal traffic. The key piece of the Firehose benchmark that models this application is a variant of the heavy-hitters problem: report a key after it has been seen a specific constant number of times. To solve this without false negatives requires $\Omega(N)$ space for partial-pattern storage for a stream of size N . We give write-optimized external-memory algorithms to accurately monitor high-speed streams with provable tunable trade-off between reporting delay and I/O overhead. Our experimental results show that a multithreaded version of our algorithm has throughput comparable to an engineered in-RAM reference implementation, but our method reports all reportable keys, while the in-RAM method can miss almost all reports for sufficiently large key space. We describe extensions to unending streams.

4.11 Optimizing Search Layouts in Packed Memory Arrays

Helen Xu (Lawrence Berkeley National Laboratory, US)

License © Creative Commons BY 4.0 International license
© Helen Xu

Joint work of Brian Wheatman, Randal C. Burns, Aydin Buluç, Helen Xu
Main reference Brian Wheatman, Randal C. Burns, Aydin Buluç, Helen Xu: “Optimizing Search Layouts in Packed Memory Arrays”, in *Proc. of the Symposium on Algorithm Engineering and Experiments, ALENEX 2023, Florence, Italy, January 22-23, 2023*, pp. 148–161, SIAM, 2023.
URL <https://doi.org/10.1137/1.9781611977561.ch13>

This talk covers Search-optimized Packed Memory Arrays (SPMAs), a collection of data structures based on Packed Memory Arrays (PMAs) that address suboptimal search via cache-optimized search layouts. Traditionally, PMAs and B-trees have tradeoffs between searches/inserts and scans: B-trees were faster for searches and inserts, while PMAs were faster for scans. Our empirical evaluation shows that SPMAs overcome this tradeoff for unsorted input distributions: on average, SPMAs are faster than B+-trees (a variant of B-trees optimized for scans) on all major operations.

4.12 Using Data Structures within Parallel Algorithms

Kunal Agrawal (Washington University – St. Louis, US)

License  Creative Commons BY 4.0 International license
 Kunal Agrawal

4.13 Dynamic Spanners

Rasmus Kyng (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
 Rasmus Kyng


Joint work of Li Chen, Yang Liu, Richard Peng, Maximilian Probst Gutenberg, Sushant Sachdeva
Main reference Li Chen, Rasmus Kyng, Yang P. Liu, Richard Peng, Maximilian Probst Gutenberg, Sushant Sachdeva: “Maximum Flow and Minimum-Cost Flow in Almost-Linear Time”, CoRR, Vol. abs/2203.00671, 2022.
URL <https://doi.org/10.48550/arXiv.2203.00671>

I gave a presentation on two topics:

1. How to use an L1 IPM to turn the task of solving an LP into a sequence data structure queries, and how this can be applied to solve minimum-cost flow problems via min-ratio cycle updates.
2. How to design dynamic spanners that allow for edge insertions and deletions and vertex splits.

4.14 Batch Parallel fast Worst Case Dynamic Connectivity

Valerie King (University of Victoria, CA)

License  Creative Commons BY 4.0 International license
 Valerie King

The dynamic connectivity problem is to process an online sequence of edge insertions and deletions in a graph while answering connectivity queries. Here we simplify and parallelize the sequential Monte Carlo algorithms of King et al. as improved by Wang for dynamic connectivity, which requires polylogarithmic time per update and per query in the worst-case. Our simplification adds no cost to the asymptotic sequential running time. It enables us to rely strictly on ET-trees, rather than more complicated path compression data structures, making it simpler to perform batch-parallel updates.

4.15 Sparsity-adaptive dynamic graph algorithms

Eva Rotenberg (Technical University of Denmark – Lyngby, DK)

License  Creative Commons BY 4.0 International license
 Eva Rotenberg

Joint work of Aleksander B. G. Christiansen, Jacob Holm, Ivor van der Hoog, Krzysztof Nowicki, Eva Rotenberg, Chris Schwiegelshohn, Carsten Thomassen

The *arboricity* α of a graph is the number of forests it takes to cover all its edges. Being asymptotically related to the graph’s degeneracy and maximal subgraph density, arboricity is considered a good measure of the sparsity of a graph. Natural computational questions about arboricity include: computing the arboricity, obtaining a decomposition of the edges into few forests, and orienting the edges so that each vertex has only close to α out-edges.

In this talk, we will address these questions in the *dynamic* setting, in which the graph is subject to arbitrary, adversarial insertions and deletions of edges. We will see how maintaining an orientation with few out-edges from each vertex leads to efficient dynamic algorithms for *matching*, *colouring*, and decomposing into $O(\alpha)$ forests; And we will see how to efficiently balance the number of out-edges in an orientation of the dynamic graph, via an almost local reconciliation between neighbouring vertices.

References

- 1 Aleksander B. G. Christiansen, Jacob Holm, Eva Rotenberg, and Carsten Thomassen. On Dynamic $\alpha + 1$ Arboricity Decomposition and Out-Orientation. In Stefan Szeider, Robert Ganian, and Alexandra Silva, editors, *47th MFCS*, volume 241 of *LIPICs*, pages 34:1–34:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.
- 2 Aleksander B. G. Christiansen, Jacob Holm, Ivor van der Hoog, Eva Rotenberg, and Chris Schwegelshohn. Adaptive Out-Orientations with Applications. *CoRR*, abs/2209.14087, 2022.
- 3 Aleksander Bjørn Grodt Christiansen, Krzysztof Nowicki, and Eva Rotenberg. Improved Dynamic Colouring of Sparse Graphs. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual STOC*, pages 1201–1214. ACM, 2023.
- 4 Aleksander B. G. Christiansen and Eva Rotenberg. Fully-Dynamic $\alpha + 2$ Arboricity Decompositions and Implicit Colouring. In Mikolaj Bojanczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th ICALP 2022*, volume 229 of *LIPICs*, pages 42:1–42:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

4.16 The k-Dyck Edit Distance Problem

Tsvi Kopelowitz (Bar-Ilan University – Ramat Gan, IL)

License © Creative Commons BY 4.0 International license

© Tsvi Kopelowitz

Joint work of Dvir Fried, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, Ely Porat, Tatiana Starikovskaya

Main reference Dvir Fried, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, Ely Porat, Tatiana Starikovskaya: “An Improved Algorithm for The k-Dyck Edit Distance Problem”, in Proc. of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 – 12, 2022, pp. 3650–3669, SIAM, 2022.

URL <https://doi.org/10.1137/1.9781611977073.144>

A Dyck sequence is a sequence of opening and closing parentheses (of various types) that is balanced. The Dyck edit distance of a given sequence of parentheses S is the smallest number of edit operations (insertions, deletions, and substitutions) needed to transform S into a Dyck sequence. We consider the threshold Dyck edit distance problem, where the input is a sequence of parentheses S and a positive integer k , and the goal is to compute the Dyck edit distance of S only if the distance is at most k , and otherwise report that the distance is larger than k . Backurs and Onak [PODS’16] showed that the threshold Dyck edit distance problem can be solved in $O(n + k^{16})$ time. In this work, we design new algorithms for the threshold Dyck edit distance problem which costs $O(n + k^{4.782036})$ time with high probability or $O(n + k^{4.853059})$ deterministically. Our algorithms combine several new structural properties of the Dyck edit distance problem, a refined algorithm for fast (min, +) matrix product, and a careful modification of ideas used in Valiant’s parsing algorithm.

4.17 Heaps

Robert E. Tarjan (Princeton University, US)

License © Creative Commons BY 4.0 International license
© Robert E. Tarjan

The heap, or priority queue data structure supports fast access to the smallest of a set of items subject to insertions, deletions, and decreases in value. We describe recent results giving tight and almost-tight amortized efficiency bounds for three self-adjusting heap implementations, the slim heap, the smooth heap, and the multipass pairing heap.

4.18 An Extendable Data Structure for Incremental Stable Perfect Hashing

Ioana Oriana Bercea (IT University of Copenhagen, DK)

Joint work of Ioana Oriana Bercea, Guy Even

License © Creative Commons BY 4.0 International license
© Ioana Oriana Bercea

Main reference Ioana Oriana Bercea, Guy Even: “An extendable data structure for incremental stable perfect hashing”, in Proc. of the STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 – 24, 2022, pp. 1298–1310, ACM, 2022.

URL <https://doi.org/10.1145/3519935.3520070>

The talk is about recent advancements in the design of dictionaries and other related data structures. A dynamic dictionary is a data structure that maintains sets under insertions and deletions and supports membership queries of the form “is an element x in the set or not?”. A related problem is that of maintaining a perfect hash function over the set, in which the data structure assigns a unique hashcode to each element in the set (but does not need to support membership queries). We specifically focus on showing a perfect hashing data structure whose space is proportional to the cardinality of the set at all points in time and whose hashcodes are stable in the incremental setting (i.e., when only insertions are allowed, the hashcode of an element does not change while the element is in the set). This is joint work with Guy Even, based on our STOC 2022 paper.

4.19 How to design algorithms for GPUs

Nodari Sitchinava (University of Hawaii at Manoa – Honolulu, US)

License © Creative Commons BY 4.0 International license
© Nodari Sitchinava

In this talk we will briefly review the models of computation used for designing algorithms on GPUs. We will also discuss recent algorithmic results on how to avoid bank conflicts when designing algorithms for GPUs.

4.20 Sparse Regular Expression Matching

Inge Li Gørtz (Technical University of Denmark – Lyngby, DK)

License © Creative Commons BY 4.0 International license
© Inge Li Gørtz

Joint work of Philip Bille, Inge Li Gørtz

Main reference Philip Bille, Inge Li Gørtz: “Sparse Regular Expression Matching”, CoRR, Vol. abs/1907.04752, 2019.

URL <http://arxiv.org/abs/1907.04752>

A regular expression specifies a set of strings formed by single characters combined with concatenation, union, and Kleene star operators. Given a regular expression R and a string Q , the regular expression matching problem is to decide if Q matches any of the strings specified by R . Regular expressions are a fundamental concept in formal languages and regular expression matching is a basic primitive for searching and processing data. A standard textbook solution [Thompson, CACM 1968] constructs and simulates a nondeterministic finite automaton, leading to an $O(nm)$ time algorithm, where n is the length of Q and m is the length of R . Despite considerable research efforts only polylogarithmic improvements of this bound are known. Recently, conditional lower bounds provided evidence for this lack of progress when Backurs and Indyk [FOCS 2016] proved that, assuming the strong exponential time hypothesis (SETH), regular expression matching cannot be solved in $O((nm)^{1-\epsilon})$, for any constant $\epsilon > 0$. Hence, the complexity of regular expression matching is essentially settled in terms of n and m .

In this paper, we take a new approach and go beyond worst-case analysis in n and m . We introduce a *density* parameter, Δ , that captures the amount of nondeterminism in the NFA simulation on Q . The density is at most $nm + 1$ but can be significantly smaller. Our main result is a new algorithm that solves regular expression matching in

$$O\left(\Delta \log \log \frac{nm}{\Delta} + n + m\right)$$

time.

This essentially replaces nm with Δ in the complexity of regular expression matching. We complement our upper bound by a matching conditional lower bound that proves that we cannot solve regular expression matching in time $O(\Delta^{1-\epsilon})$ for any constant $\epsilon > 0$ assuming SETH.

4.21 Sorting Pattern-avoiding Permutations and Forbidden 0-1 Matrices

Seth Pettie (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 4.0 International license
© Seth Pettie

Joint work of Parinya Chalermsook, Seth Pettie, Sorrachai Yingchareonthawornchai


Main reference Parinya Chalermsook, Seth Pettie, Sorrachai Yingchareonthawornchai: “Forbidden 0-1 Matrices and the Complexity of Sorting Pattern-Avoiding Permutations”. Manuscript (2023).

An n -permutation S “avoids” a k -permutation π if there are no k indices $i_1 < \dots < i_k$ such that $S(i_j) < S(i_{j'})$ iff $\pi(j) < \pi(j')$. Chalermsook et al. (2015) and Kozma and Saranurak (2019) presented two algorithms for sorting such permutations in $O(n \cdot 2^{(\alpha(n))^{3k/2}})$ time. Their upper bound was derived by transcribing the behavior of the algorithm as an $n \times n$ 0-1 matrix and proving that this matrix avoids a certain pattern, which is the Kronecker product of a permutation (encoding π) and a “hat” pattern. In this talk I prove that both of

these algorithms actually run in $O(n \cdot 2^{(1+o(1))\alpha(n)})$ time, by bounding the extremal function of all such patterns. I also show that this bound is essentially tight, and that most such patterns have an extremal function that is $\Omega(n \cdot 2^{\alpha(n)})$.

4.22 Maplets and their Application


Rob Johnson (VMware – Palo Alto, US)

License  Creative Commons BY 4.0 International license
© Rob Johnson

Filters, such as Bloom, quotient, cuckoo, xor, and ribbon filters, are space-efficient, lossy representations of sets. They have become widely used in systems and extensively researched by data structures designers. In this talk, we argue that most applications would be better served by maplets, i.e. space efficient, lossy maps, rather than filters. We explain how to generalize the filter notion of lossiness to maps, show how several classic filter applications can be dramatically improved by using maplets, and show how to construct maplets straightforwardly from many of today's filters.

4.23 Morphing planar drawings of graphs: main results, morphing queries and a progress report

Elena Arseneva (University of Lugano, CH)

License  Creative Commons BY 4.0 International license
© Elena Arseneva
Joint work of Elena Arseneva, Eunjin Oh

Given two planar (straight-line) drawings of a planar graph, can one drawing be transformed to the other in a little number of steps, where during each step every vertex moves along a straight line segment with a uniform speed? It is crucially required that there is no crossing between any elements of a drawing at any moment. Such transformation is called a morph. If the drawings are topologically equivalent, then a 2D morph is always possible in $O(n)$ steps, and sometimes a linear number of steps is necessary [1]. Allowing intermediate drawings to lie in 3D reduces the upper bound on the number of steps to $O(\log n)$ for trees [2], and lifts the topologically equivalent requirement for arbitrary graphs, however at a cost of quadratic number of steps [3]. No non-trivial lower bound for this setting is known. During the seminar I proposed a query variant of the graph morphing problem, and jointly with Eunjin Oh we obtained the first result in this direction:

A planar drawing Γ of an n -vertex graph G can be preprocessed in $O(n^{5/3} \log^2 n)$ expected time and $O(n \log n)$ space for the following queries: given a vertex v of G , and a point t in R^2 , can v be moved from its position in Γ to the point t , such that the morph is non-crossing?

After the above preprocessing, this query can be answered in $O(\log m \log^2 n)$ time, when m is the degree of vertex v . We hope to further improve this runtime and generalise our data structure to availability region queries, updates and 3D morphs.

This result is a joint work during the seminar with its participant Eunjin Oh.

References

- 1 S. Alamdari, P. Angelini, F. Barrera-Cruz, T. Chan, G. Da Lozzo, G. Di Battista, F. Frati, P. Haxell, A. Lubiw, M. Patrignani, V. Roselli How to morph planar graph drawings. *SIAM Journal on Computing*, 46(2):824-52, 2017.
- 2 E. Arseneva, P. Bose, P. Cano, A. D'Angelo, V. Dujmovic, F. Frati, S. Langerman, and A. Tappini. Pole Dancing: 3D Morphs for Tree Drawings. *Journal of Graph Algorithms and Applications (Special issue of selected papers from GD'18)*, 23(3), pages 579–602 DOI: 10.7155/jgaa.00503, 2019.
- 3 K. Buchin, W. Evans, F. Frati, I. Kostitsyna, M. Löffler, T. Ophelders, A. Wolff. Morphing planar graph drawings through 3d. In *Proc. International Conference on Current Trends in Theory and Practice of Computer Science 2023 Jan 1 (pp. 80-95)*.

4.24 The Ordered Majority Problem

Ian Munro (University of Waterloo, CA)

License © Creative Commons BY 4.0 International license
© Ian Munro

Joint work of Louisa Seelbach Benkner, Ben Baral, Siwei Yang, Ian Munro

We examine the well-known problem of determining whether some value occurs in the majority of positions in an array. The twist is that in the past comparisons have been ($=, \neq$), while here we focus on a three way outcome ($<, =, >$). We show, that like the ($=, \neq$) version, $3n/2 - o(1)$ comparisons are necessary (and sufficient) in the worst case, but give a Las Vegas algorithm requiring an expected $n + o(1)$ comparisons and $n - o(n)$ lower bound, in contrast with a $1.059\dots n$ lower bound for the ($=, \neq$) version.

4.25 Parameterized algorithm for the planar disjoint paths problem

Eunjin Oh (POSTECH – Pohang, KR)

License © Creative Commons BY 4.0 International license
© Eunjin Oh

Joint work of Kyungjin Cho, Eunjin Oh, Seunghyeok Oh


Main reference Kyungjin Cho, Eunjin Oh, Seunghyeok Oh: “Parameterized Algorithm for the Disjoint Path Problem on Planar Graphs: Exponential in k^2 and Linear in n ”, in Proc. of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023, pp. 3734–3758, SIAM, 2023.

URL <https://doi.org/10.1137/1.9781611977554.ch144>

Given a planar graph G with n vertices and a set $T = \{(s_1, t_1), \dots, (s_k, t_k)\}$ of k terminal pairs, the disjoint paths problem asks for computing a set of vertex disjoint paths P_1, \dots, P_k such that P_i connects s_i and t_i . In this talk, I will introduce a $2^{O(k^2)}n$ -time algorithm for the planar disjoint paths problem. This improves the previously best-known algorithm with running times of $2^{O(k^2)}n^6$ and $2^{2^{O(k)}}n$.

4.26 Algorithms for Processing-In-Memory

Francesco Silvestri (University of Padova, IT)


License  Creative Commons BY 4.0 International license
© Francesco Silvestri

Joint work of Lorenzo Asquini, Juan Gomez-Luna, Francesco Silvestri

Processing-In-Memory (PIM) is a hardware architecture that allows reducing the memory bottleneck: while data are sent from the memory to the CPU in the traditional memory hierarchy, in a PIM architecture the computation is sent to the memory. The main idea is to add a small computing unit within each memory module: this is an almost 40-year-old theoretical idea; however, only recently PIM architectures have been successfully implemented and commercialized (e.g., by UPMEM). In this talk, we will see a computational model for designing efficient algorithms that fully exploit PIMs, and introduce PIM algorithms for binary search and triangle counting.

4.27 Sorting with Priced Comparisons: The General, the Bichromatic, and the Universal

Riko Jacob (IT University of Copenhagen, DK)

License  Creative Commons BY 4.0 International license
© Riko Jacob

Joint work of Mayank Goswami, Riko Jacob

Main reference Mayank Goswami, Riko Jacob: “Universal Sorting: Finding a DAG using Priced Comparisons”, CoRR, Vol. abs/2211.04601, 2022.

URL <https://doi.org/10.48550/arXiv.2211.04601>

We address two open problems in sorting with priced information, introduced by [Charikar, Fagin, Guruswami, Kleinberg, Raghavan, Sahai (CFGKRS), STOC 2000]. In this setting, different comparisons have different (potentially infinite) costs. The goal is to find a sorting algorithm with small competitive ratio, defined as the (worst-case) ratio of the algorithm’s cost to the cost of the cheapest proof of the sorted order.

1) When all costs are in $\{0, 1, n, \infty\}$, we give an algorithm that has $\tilde{O}(n^{3/4})$ competitive ratio. Our result refutes the hypothesis that a widely cited $\Omega(n)$ lower bound on the competitive ratio for finding the maximum extends to sorting. This lower bound by [Gupta, Kumar, FOCS 2000] uses costs in $\{0, 1, n, \infty\}$ and was claimed as the reason why sorting with arbitrary costs seemed bleak and hopeless. Our algorithm also generalizes the algorithms for generalized sorting (all costs in $\{1, \infty\}$), a version initiated by [Huang, Kannan, Khanna, FOCS 2011] and addressed recently by [Kuszmaul, Narayanan, FOCS 2021].

2) We answer the problem of bichromatic sorting posed by [CFGKRS]: We are given two sets A and B of total size n , and the cost of an $A - A$ comparison or a $B - B$ comparison is higher than an $A - B$ comparison. The goal is to sort $A \cup B$. An $\Omega(\log n)$ lower bound on competitive ratio follows from unit-cost sorting. We give a randomized algorithm with an almost-optimal w.h.p. competitive ratio of $O(\log^3 n)$.

We also study generalizations of the problem *universal sorting* and *bipartite sorting* (a generalization of nuts-and-bolts). Here, we define a notion of *instance optimality*, and develop an algorithm for bipartite sorting which is $O(\log^3 n)$ instance-optimal. Our framework of instance optimality applies to other static problems and may be of independent interest.

4.28 Online List Labeling: Breaking the $\log^2 n$ Barrier

Nicole Wein (Rutgers University – Piscataway, US)

License © Creative Commons BY 4.0 International license
© Nicole Wein

Joint work of Michael Bender, Alexander Conway, Martin Farach-Colton, Hanna Komlos, William Kuszmaul, Nicole Wein

Main reference Michael A. Bender, Alexander Conway, Martin Farach-Colton, Hanna Komlós, William Kuszmaul, Nicole Wein: “Online List Labeling: Breaking the $\log^2 n$ Barrier”, CoRR, Vol. abs/2203.02763, 2022.

URL <https://doi.org/10.48550/arXiv.2203.02763>

The online list labeling problem is a basic primitive in data structures. The goal is to store a dynamically-changing set of n items in an array of m slots, while keeping the elements in sorted order. To do so, some items may need to be moved over time, and the goal is to minimize the number of items moved per insertion/deletion. When $m = Cn$ for some constant $C > 1$, an upper bound of $O(\log^2 n)$ items moved per insertion/deletion has been known since 1981. There is a matching lower bound for deterministic algorithms, but for randomized algorithms, the best known lower bound is $\Omega(\log n)$, leaving a gap between upper and lower bounds. We improve the upper bound, providing a randomized data structure with expected $O(\log^{3/2} n)$ items moved per insertion/deletion.

4.29 Dynamic Filter and Retrieval with One Memory Access

Guy Even (Tel Aviv University, IL)

License © Creative Commons BY 4.0 International license
© Guy Even

Joint work of Ioana-Oriana Bercea, Guy Even, Tomer Even

We present two dynamic data-structures in the word RAM model. The first data structure is a filter that supports approximate membership queries that is characterized by the following properties:

1. Dynamic: supports insert, delete, and membership-query operations.
2. Can store a dataset of cardinality at most n (insertions may fail with probability $o(1/\text{poly}(n))$).
3. Adjustable false-positive probability ε , provided that $\varepsilon = \Omega(1/(\text{polylog } n))$.
4. Space-efficient: $(1 + o(1)) \cdot n \log_2(1/\varepsilon) + O(n)$ bits.
5. Worst case constant time for every operation.
6. Single memory access per operation (not including hash function evaluation).

The second data structure is a retrieval data structure that supports value queries and is characterized by the following properties:

1. Supports the following operations: insert key-value pairs, delete a key, and query the value of a key.
2. Can store a set of at most n key-value pairs (insertions may fail with probability $1/\text{poly}(n)$).
3. Values are binary strings of length $O(\log \log n)$ bits.
4. Space-compact: $O(n \log \log n)$ bits.
5. Worst case constant time for every operation.
6. The expected number of memory accesses per operation is $1 + 1/\text{polylog}(n)$ (not including hash function evaluation).

The retrieval data structure has an additional “false-positive” feature: the response to a query for a key not in the dataset is NULL with probability at least $1 - 1/\text{polylog}(n)$.

4.30 Conditional Lower Bounds for Dynamic Geometric Measure Problems

Justin Dallant (UL – Brussels, BE)

License © Creative Commons BY 4.0 International license
© Justin Dallant

Joint work of Justin Dallant, John Iacono

Main reference Justin Dallant, John Iacono: “Conditional Lower Bounds for Dynamic Geometric Measure Problems”, in Proc. of the 30th Annual European Symposium on Algorithms, ESA 2022, September 5-9, 2022, Berlin/Potsdam, Germany, LIPIcs, Vol. 244, pp. 39:1–39:17, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <https://doi.org/10.4230/LIPIcs.ESA.2022.39>

We give new polynomial lower bounds for a number of dynamic measure problems in computational geometry. These lower bounds hold in the Word-RAM model, conditioned on the hardness of either 3SUM, APSP, or the Online Matrix-Vector Multiplication problem [Henzinger et al., STOC 2015]. In particular we get lower bounds in the incremental and fully-dynamic settings for counting maximal or extremal points in \mathbb{R}^3 , different variants of Klee’s Measure Problem, problems related to finding the largest empty disk in a set of points, and querying the size of the i ’th convex layer in a planar set of points. We also answer a question of Chan et al. [SODA 2022] by giving a conditional lower bound for dynamic approximate square set cover. While many conditional lower bounds for dynamic data structures have been proven since the seminal work of Patrascu [STOC 2010], few of them relate to computational geometry problems. This is the first paper focusing on this topic. Most problems we consider can be solved in $O(n \log n)$ time in the static case and their dynamic versions have only been approached from the perspective of improving known upper bounds. One exception to this is Klee’s measure problem in \mathbb{R}^2 , for which Chan [CGTA 2010] gave an unconditional $\Omega(\sqrt{n})$ lower bound on the worst-case update time. By a similar approach, we show that such a lower bound also holds for an important special case of Klee’s measure problem in \mathbb{R}^3 known as the Hypervolume Indicator problem, even for amortized runtime in the incremental setting.

Participants

- Kunal Agrawal
Washington University –
St. Louis, US
- Elena Arseneva
University of Lugano, CH
- Michael A. Bender
Stony Brook University, US
- Ioana Oriana Bercea
IT University of
Copenhagen, DK
- Guy E. Blelloch
Carnegie Mellon University –
Pittsburgh, US
- Gerth Stølting Brodal
Aarhus University, DK
- Alexander Conway
VMware Research –
Palo Alto, US
- Justin Dallant
UL – Brussels, BE
- Guy Even
Tel Aviv University, IL
- Rolf Fagerberg
University of Southern Denmark –
Odense, DK
- Martin Farach-Colton
Rutgers University –
Piscataway, US
- Inge Li Gørtz
Technical University of Denmark
– Lyngby, DK
- John Iacono
UL – Brussels, BE
- Riko Jacob
IT University of
Copenhagen, DK
- Rob Johnson
VMware – Palo Alto, US
- Valerie King
University of Victoria, CA
- Tsvi Kopelowitz
Bar-Ilan University –
Ramat Gan, IL
- László Kozma
FU Berlin, DE
- Rasmus Kyng
ETH Zürich, CH
- Moshe Lewenstein
Bar-Ilan University –
Ramat Gan, IL
- Quanquan C. Liu
Northwestern University –
Evanston, US
- Ulrich Carsten Meyer
Goethe-Universität –
Frankfurt am Main, DE
- Ian Munro
University of Waterloo, CA
- Markus E. Nebel
Universität Bielefeld, DE
- Eunjin Oh
POSTECH – Pohang, KR
- Rotem Oshman
Tel Aviv University, IL
- Seth Pettie
University of Michigan –
Ann Arbor, US
- Cynthia A. Phillips
Sandia National Labs –
Albuquerque, US
- Vijaya Ramachandran
University of Texas – Austin, US
- Rajeev Raman
University of Leicester, GB
- Eva Rotenberg
Technical University of Denmark
– Lyngby, DK
- Peter Sanders
KIT – Karlsruher Institut für
Technologie, DE
- Francesco Silvestri
University of Padova, IT
- Shikha Singh
Williams College –
Williamstown, US
- Nodari Sitchinava
University of Hawaii at Manoa –
Honolulu, US
- Yihan Sun
University of California –
Riverside, US
- Robert Endre Tarjan
Princeton University, US
- Nicole Wein
Rutgers University –
Piscataway, US
- Sebastian Wild
University of Liverpool, GB
- Helen Xu
Lawrence Berkeley National
Laboratory, US



Designing the Human-Machine Symbiosis

Ellen Yi-Luen Do^{*1}, Pattie Maes^{*2}, Florian ‘Floyd’ Mueller^{*3}, and Nathan Semertzidis^{†4}

1 University of Colorado – Boulder, US. ellen.do@colorado.edu

2 MIT – Cambridge, US. pattie@media.mit.edu

3 Monash University – Clayton, AU. floyd@floydmueller.com

4 Monash University – Clayton, AU. nathan@exertiongameslab.org

Abstract

Our understanding of computers simply executing tasks is changing towards one where the human and machine enter a symbiosis: computers are increasingly extending human capacity by integrating with bodily senses, thanks to sensor and actuator advances as well as enhanced software developments. Wearables, augmented reality, exoskeletons and implantable devices are all emerging trends that mark the beginning of such a human-machine symbiosis. What is still missing, though, is a thorough understanding of how to design such symbiotic user experiences in a systematic way, as, despite the increase of associated systems entering the market, there is a lack of understanding of how such a human-machine symbiosis emerges and what theoretical frameworks underlie it. Open questions around this topic are concerned with whether such systems can enhance human empowerment, what role a sense of control plays in the associated user experiences, and how to responsibly design devices that all people can benefit from. To begin answering such questions, this Dagstuhl Seminar invites experts from both industry and academia in order to bring together leaders from so far independent streams of investigation to work on a coherent approach to human-machine symbiosis that engages a holistic perspective while considering also societal and ethical issues.

Seminar May 21–26, 2023 – <https://www.dagstuhl.de/23212>

2012 ACM Subject Classification Networks → Network performance analysis; Applied computing → Computer-aided design; Information systems → Data streams; Software and its engineering → Concurrent programming languages

Keywords and phrases Human-Machine Symbiosis, Embodiment, Wearables, Bodily Extensions

Digital Object Identifier 10.4230/DagRep.13.5.136

1 Summary

Florian ‘Floyd’ Mueller (Monash University – Melbourne, AU, floyd@exertiongameslab.org)

License  Creative Commons BY 4.0 International license
© Florian ‘Floyd’ Mueller

In May 2023, a Dagstuhl Seminar took place in which 22 researchers and academics from across the world gathered for a week in Schloss Dagstuhl, Saarland, Germany, to discuss the future of the changing relationship between humans and computational machines, appraised by the group as an emerging form of “human-machine symbiosis”. The following manuscript documents said seminar and the efforts of its participant to investigate the underlying mechanisms of human-machine symbiosis, as well as knowledge guiding the design of such technologies, and the challenges the field of human-machine symbiosis faces moving forward.

* Editor / Organizer

† Editorial Assistant / Collector

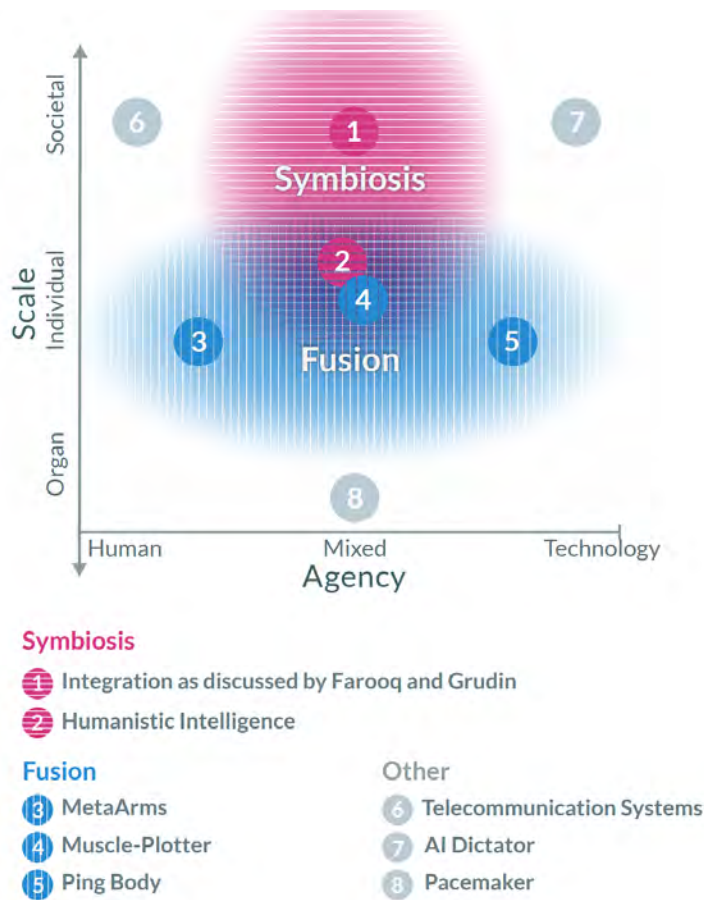


Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Designing the Human-Machine Symbiosis, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 136–164
Editors: Ellen Yi-Luen Do, Pattie Maes, Florian ‘Floyd’ Mueller, and Nathan Semertzidis



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



■ **Figure 1** An earlier map of Human-Computer Integration. Figure from “Next Steps in Human-Computer Integration,” by F. Mueller et al., 2020, Proceedings of the ACM 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-15.

In 1960 Licklider predicted that the future of computing would be one in which humans and machines would become tightly coupled and form a complimentary “symbiosis”, extending human capabilities [19]. Today, it appears we may be witnessing the advent of such a future. The traditional conceptual frameworks that attempted to describe the human-machine relationship over the last century of computing in terms of command and response, of master and slave, are becoming ever more antiquated. With the emergence of increasingly intelligent algorithms, complex computer architecture, and advanced sensor technologies, machines are now able to sense, understand, and act on the world as agents independent of human input and oversight. Describing this growing trend as “human-computer integration” in 2017 [12, 13], Farooq and Gruden deemed it to be a new paradigm within HCI, with “human-computer integration” referring to a move in technology away from the “stimulus-response” paradigm we commonly think of when we talk about interaction, and toward a “symbiotic partnership” between humans and computers, in which both parties are integrated and must be considered holistically.

The notion of human-computer integration was further developed at a 2018 Dagstuhl Seminar [26], in which 29 leading experts from industry and academia came together over a five-day seminar to develop and discuss the future of human-computer integration

(HInt). The discussions during this seminar ultimately produced a foundational work titled “Next Steps in Human-Computer Integration” [25] which was presented at CHI 2020. The work, articulating a synthesis of contributions made toward human-Computer integration, summarised the contemporary state of the emerging paradigm. The paper defined HInt as “a new paradigm with the key property that computers become closely integrated with the user”, which included examples in which “humans and digital technology work together, either towards a shared goal or towards complementary goals (symbiosis)”; and “integration in which devices extend the experienced human body or in which the human body extends devices (fusion)”. Through this updated rendition, it became clear that this new state of the human-machine relationship was not only marked by the newly emerging capacity for machines to exhibit agency and work in tightly coupled collaborative partnerships with humans [20, 5, 18, 21, 29]; but also by the tendency for emerging technologies to extend human capabilities by bidirectionally sensing and actuating human physiology and act as extensions of the human body [7, 15, 16, 22, 33, 34].

Moving forward, much research has since been made in contribution to furthering our understanding of human-computer integration theory and the design of integration systems, including a definitive book on human-computer integration titled: “Human-Computer Integration: Towards Integrating the Human Body with the Computational Machine” [28]. These contemporary conceptualizations and theoretical contributions to HInt have largely taken on a more “bodily” focus, centering on how machines can integrate with the human body, physiological processes, and experiences [31, 32]. Such theoretical contributions include: the bodily integration framework [24], which describes the user experience of integration systems that both sense and actuate the human body by considering the human’s sense of bodily agency and sense of bodily ownership; experiential integration [6, 7], which seeks to understand how machines can be integrated into an individual’s pre-reflective experience of “self” as opposed two “other”; and the brain-computer integration framework [30], which provides a design space for brain-computer interfaces that bidirectionally sense and actuate the human brain, as well as the user experiences they create. Taken together, these works provide the conceptual tools to understand and design technologies that are physically close to or conform to the body [16, 33, 34, 2], extend the body’s sensory and motor capabilities [15, 22], augment cognitive abilities [4, 14, 10], and provide novel modalities for human expression and play [23, 3, 9, 17, 11, 21], facilitating empowerment and self-actualization.

More recently, contributions to HInt theory are increasingly concerned with the inter-agential dynamics and relationships that arise from systems that are closely coupled with the human body but are able to act with their own agency [20, 11], often sharing agency with, or borrowing agency from their human counterpart. Such theoretical works include: the intertwined integration framework [27], which maps the possible user experiences that arise from a combination of the alignment of the system with the user, and the user’s awareness of the agency of the system; integrated embodiment [18], which explores how human-ai partnerships can be embodied within a single physical body; and the integrated exertion framework [1], which describes the different experiences of human-machine partnerships that may arise in an exertion context. Considering these newer contributions, it becomes increasingly evident that the original conceptualization of symbiosis and fusion as two opposing forms of integration, one social and one bodily, no longer completely describes the new and evolving relationships we are beginning to see emerge between humans and computers, highlighting a significant gap in our understanding of human-machine symbiosis theory and design.

Considering the recent acceleration in technological advances that enable autonomous agents and artificial intelligence to form new and previously unseen relationships with humans, both socially and physiologically, the importance of having a strongly developed and fully articulated definitive theory of human-machine symbiosis and their design becomes extremely pertinent. While the growing presence of human-machine symbiosis in our daily lives can greatly extend human capacity, capability, experience, and culture, recent research has also highlighted that human-computer symbiosis holds the strong potential to develop into parasitic relationships which can be detrimental to humanity [8]. Thus, considering the growing relevance of symbiosis, in contrast to the relatively nascent state of the field in terms of theoretical understanding and design knowledge, the present seminar sought to lead a concerted effort in developing the theory of human-machine symbiosis, articulating its underlying mechanism, and outlining the grand challenges facing the field moving forward. In doing so, the seminar asked: What symbiosis is, the mechanisms through which it operates, and what kinds of symbioses are possible? Then with this knowledge, the seminar aimed to build an understanding of how to guide the design of symbiosis. Finally, through acknowledging the limitations to our current understanding of symbiosis and our abilities in designing symbiotic systems, the seminar articulated a set of grand challenges that we can work toward to move the field forward. The following documents the activities undertaken by the participants of the seminar in an effort to actualize these aims.

References

- 1 Josh Andres, Nathan Semertzidis, Zhuying Li, Yan Wang, and Florian Mueller. Integrated exertion–understanding the design of human–computer integration in an exertion context. *ACM Transactions on Computer-Human Interaction*, 29(6):1–28, 2023.
- 2 Andrea Bianchi, Steve Hodges, David J Cuartielles, Hyunjoo Oh, Mannu Lambrichts, and Anne Roudaut. Beyond prototyping boards: future paradigms for electronics toolkits. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2023.
- 3 Oğuz ‘Oz’ Buruk, Louise Petersen Matjeka, and Florian Mueller. Towards designing playful bodily extensions: Learning from expert interviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.
- 4 Yi Fei Cheng, Hang Yin, Yukang Yan, Jan Gugenheimer, and David Lindlbauer. Towards understanding diminished reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.
- 5 S Chien, Seungyeon Choo, Marc Aurel Schnabel, Walaiporn Nakapan, Mi Jeong Kim, and Stanislav Roudavski. Living systems and micro-utopias: towards continuous designing. In *Proceedings of the 21st International Conference of the Association for Computer-Aided Architectural Design Research in Asia CAADRIA 2016*, pages 713–722, 2016.
- 6 Valdemar Danry, Pat Pataranutaporn, Adam Haar Horowitz, Paul Strohmeier, Josh Andres, Rakesh Patibanda, Zhuying Li, Takuto Nakamura, Jun Nishida, Pedro Lopes, et al. Do cyborgs dream of electric limbs? experiential factors in human-computer integration design and evaluation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.
- 7 Valdemar Danry, Pat Pataranutaporn, Florian Mueller, Pattie Maes, and Sang-won Leigh. On eliciting a sense of self when integrating with computers. In *Augmented Humans 2022*, pages 68–81. 2022.
- 8 Rod Dickinson, Nathan Semertzidis, and Florian Mueller. Machine in the middle: Exploring dark patterns of emotional human-computer integration through media art. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

- 9 Ellen Yi-Luen Do. Design for assistive augmentation – mind, might and magic. *Assistive augmentation*, pages 99–116, 2018.
- 10 Barrett Ens, Maxime Cordeil, Chris North, Tim Dwyer, Lonni Besançon, Arnaud Prouzeau, Jiazhou Liu, Andrew Cunningham, Adam Drogemuller, Kadek Ananta Satriadi, et al. Immersive analytics 2.0: Spatial and embodied sensemaking. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.
- 11 Xiao Fang, Nathan Semertzidis, Michaela Scary, Xinyi Wang, Josh Andres, Fabio Zambetta, and Florian Mueller. Telepathic play: Towards playful experiences based on brain-to-brain interfacing. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, pages 268–273, 2021.
- 12 Umer Farooq, Jonathan Grudin, Ben Shneiderman, Pattie Maes, and Xiangshi Ren. Human computer integration versus powerful tools. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1277–1282, 2017.
- 13 Umer Farooq and Jonathan T Grudin. Paradigm shift from human computer interaction to integration. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1360–1363, 2017.
- 14 Lukas Gehrke, Pedro Lopes, and Klaus Gramann. Toward human augmentation using neural fingerprints of affordances. In *Affordances in Everyday Life: A Multidisciplinary Collection of Essays*, pages 173–180. Springer, 2022.
- 15 Masahiko Inami, Daisuke Uriu, Zendai Kashino, Shigeo Yoshida, Hiroto Saito, Azumi Maekawa, and Michiteru Kitazaki. Cyborgs, human augmentation, cybernetics, and jizai body. In *Augmented Humans 2022*, pages 230–242. 2022.
- 16 Hsin-Liu Cindy Kao. Hybrid body craft: toward culturally and socially inclusive design for on-skin interfaces. *IEEE Pervasive Computing*, 20(3):41–50, 2021.
- 17 Sang-won Leigh, Abhinandan Jain, and Pattie Maes. Exploring human-machine synergy and interaction on a robotic instrument. In *NIME*, pages 437–442, 2019.
- 18 Zhuying Li, Tianze Huang, Rakesh Patibanda, and Florian Mueller. Ai in the shell: Towards an understanding of integrated embodiment. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2023.
- 19 Joseph CR Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11, 1960.
- 20 Dominika Lisy. In-corporo-real robot-dreams: Empathy, skin, and boundaries. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–05. IEEE, 2021.
- 21 Azumi Maekawa, Hiroto Saito, Daisuke Uriu, Shunichi Kasahara, and Masahiko Inami. Machine-mediated teaming: Mixture of human and machine in physical gaming experience. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.
- 22 Azumi Maekawa, Hiroto Saito, Daisuke Uriu, Shunichi Kasahara, and Masahiko Inami. Machine-mediated teaming: Mixture of human and machine in physical gaming experience. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2022.
- 23 Florian Mueller, Tuomas Kari, Zhuying Li, Yan Wang, Yash Dhanpal Mehta, Josh Andres, Jonathan Marquez, and Rakesh Patibanda. Towards designing bodily integrated play. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 207–218, 2020.
- 24 Florian Mueller, Pedro Lopes, Josh Andres, Richard Byrne, Nathan Semertzidis, Zhuying Li, Jarrod Knibbe, Stefan Greuter, et al. Towards understanding the design of bodily integration. *International Journal of Human-Computer Studies*, 152:102643, 2021.

- 25 Florian Mueller, Pedro Lopes, Paul Strohmeier, Wendy Ju, Caitlyn Seim, Martin Weigel, Suranga Nanayakkara, Marianna Obrist, Zhuying Li, Joseph Delfa, et al. Next steps for human-computer integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.
- 26 Florian Mueller, Pattie Maes, and Jonathan Grudin. Human-computer integration (dagstuhl seminar 18322). In *Dagstuhl Reports*, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
- 27 Florian Mueller, Nathan Semertzidis, Josh Andres, Joe Marshall, Steve Benford, Xiang Li, Louise Matjeka, and Yash Mehta. Towards understanding the design of intertwined human-computer integrations. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.
- 28 Florian Mueller, Nathan Semertzidis, Josh Andres, Martin Weigel, Suranga Nanayakkara, Rakesh Patibanda, Zhuying Li, Paul Strohmeier, Jarrod Knibbe, Stefan Greuter, Marianna Obrist, et al. Human-computer integration: Towards integrating the human body with the computational machine. *Foundations and Trends® in Human-Computer Interaction*, 16(1):1–64, 2022.
- 29 Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsonon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.
- 30 Nathan Semertzidis, Fabio Zambetta, and Florian Mueller. Brain-computer integration: A framework for the design of brain-computer interfaces from an integrations perspective. *ACM Transactions on Computer-Human Interaction*, 2023.
- 31 Nathan Arthur Semertzidis, Zoe Xiao Fang, Pedro Lopes, Kai Kunze, Paul Pangaro, Florian Mueller, and Pattie Maes. What we talk about when we talk about human-computer integration. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–4, 2022.
- 32 Nathan Arthur Semertzidis, Michaela Scary, Xiao Fang, Xinyi Wang, Rakesh Patibanda, Josh Andres, Paul Strohmeier, Kai Kunze, Pedro Lopes, Fabio Zambetta, et al. Sighint: Special interest group for human-computer integration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–3, 2021.
- 33 Jürgen Steimle, Marie Muehlhaus, Madalina Luciana Nicolae, Aditya Shekhar Nittala, Narjes Pourjafarian, Adwait Sharma, Marc Teyssier, Marion Koelle, Bruno Fruchard, and Paul Strohmeier. Design and fabrication of body-based interfaces (demo of saarland hci lab). In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2023.
- 34 Anusha Withana, Daniel Groeger, and Jürgen Steimle. Tacttoo: A thin and feel-through tattoo for on-skin tactile output. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 365–378, 2018.

2 Table of Contents

Summary

<i>Florian ‘Floyd’ Mueller</i>	136
--	-----

Introductions

Defining human-machine symbiosis	144
Interactivity session	146
Domains of symbiosis	147
Symbiosis special interest groups	150
Hike	150
Human-AI Symbiosis workshop	151
Collating and writing	152
Conclusion	152

Overview of Talks

If all you have is a hammer <i>Andrea Bianchi</i>	153
Symbiosis over time with physical computing <i>Anusha Withana</i>	154
Design of dynamic human-machine coupling system <i>Azumi Maekawa</i>	155
Making sense of information anywhere <i>Barrett Ens</i>	155
Hybrid skins for symbiosis <i>Cindy Hsin-Liu Kao</i>	156
Dermal layers in between the human self and the robot other <i>Dominika Lisy</i>	157
Designing the human-machine symbiosis? fun with creative technology and design <i>Ellen Yi-Luen Do</i>	157
Symbiosis is bodily <i>Florian ‘Floyd’ Mueller</i>	158
Skin as an interface for human-machine symbiosis <i>Jürgen Steimle</i>	158
Toward identifying features that make human-machine relationships symbiotic <i>Kumiyo Nakakoji</i>	159
Agency-preserving action augmentation using brain-computer interfaces <i>Lukas Gehrke</i>	159
JIZAI body and symbiosis <i>Masahiko Inami</i>	160
Speculation on a world with social digital cyborgs <i>Nahoko Yamamura</i>	160

Brain-computer symbiosis
Nathan Semertzidis 160

Tools, medium, mediator, partner, and beyond...
Sheng-Fen ‘Nik’ Chien 161

Understanding games and play in a posthuman era
Oğuz ‘Oz’ Buruk 161

How will people live symbiotically with AI?
Pattie Maes 162

Machine poetics
Sang-won Leigh 162

Human-Machine symbiosis via a mixed reality
Yi Fei Cheng 162

Human-machine symbiosis
Zhuying Li 163

Participants 164

3 Introductions

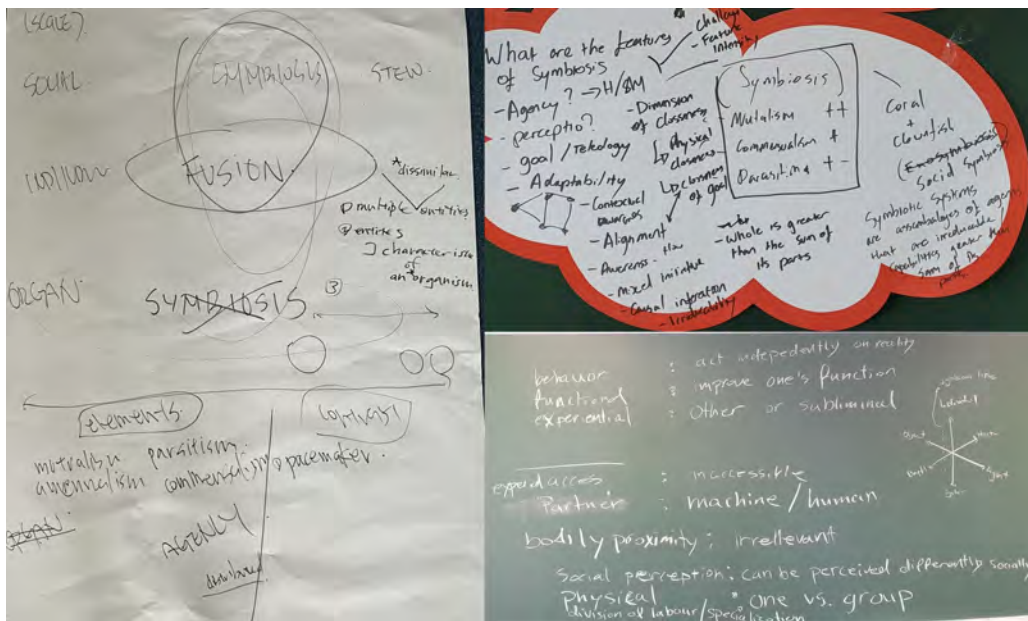
To begin the seminar, an embodied introductory activity was first undertaken in order to facilitate an accelerated familiarisation between seminar participants, while also establishing an atmosphere of playfulness, openness, and enthusiasm for group interaction. This involved the seminar participants standing in a circle and introducing themselves with their name and a bodily gesture, with all other participants needing to repeat the name and gesture. This was completed clockwise around the circle of participants, and for each new participant introduction, the introduction of all previous participants would also be repeated in sequence. This accumulated to 253 introductions for the 22 participants present by the end.

The introductory activity was followed by opening remarks from organizer Florian ‘Floyd’ Mueller, who provided theoretical context for the seminar and set the basis for the discussion of designing the human-machine symbiosis moving forward. At this point, several whiteboards were also established that participants could add to at any point during the town hall sessions. This included a whiteboard for listing the “challenges and opportunities” of designing the human-machine symbiosis, which anyone could append to as challenges and opportunities became evident through presentations and discussion. Similarly, a whiteboard titled “the marketplace” was established for anyone to post ideas for potential collaborative future projects, papers, workshops, or other initiatives with the intention that other participants could express their interest to join said initiative. The marketplace specifically produced several offshoot human-machine symbiosis research initiatives that are now currently being continued after the completion of the seminar, including: the organization of a posthuman theory and design workshop; the organization of a brain-computer interface workshop, and research paper; research projects involving symbiotic systems that utilize biological materials; and the writing of several other Dagstuhl proposals.

Each seminar participant then gave a prepared presentation introducing their research. In order to preserve a spirit of open and spontaneous group ideation, discourse, and collaboration, we strove to avoid a conference-like format of dry, dense lengthy presentations and instead adopted a “PechaKutcha” inspired presentation format; a rapid-fire series of short six-minute, visually oriented (picture and video) presentations. In addition to showcasing their work and its relation to human-machine symbiosis, the content of each presentation also included an articulation of what the presenter expected from the seminar, the grand challenges facing their niche of human-computer symbiosis (which were appended to the challenges whiteboard), and what prior work the group should read and why, with the interest of establishing a common conceptual toolbox which the group can draw from to develop human-machine symbiosis theory and discussion. An abstract of each presentation can be found at the end of the present report.

3.1 Defining human-machine symbiosis

Through the presentations above, a comprehensive list of “challenges” was collated by the time every participant had presented, providing a strong foundation for steering discussion during the remainder of the seminar toward topics that require further elaboration. It has been argued that HCI requires “grand challenges”, namely in that it provides a steering force to drive coordinated action in guiding future research, theory, design, and commercial development [3, 1, 2]. Acknowledging this need, these challenges were further consolidated into a set of concrete “Grand Challenges for Human-Machine Symbiosis”, with the intention

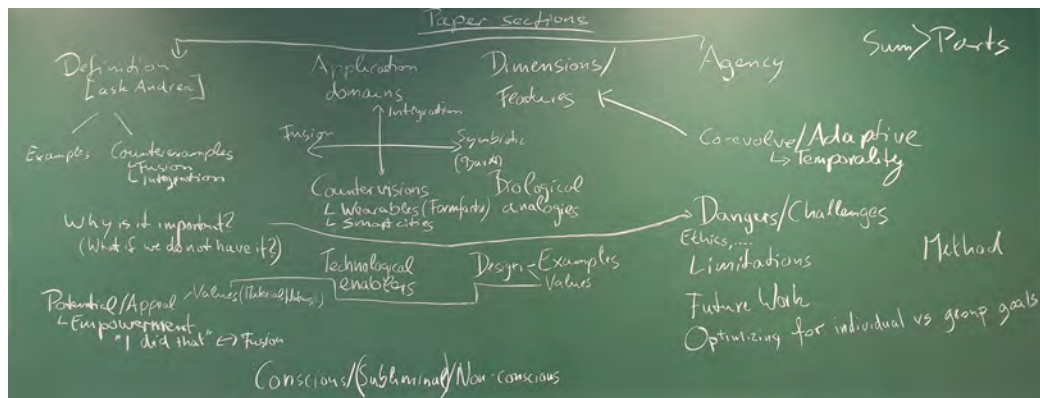


■ **Figure 2** The three articulations made by the breakout groups in an effort to define human-machine symbiosis.

their articulation would give guidance to researchers wishing to contribute to the development of the theory by providing specific and actionable gaps in knowledge or capability to which they can contribute through future research. Furthermore, the collation of recommended reading provided by each participant of the seminar now serves as a library curated by experts that may provide entrants into human-machine symbiosis research a foundational understanding of core concepts, theory, and knowledge that may guide their own future investigations, making the field more accessible to potential new contributors.

Following the completion of all participant presentations, a discussion took place in which the participants synthesized the concepts brought to light during the presentations, as well as the opportunities and challenges that were also highlighted. Through this discussion, it became evident that many challenges and opportunities pointed toward there being a significant need to articulate a clearer, more grounded, more comprehensive definition of what human-machine symbiosis is. This sentiment was further reinforced by many participants showing eagerness to draw knowledge from existing biological science explanations of symbiotic relationships, in conjunction with existing theories that deal with concepts like “self” (such as philosophy of mind) and describe “agents”, “agency”, and inter-agent interactions. With this considered, all 22 participants were broken into three breakout groups, with each group tasked with individually providing an articulation of what “human-machine symbiosis” is. Each group approached this in its own unique way, emerging from how its members organized to address the task of defining human-machine symbiosis, which resulted in it being articulated in terms of its dimensions, features, and underlying postulates, which each breakout group presented to the rest of the seminar in its entirety.

All participants then gathered to return to the main seminar room and a town hall discussion followed in which the participants attempted to integrate the three definitions into a unifying framework of human-machine symbiosis. A summary of the consensus suggested that central to symbiosis is a dynamic relationship between tightly coupled agents who



■ **Figure 3** The initial skeleton of the sections of the human-machine symbiosis manuscript based on the discussion at the conclusion of the first day.

are able to act independently on reality, yet assemble to exhibit emergent properties or capabilities that would otherwise not be possible with the constituent agents in isolation. These relationships would include at least one human and one machine agent but can include many of either. It was also agreed there could be a number of different types of these relationships that describe dynamics between the symbiotic agents. These types were drawn from biology and include: Mutualism (where both agents benefit); commensalism (where one agent benefits and one is unaffected); and parasitism (where one agent benefits at the expense of the other) however some suggested that it may be best to focus first on mutualism before exploring other types of relationships for the sake of simplicity. However, the attempt to define human-machine symbiosis also highlighted some theoretical challenges which were unable to be solved by the conclusion of the discussion, including where the line is between an algorithm and an agent (i.e. when does a program become an agent with agency?), where does symbiosis sit in relation to human-computer integration (i.e. are fusion and symbiosis still two subsets of integration?). Another ongoing question was whether symbiotic systems needed to be close to the human body, with the common concession being that on-body systems were not a requisite, but seen as more likely to form symbiotic relationships, or lend themselves to forming stronger symbiosis. Finally, the major themes and outcomes of this discussion were transcribed and translated into sections and subtitles of a future “Designing the human-machine symbiosis” paper, resulting in an initial skeleton of the manuscript, concluding the first day of the seminar.

3.2 Interactivity session

The second day of the seminar was introduced by Ellen Yi-Luen Do, starting off with an “interactivity session” that involved live interactive demonstrations of systems and technologies relating to human-machine symbiosis. Several of the participants of the seminar had brought with them prototypes or technologies which were set up around the main seminar room. Other participants could then move about the room to try out systems and ask questions, with the intention that this hands-on demonstration of symbiosis-enabling technologies may inspire participants to ideate their own symbiosis systems.

Nathan Semertzidis demonstrated a novel electroencephalography (EEG) headset as well as transcranial electrical stimulation (tES) device, which his projects often employ in combination with a closed-loop to bilaterally sense and stimulate the brain. Participants

were given the opportunity to experience a phosphine, a flash of light experienced in the visual field evoked by the stimulation of the optic nerve using the tES device. Nathan argued that these technologies employed in human-machine symbiosis systems could serve to provide a bridge between biological neural networks (the brain) and artificial neural networks (A.I.) opening up many possibilities for mental mergers with A.I. systems.

Nahoro Yamamura distributed to each participant a Lego set which, when assembled, formed a prosthetic extra finger participants could wear as a bodily extension. The finger was indented as an exercise to explore how the body might react to a prosthetic extension, and whether the brain would integrate it into its body schema. Some participants wore the finger for the entire duration of the seminar, reporting that sometimes it would be experienced as part of their body (for example describing that they got confused when they would move the rest of their fingers and the prosthetic would not move).

Zoe Xiao Fang demonstrated a lollipop that, when bitten, could deliver a tune to the brain through its bone-conductive lollipop stick. This demonstration highlighted the potential for food-related technologies to integrate with human physiology.

Andrea Bianchi demonstrated a collaborative microcontroller remote development toolkit. It was explained that the system was originally developed to facilitate the teaching of microcontroller development remotely during COVID-19-related lockdowns, however, it was hypothesized that such systems could also be used to teach and prototype symbiotic systems.

Jürgen Steimle demonstrated a software toolkit that enables the prototyping of bodily extensions and on-body interactions. Similarly, it was considered such toolkits could be used to rapidly prototype symbiotic systems.

Masahiko Inami demonstrated webcams that had been anthropomorphized as eyes, and speakers that had been anthropomorphized as mouths, which could both be fixed to any surface. This demonstration was intended to be a playful exploration of the embodiment and integration of inanimate objects.

Valdemar Danry demonstrated two AI conversational agents using large language models to duplicate existing people, allowing participants to talk with them. One was a duplication of Leonardo da Vinci, and another was a duplication of himself at a younger age. It was discussed that such AI systems could in the future form social symbiosis with their human counterparts, freeing cognitive demand by sharing tasks, appearing as one but distributing the workload.

3.3 Domains of symbiosis

Following on from the previous days defining of symbiosis and inspired by the interactivity session, the participants were then prompted to consider what types of symbiosis systems, areas of symbiosis theory, or application domains they would most like to further develop or contribute to. Participants then wrote their answers down on sticky notes, which were then posted on the main blackboard of the seminar town hall room. All participants then attempted to group these sticky notes in a self-organized manner until higher-level groupings emerged.

These high-level groupings were then reorganized into a set of application domains (applications human-machine symbiosis systems could be designed for). Participants were then instructed to form breakout groups with each group focusing on one application domain. Initially, these were 1) augmented body: sensory-motor 2), augmented cognition, 3) social symbiosis, 4) biological symbiosis, and 5) symbiotic play. However social symbiosis and



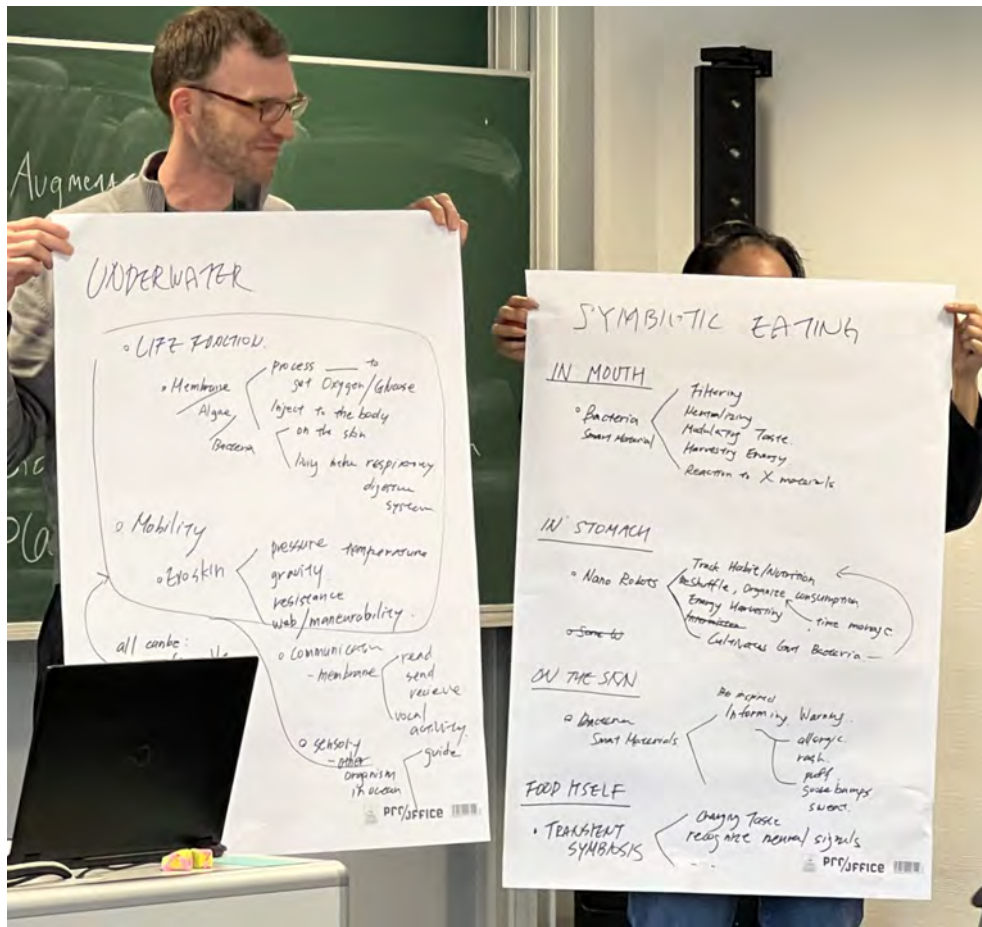
■ **Figure 4** Participant Oğuz ‘Oz’ Buruk trying the prosthetic sixth lego finger on the left, and the anthropomorphic webcam eye on the right.

biological symbiosis were merged into a single group due to having a low number of members in each. Each group was then tasked with ideating hypothetical examples of human-machine symbiosis systems designed for their given application domains. Each of the groups generated a great number of designs, that were then further refined into two to four designs that were explored more deeply. The groups then reconvened in the main seminar room and presented their designs to the rest of the seminar participants, concluding the day’s activities. The following presents a summary of what each group presented.

Group 1) “augmented body” ideated several systems that symbiotically augmented the body to enhance its capabilities in a variety of contexts. This included symbiosis for underwater living, which proposed engineering biological symbiotes that support human homeostasis as well as mobility in aquatic environments through technologies such as microbe-filled wearable membrane bioreactors, and “exoskins” which can sense and respond to properties of the water around it. Group 1 also ideated systems for “symbiotic eating” including microbes, smart materials, and nanorobots that can perform functions in the human mouth, stomach, skin, and on the food itself in order to modulate energy and nutrition intake, and alter the experience of eating such as taste.

Group 2) “augmented cognition” proposed several systems that work toward enhancing the capabilities of various domains of human cognition. One example was a system that augments memory processes by monitoring the brain for occurrences of error-related potentials or inability to recall information, reflexively rewinding through a log the system has kept of the day (e.g. through video or auditory recording) to find the information in its correct form. Similarly, the system could also employ predictive algorithms to allow the user to “recall” information from events yet to occur. The group also suggested that systems could capitalize on how affective states can alter memory encoding and recall to improve learning and facilitate desired forgetting.

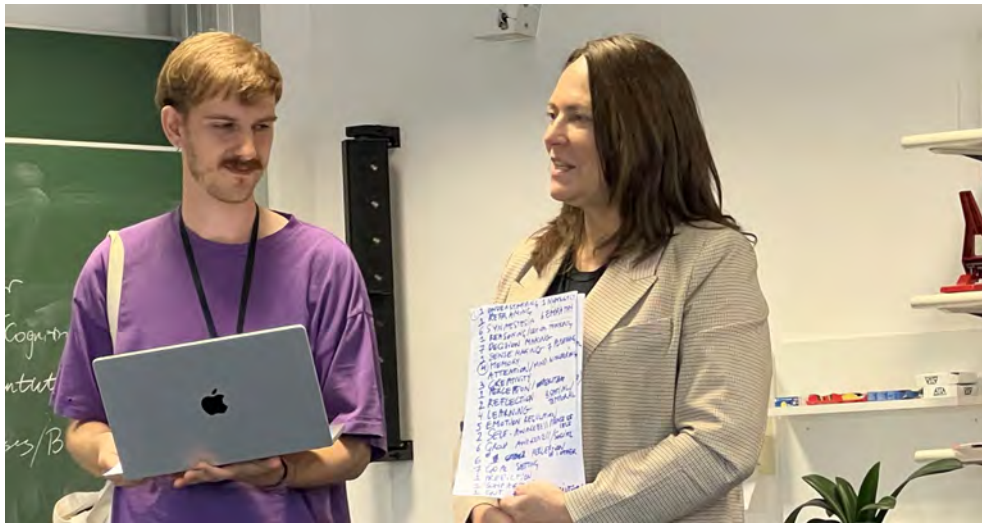
Group 3) “biosocial symbiosis” was a combination of participants interested in investigating the social applications and consequences of symbiosis, and participants interested in symbiotic systems that have biological components. As such, the designs that came out of this group typically involved some kind of biological symbiotic system, whose use brought with it social



■ **Figure 5** Two participants from the augmented body group reporting their results.

consequences, both between the symbiont and the wearer, as well as between other members of society. These included: a second spine that gives the wearer courage through the direct injection of adrenaline which it produces but must be kept in a fish tank during non-use to avoid adrenal overdose; an implantable photosynthetic organism that gives the wearer the ability to photosynthesis but must constantly be managed else it will overgrow (and its use is also associated with a specific socioeconomic status); and a subdermal bioreactor that consumes excess nutrients from its wearer who can then sell the energy stored in the bioreactor for income.

Group 4) “symbiotic play” ideated examples that explored how symbiotic systems could be designed for games or playful experiences. This highlighted that symbiotic games and play can serve as a lens for safely identifying and exploring potential dystopic futures symbiotic systems may one day actualize. Some examples included a “prank” exoskeleton that forces its wearer to run or dance benefiting them through exercise; and swarms of interactive prosthetic “eye entities” that the user can see through, providing them with emergent abilities based on how they configure the eyes.



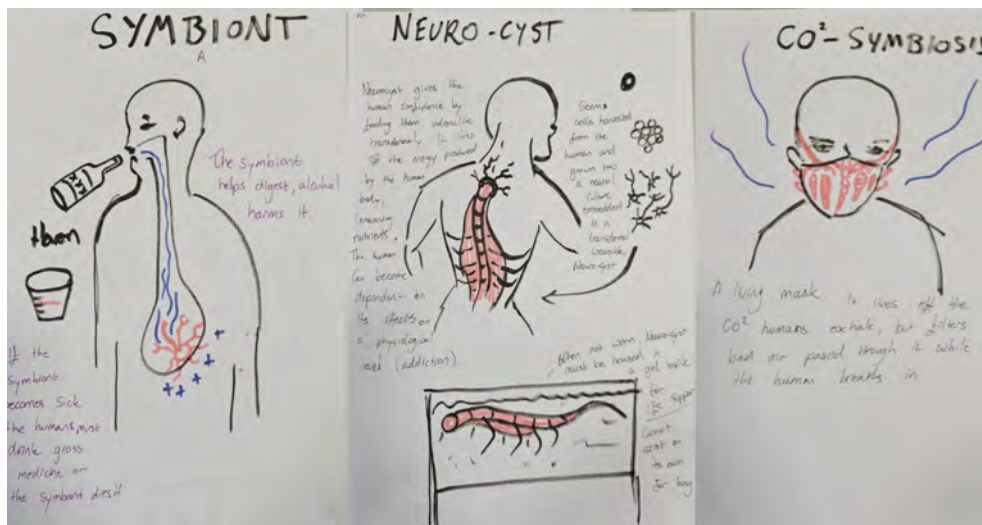
■ **Figure 6** Two participants from the augmented cognition group reporting their results.

3.4 Symbiosis special interest groups

The third day of the seminar was initiated with a town hall discussion that reviewed the progress we had made so far in developing human-machine symbiosis theory and in the identification and articulation of its grand challenges and next steps, while also highlighting what conceptual areas required further development. Considering these points, participants then broke up into groups centered around projects, initiatives, and special interests that had accumulated on the “marketplace” whiteboard throughout the duration of the seminar. What each group did during this breakout session depended on the specific topic it was aligned with, with each group autonomously working toward its self-organized agenda. The topics covered by each group included: further developing human-machine symbiosis theory, with a specific focus on the different “levels” of symbiosis; further developing human-machine symbiosis theory, with a specific focus on the temporal, spatial, and structural properties and outcomes of symbiotic relationships; exploration of how human-machine symbiosis could facilitate deeper connections with nature, with a focus on writing a Dagstuhl seminar proposal around the topic; exploring the idea of human-machine symbiosis using biological materials, with a focus on ideating possible future research projects; and an exploration of the concept of “body hijacking” as an example of human-machine symbiosis, specifically referring to technological systems that allow one agent to take over the body of another. Each breakout group reconvened in the main seminar room and presented what they produced to the rest of the participants.

3.5 Hike

The seminar participants then embarked on a hike in Saarschleife along Mettlach, which lasted for most of the remainder of the day. The hike was intended to revive, refresh and inspire participants by breaking from the constant intensive discussion and conceptualization of the days prior, while also giving them an opportunity to exchange thoughts, ideas, and career advice, and to forge new research relationships and collaborations. Furthermore, during the hike participants saw that it was a good opportunity to keep an eye out for



■ **Figure 7** A sample of three example systems from the biosocial group.

naturally occurring examples of symbiosis, which lead to participants drawing inspiration from examples that were encountered such as lichens, pollinator symbiosis, symbiosis between ants and plants, and the parasitism of vines growing on trees. The hike ended with drinks and dinner at a local restaurant, concluding the day.

3.6 Human-AI Symbiosis workshop

The next day began with the initiation of a workshop led by Valdemar Danry and Pattie Maes, focusing on how recent advances in generative AI can be applied toward the development of AI-centered human-machine symbiosis relations (human-AI symbiosis). The workshop began with a presentation showcase demonstrating work the MIT Media lab's fluid interfaces group has been doing to augment human cognition through offloading cognitive resources from the human to often wearable AI symbiotic systems, enabling the experience of enhanced memory, attention, and sensory abilities. Similarly, research conducted by others using generative AI was also showcased in order to demonstrate the widespread potential such technology may offer through symbiosis with a human partner (e.g. such as enhanced creative ability). Participants were then introduced to resources they could utilise to begin work on such systems themselves.

Following the presentation, participants then formed breakout groups around different applications that would benefit from human-AI symbiosis. The groups included: sensory augmentation and assistive tech; cognitive augmentation and assistive tech; motor augmentation and assistive tech; and play, games and sports. Each group ideated novel designs that exemplified how generative AI could be employed toward the ends of the specific application domain they were assigned with. Most groups generated a large number of designs, which were then refined to a single design that would be demonstrated to the rest of the participants after reconvening in the main seminar room. Most groups chose to present their designs and ideas in the form of a roleplay, usually with different participants acting out the role of the human user, the AI symbiont, and various components of the system.



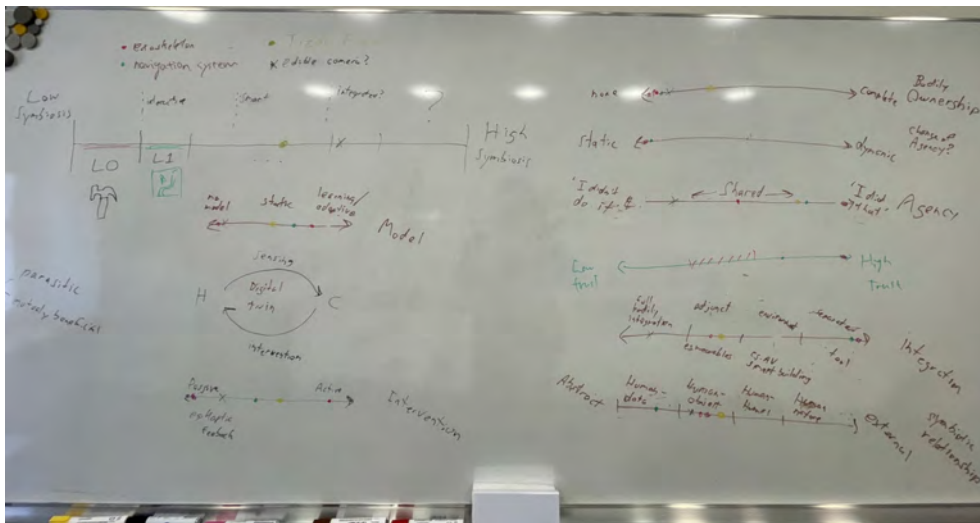
■ **Figure 8** The symbiotic play group reporting their results.

3.7 Collating and writing

A town hall discussion was then held focused on the writing of the “Designing the Human-Machine Symbiosis” manuscript. Participants further refined the outlining sections of the manuscript and identified other areas requiring further development the group could contribute toward. Based on these sections, breakout groups formed around the topics: defining human-machine symbiosis; methods and measures to study human-machine symbiosis; topologies of human-machine symbiosis; and the grand challenges facing human-machine symbiosis. Each of these groups contributed to the writing of the main manuscript by focusing on their aligned section. In addition, another breakout group was formed to draft a new Dagstuhl proposal for Human-Nature Interaction. Groups continued writing their sections until the completion of the working day.

3.8 Conclusion

The final day of the seminar began with closing remarks from the organizers. A final town hall discussion was held in which participants reiterated what we have established so far in building a theory of human-machine symbiosis and articulating a set of challenges that future research could address to make further contributions to the field. We also highlighted where there might be gaps in the concept, and what next steps could be taken to further crystalize the concept. It was also decided a slack group be formed to maintain contact between participants and facilitate ongoing discussion in the human-machine symbiosis research community. Following these closing remarks, participants spent the remainder of their time at Dagstuhl contributing to the “Designing the human-machine symbiosis” manuscript. As of writing this report, the manuscript is still currently in development.



■ **Figure 9** A sample of the results of the “levels of symbiosis” group.

References

- 1 Jason Alexander, Anne Roudaut, Jürgen Steimle, Kasper Hornbæk, Miguel Bruns Alonso, Sean Follmer, and Timothy Merritt. Grand challenges in shape-changing interface research. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.
- 2 Barrett Ens, Benjamin Bach, Maxime Cordeil, Ulrich Engelke, Marcos Serrano, Wesley Willett, Arnaud Prouzeau, Christoph Anthes, Wolfgang Büschel, Cody Dunne, et al. Grand challenges in immersive analytics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.
- 3 Juliet Norton, Ankita Raturi, Bonnie Nardi, Sebastian Prost, Samantha McDonald, Daniel Pargman, Oliver Bates, Maria Normark, Bill Tomlinson, Nico Herbig, et al. A grand challenge for hci: Food+ sustainability. *interactions*, 24(6):50–55, 2017.

4 Overview of Talks

4.1 If all you have is a hammer

Andrea Bianchi (Korea Advanced Institute of Science and Technology – Daejeon, KR, andrea@kaist.ac.kr)

License © Creative Commons BY 4.0 International license
© Andrea Bianchi

Since the dawn of mankind, the history of the human race is reflected in the history of their tools and their usage. Many of these tools provide augmentation to our physical capabilities: power tools increase the body’s strength, bikes increase locomotion efficiency, and glasses and microscopes increase vision and the human ability to explore the world. However, more interestingly, tools also shape the way we think. It is known that “if all you have is a hammer, everything looks like a nail” (Maslow’s hammer), and to some extent, this is true for any type of tool, as they unconsciously reshape our perception of reality, our consciousness, and our understanding of how to interact with the world surrounding. In this short presentation,



■ **Figure 10** A snapshot of each breakout group roleplaying their symbiotic AI concept. In clockwise order starting from top left the groups are 1) motor augmentation, 2) play games and sports, 3) sensory augmentation, and 4) cognitive augmentation.

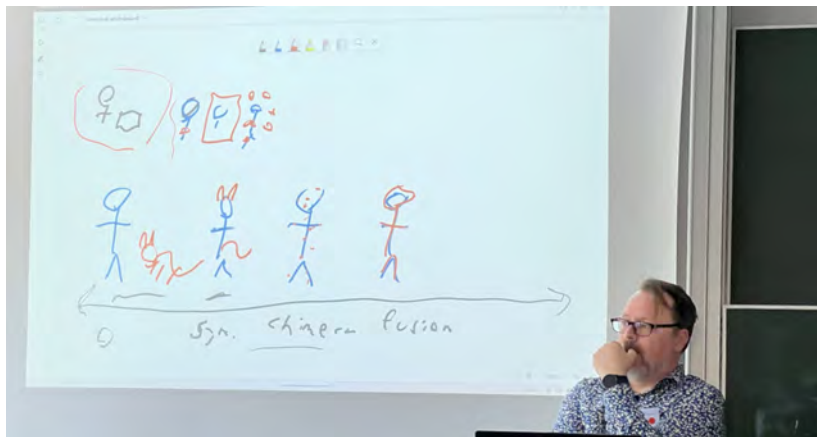
I show examples of digitally augmented physical tools that shape our perception of reality and give us new perspectives on how to design for supporting prototyping as an exploration activity, and virtual-physical interactions.

4.2 Symbiosis over time with physical computing

Anusha Withana (University of Sydney, AU, anusha.withana@sydney.edu.au)

License  Creative Commons BY 4.0 International license
© Anusha Withana

In biology, symbiosis sustains over long durations of time. With AI and reinforced learning, now we can see how software systems can co-evolve, somewhat similar to what Licklider envisioned in his visionary paper. However, how physical or tangible interfaces can co-evolve with humans over long periods is underexplored. I believe that computational and personal fabrication has a lot to offer in this context. Considering the slow development of hardware, we may need symbiotic software systems that model human behavior and can anticipate or predict future needs. These predictive or anticipatory symbiotic models will help us to create a seamless symbiotic adaptation of physical interfaces such as wearable devices, implantables, and other physical devices. Furthermore, the emerging domain of bio-fabrication could add a lot of value in creating co-evolving systems. In our work, we explore these directions as mechanisms to create co-evolving symbiotic physical interfaces.



■ **Figure 11** Barrett Ens presenting a reiteration of the levels of symbiosis.

4.3 Design of dynamic human-machine coupling system

Azumi Maekawa (The University of Tokyo, JP, azumi@star.rcast.u-tokyo.ac.jp)

License © Creative Commons BY 4.0 International license
© Azumi Maekawa

We humans have had a desire to expand our own abilities through technology since ancient times. With the development of technology, autonomous machines that can achieve physical performance and information processing beyond human abilities have become possible. These autonomous machines have the potential to extend our abilities and fundamentally change our lives. However, it is not yet clear how machines can engage with the dynamic and high-intensity movements of humans, and how humans and machines can function as a unified entity. Due to differences in the characteristics and nature of motion capabilities between humans and machines, simply coupling the two is currently challenging. In this talk, I will introduce prototypes that explore how humans and machines can integrate and harmonize in dynamic motion. I will also briefly present the insights and challenges obtained from this exploration.

4.4 Making sense of information anywhere

Barrett Ens (Monash University – Melbourne, AU, barrett.ens@monash.edu)

License © Creative Commons BY 4.0 International license
© Barrett Ens

The miniaturisation of sensing, networking, and processing technologies has increasingly made information readily available. Taking this further, emerging Augmented Reality (AR) technologies and near-future holographic displays (such as light field and laser plasma displays) will soon allow rich visual information to be displayed anywhere, beyond the confines of small 2D screens. On one hand, these advances will allow relevant information to be more directly integrated with the activities, places or objects to which it is related. However, they will also bring significant challenges in designing useful and productive interfaces for visualising information and interacting with it. Given these coming developments, how can we leverage



■ **Figure 12** The results of one of the breakout groups dedicated toward summarising the grand challenges of human-machine symbiosis based on the conversations throughout the seminar.

spatial interaction and situated information spaces to improve the way we perceive, interact with, and understand information? In this talk I will briefly introduce the motivations for my work on spatial interface design for data exploration and sensemaking.

4.5 Hybrid skins for symbiosis

Cindy Hsin-Liu Kao (Cornell University – Ithaca, USA, cindykao@cornell.edu)

License © Creative Commons BY 4.0 International license
© Cindy Hsin-Liu Kao

Sensor device miniaturization and novel material breakthroughs have enabled hybrid skins as an emerging wearable form. These conformable skin interfaces are hybrid in their integration of technological function with existing cultural body art form factors and often involve social,

design, and engineering challenges in their realization. These hybrid skins create a symbiosis with the human body through their direct skin contact, serving as one of the most intimate physical interactive interfaces in human-computer interaction. While engineering these hybrid skins has received increased interest in the past decade, open questions remain regarding the symbiosis relationship between these hybrid skins and the human wearer. Specifically, how do we design the agency and control of our interactions for, with, and by these hybrid skins? As hybrid skins evolve beyond the skin layer and even onto our bodily organs and cellular surfaces, what does it mean to be human in the age of symbiosis?

4.6 Dermal layers in between the human self and the robot other

Dominika Lisy (Linköping University, Sweden, dominika.lisy@liu.se)

License © Creative Commons BY 4.0 International license
© Dominika Lisy

My interest lies in exploring material boundaries to understand conceptual relations of dualisms and how they intra-actively co-constitute each other (see Barad 1998, 2003, 2007, 2014). In my work with the skin as a feminist figuration, I am exploring different dermal characteristics such as processes of keratinisation, layeredness, and permeability, which allow me to reconfigure dualistic relations as non-oppositional. Being and knowing through the skin makes apparent how boundaries harden/keratinise or dissolve to determine the potentiality of affective mingling with other bodies (see Serres 1985/2008) – also, other “bodies of knowledge”. Their material-discursive touching might provide an ethical space for reconfiguring human-robot-relationalities. In my work I argue that the process of asymmetrical agential cutting (see Suchman 2006) is an embodied ethical practice, and that the figuration of the skin makes apparent how different ways of knowing are layered.

4.7 Designing the human-machine symbiosis? fun with creative technology and design

Ellen Yi-Luen Do (University of Colorado – Boulder, USA, ellen.do@colorado.edu)

License © Creative Commons BY 4.0 International license
© Ellen Yi-Luen Do

A way to Design the Human-Machine Symbiosis is through Fun with Creative Technology and Design. To facilitate cross-disciplinary research we need to provide the environments for creativity, a lab for making things [2]. We build physical and computational artifacts that are “objects” to think with, and this way of working helps us develop methods and tools to make better things. I argued earlier that the Design for Assistive Augmentation should take 3M’s into consideration: Mind, Might, and Magic [3], to have technology wonderfully blended in everyday life activities. Meanwhile, isn’t it time we discuss the art and science of the Human-Machine Symbiosis, and to investigate the engineering and design of such? We could be reflecting (art), understanding, and explaining (science), solving problems (engineering), and inventing and making (design). Let’s remember that the words design and program are remarkably close in their Greek and Latin roots. De (meaning out) and sign (meaning mark) and pro (meaning forward or out) and gram (meaning writing) both mean to mark out or


make an explicit representation. Finally, let me quote Paul Graham here: What hackers and painters have in common is that they're both makers. Along with composers, architects, and writers, what hackers and painters are trying to do is make good things [1].

References

- 1 Paul Graham. *Hackers & painters: big ideas from the computer age*. “ O'Reilly Media, Inc.”, 2004.
- 2 Ellen Yi-Luen Do and Mark D Gross. Environments for creativity: a lab for making things. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*, pages 27–36, 2007.
- 3 Ellen Yi-Luen Do. Design for assistive augmentation – mind, might and magic. *Assistive augmentation*, pages 99–116, 2018.

4.8 Symbiosis is bodily

Florian 'Floyd' Mueller (Monash University – Melbourne, AU, floyd@exertiongameslab.org)

License  Creative Commons BY 4.0 International license
© Florian 'Floyd' Mueller

Symbiosis is concerned with the intertwinedness of the human body and interactive technology. We demonstrate this through a series of research design works around symbiotic cycling experiences, symbiotic entertainment experiences, and symbiotic arts experiences. The results of these works suggest interesting ways forward for symbiosis research, in particular how the design of symbiosis can highlight experiential aspects, facilitating playful experiences. Ultimately, with our work, we want to enhance our knowledge around the design of symbiosis experiences to help people understand who they are, who they want to become, and how to get there.

4.9 Skin as an interface for human-machine symbiosis

Jürgen Steimle (Saarland University – Saarbrücken, DE, steimle@cs.uni-saarland.de)

License  Creative Commons BY 4.0 International license
© Jürgen Steimle

Skin is the largest organ in the human body and the primary interface between the body and its environment. It is through skin contact that we interact with objects; it is through skin that we perceive multi-modal haptic cues; and it is through our skin that we communicate through touch and visual appearance. I argue that skin is a promising platform for human-machine symbiosis, as devices can be deployed in intimate proximity to the human body. I will present recent results from my group's research on ultra-thin, skin-conformal devices that can be ergonomically worn on the skin for physiological sensing, touch interaction and haptic output. Furthermore, my talk will emphasise the importance of computational models and tools for designing wearable devices that are tailored to human anatomical properties, wearability and desired applications.

4.10 Toward identifying features that make human-machine relationships symbiotic

Kumiyo Nakakoji (Future University – Hakodate, JP, kumiyo@fun.ac.jp)

License  Creative Commons BY 4.0 International license
© Kumiyo Nakakoji

By reflecting on six different types of systems and projects that we have worked on over the years, I would like to explore what would be a common framework for the human-machine symbiosis, where individuals and computational environments “evolve” together through their interactions. The “evolution” takes different forms in different contexts. The six types of symbiosis in our versions include: (1) human-computer cooperative problem solving systems, (2) knowledge interaction design for amplifying representational talkback, (3) an online knowledge community where artifacts, stakeholders, the community, and the roles of participants play in the community would evolve in parallel, (4) data experience and engagement platforms for interactive data visualization systems to help users in sense-making and story composition, (5) pseudo-haptics by touch-centric interaction embodiment, where a user perceives what physically does not exist through pseudo-haptics, and finally (6) MR (mixed-reality) world engagement to explore how an individual constructs a congruent image of the “reality” while interacting with a mixed reality environment where physical objects and virtual objects interact with each other.

4.11 Agency-preserving action augmentation using brain-computer interfaces

Lukas Gehrke (Technische Universität Berlin, DE, info@lukasgehrke.com)

License  Creative Commons BY 4.0 International license
© Lukas Gehrke

Advances in hardware that augment users’ motions have reignited dreams of overcoming human limitations, recovering lost abilities as well as simplifying learning new skills. Yet, augmented users report dissociative experiences during or following augmented action. They do not experience agency. To drive adoption of human action augmentation, one of the grand challenges then is to design for agency experience, so users feel as though they are in the “driving seat” once again. One way to preserve users’ sense of agency is by keeping the physical impact on their body in line with their intention to move. Using brain-computer interfaces, our work sets out to preserve the experience of agency by establishing a fast communication channel between the augmentation hardware and users’ brain signals reflecting their intent to act.

4.12 JIZAI body and symbiosis


Masahiko Inami (University of Tokyo, JP, drinami@star.rcast.u-tokyo.ac.jp)

License  Creative Commons BY 4.0 International license
© Masahiko Inami

The harmony of humans and computers is critical, hinging on three crucial aspects. Firstly, cognitive transparency is required, meaning that individuals should be able to manipulate the entire system seamlessly without cognitive barriers, akin to the control one has over their own body. Secondly, we introduce the concept of authority delegation. As Don Norman metaphorically referred to in the context of "horse reins", we should be capable of controlling it as an extension of ourselves when we pull the reins, and seamlessly delegate authority to the computer system when we loosen them. We refer to this as JIZAI-ness. Lastly, the principle of co-emergence comes into play, where a relationship similar to a jazz session is established – the computer system inspires humans and humans inspire the computer system, in turn fostering a more innovative environment. This is integral to the symbiosis of humans and computers.

4.13 Speculation on a world with social digital cyborgs

Nahoko Yamamura (University of Tokyo, JP, yamamura@star.rcast.u-tokyo.ac.jp)

License  Creative Commons BY 4.0 International license
© Nahoko Yamamura

Half a century since the concept of a cyborg was introduced, digital cyborgs, enabled by the spread of wearable robotics, are the focus of much research in recent times. We introduce JIZAI ARMS, a supernumerary robotic limb system consisting of a wearable base unit with six terminals and detachable robotic arms. The system was designed to enable social interaction between multiple wearers, such as an exchange of arm(s), and explore possible interactions between digital cyborgs in a cyborg society. Human augmentation researchers, product designers, system architects and manufacturers have collaborated in an interdisciplinary approach to create a technically complex system while considering the aesthetics of a digital cyborg. As a next step, we have begun to explore the bodily expressions created by social digital cyborgs through dance performance.

4.14 Brain-computer symbiosis

Nathan Semertzidis (Monash University – Melbourne, AU, nathan@exertiongameslab.org)

License  Creative Commons BY 4.0 International license
© Nathan Semertzidis

Nathan Semertzidis is a researcher at the Exertion Games Lab, part of the Department of Human-Centred Computing at Monash University, Australia. Diagnosed with genetic hearing loss, Nathan has depended on hearing aids for most of his life, which from a young age inspired his interest in how technology can be experienced as an extension of the human body. His research focuses on investigating the design of brain-computer integration systems, technologies that both sense and actuate brains to facilitate the artificial extension of the

nervous system. Nathan’s research involves the development, deployment and evaluation of novel brain-computer integration systems through empirical studies, art installations, and game development. Through this Dagstuhl seminar, Nathan draws from his research experience, as well as his own personal relationship with “symbiosis” gained from cultivating corals and ant colonies at home, to ask: “what is human-machine symbiosis and how does it inform brain-computer integration?”

4.15 Tools, medium, mediator, partner, and beyond...

Sheng-Fen ‘Nik’ Chien (National Cheng Kung University – Tainan, TW, schien@mail.ncku.edu.tw)

License © Creative Commons BY 4.0 International license
© Sheng-Fen ‘Nik’ Chien

New technologies are often introduced as tools to support works, or complement disabilities. Tools can become extensions of us: those we act with, as well as those we think with, i.e. mediums. Artists are often pioneers to demonstrate radical uses of new technologies: those as mediators of audiences and artworks, as well as those as partners in their art-creation process. These radical uses provide foresight, which in turn initiates reflections on the impacts of new technologies, and brings insights of new social development. Recent advances in computing technologies, in particular neural network based machine learning, warrant a revisit of “Man-Computer Symbiosis” envisioned by Licklider in 1960. For the Dagstuhl Seminar, I plan to take the stance of creative reflective practitioners (artists) to explore definitions and implications of human-computer symbiosis.

4.16 Understanding games and play in a posthuman era

Oğuz ‘Oz’ Buruk (Tampere University, FI, oguz.buruk@tuni.fi)

License © Creative Commons BY 4.0 International license
© Oğuz ‘Oz’ Buruk

With the rapid advances in technology, artificial intelligence, robotic companions, bodily integrated technologies, brain-machine interfaces, and space habituation technologies are now here or on the horizon. When integrated fully at the societal level, these technologies will profoundly impact how we live and experience the world around us. Can we try to understand these futures through an antisolutionist approach by using games and play as a central lens? What will games and play be like in a posthuman age? What can designing games and play of a posthuman era tell us about future societies? Answering these questions and examining games and play in posthuman technofutures by utilizing antisolutionist design methods can reveal much about the experiences that await us in the future. In other words, through speculative and critical design methods such as design fiction, fictional probes, pastiche scenarios, creating fictional narratives, worlds and prototypes demonstrating how games and play will be like in posthuman futures will help us understand the experiential texture of symbiotic living with computers.

4.17 How will people live symbiotically with AI?

Pattie Maes (Massachusetts Institute of Technology – Cambridge, USA, pattie@media.mit.edu)

License  Creative Commons BY 4.0 International license
© Pattie Maes

The advent of AI forces us to reimagine human machine symbiosis at the Cognitive level. Symbiotic systems to date have often targeted bodily symbiosis, eg of senses and motor systems, but we are now facing a historic, unprecedented opportunity to design the symbiosis of the computer and human minds. How might AI systems augment us rather than replace us? What knowledge and skills should humans still learn to internalize, even though AI systems could easily do the job for them? How can AI systems be designed in a human centric way, with the goal of supporting people in becoming the person they'd like to be? Our research group at the MIT Media Lab has for a while now been focused on these questions, building symbiotic systems for reasoning, learning, memory, and more, but we have only scratched the surface of this important challenge, which I believe may well be one of the hardest challenges of this era.

4.18 Machine poetics

Sang-won Leigh (Samsung Design Innovation Center – San Francisco, USA & Georgia Tech – Atlanta, USA, sang.leigh@design.gatech.edu)

License  Creative Commons BY 4.0 International license
© Sang-won Leigh

Throughout history we have augmented our physical abilities with machines. Concepts for flying machines and today's exoskeletons were recorded as early as the 13th century. Today, as technology permeates every aspect of our lives, it is easy to imagine a much closer integration of machines into the tasks we carry out, or even our own existence. There would be extreme synergies between machine tools and humans, with technology essentially becoming a natural extension of our bodies. This, what I would call a symbiosis, requires a close examination given how we are already incredibly influenced by technological systems surrounding us, and how deep of a physical, cognitive or existential influence it could make to ourselves. Instead of rehashing existing lifestyles and system concepts into a word symbiosis, the exploration may need to be rooted in a critical, experiential investigation both looking at psychosomatic experience and alternative realities of human-machine symbiosis.

4.19 Human-Machine symbiosis via a mixed reality

Yi Fei Cheng (Carnegie Mellon University – Pittsburgh, USA, yifeic2@andrew.cmu.edu)

License  Creative Commons BY 4.0 International license
© Yi Fei Cheng

Mixed reality (MR) technologies promise to transform how we interact with digital information. With its capabilities to directly integrate virtuality into physical reality and blur the digital-physical divide, it serves as a promising platform to enable humans and computers to enter a symbiosis. However, while MR may enrich our lived experiences with playful integrations

of virtuality or beneficially provide just-in-time digital information to supplement decision making and communication, it also has the potential to detrimentally distract and overwhelm. Enabling a seamless integration of virtuality is therefore just the first step in enabling symbiosis. What is equally critical is to ensure that we can achieve a careful balance between virtual and physical reality. I believe significant work lies ahead in both exploring new ways for more seamless integrations of virtuality and reality and designing technologies that always retain a beneficial balance between the two.

4.20 Human-machine symbiosis

Zhuying Li (Southeast University – Nanjing, CN, zhuying9405@gmail.com)

License  Creative Commons BY 4.0 International license
© Zhuying Li

The rapid development of sensing and actuating technologies has unlocked unprecedented opportunities to blur and extend the boundaries of the human body through the integration of such technologies. In my talk, I will show my works of ingestible play where players swallow an ingestible sensor and play with their interior body data collected by the sensor. Such a play genre allows people to experience having a foreign technology inside their physical body and engaged with their bodily information they could not have known before, which might increase their bodily awareness and understanding. Furthermore, the recent advancements in AI technologies offer us the prospect of forming a sensory or cognitive symbiotic relationship with machines. For example, AI-powered mobile devices can augment our sensory channel by continuously gathering information from our bodies and surrounding environments, providing feedback that we may otherwise overlook. In this talk, I will introduce an AI-powered wearable device that can identify birds based on their songs, providing users with haptic feedback to enhance their environmental awareness and engage them with the natural environment. Moreover, I will demonstrate how AI can augment our cognitive processes, such as decision-making and reasoning. In support of this, I will share our latest work – an arm-worn exoskeleton designed to make certain decisions for the user, effectively assuming control over the user’s arm to perform specific actions.

Participants

- Andrea Bianchi
KAIST – Daejeon, KR
- Oguz Buruk
University of Tampere, FI
- Yi Fei Cheng
Carnegie Mellon University –
Pittsburgh, US
- Sheng-Fen Chien
National Cheng Kung University
– Tainan, TW
- Valdemar Danry
MIT – Cambridge, US
- Ellen Yi-Luen Do
University of Colorado –
Boulder, US
- Barrett Ens
Monash University –
Clayton, AU
- Zoe Xiao Fang
Zhejiang University –
Hangzhou, CN
- Lukas Gehrke
TU Berlin, DE
- Masahiko Inami
University of Tokyo, JP
- Cindy Hsin-Liu Kao
Cornell University – Ithaca, US
- Sang Leigh
Samsung Research America –
San Francisco, US & Georgia
Institute of Technology –
Atlanta, US
- Zhuying Li
Southeast University –
Nanjing, CN
- Dominika Lisy
Linköping University, SE
- Azumi Maekawa
University of Tokyo, JP
- Pattie Maes
MIT – Cambridge, US
- Florian ‘Floyd’ Mueller
Monash University –
Clayton, AU
- Kumiyo Nakakoji
Future University –
Hakodate, JP
- Nathan Semertzidis
Monash University –
Clayton, AU
- Jürgen Steimle
Universität des Saarlandes –
Saarbrücken, DE
- Don Anusha Withanage
The University of Sydney, AU
- Nahoro Yamamura
University of Tokyo, JP



Computational Geometry

Siu-Wing Cheng^{*1}, Maarten Löffler^{*2}, Jeff M. Phillips^{*3}, and Aleksandr Popov^{†4}

1 HKUST – Hong Kong, CN. scheng@cse.ust.hk

2 Utrecht University, NL & Tulane University – New Orleans, LA, US.
m.loffler@uu.nl

3 University of Utah – Salt Lake City, UT, US. jeffp@cs.utah.edu

4 TU Eindhoven, NL. a.popov@tue.nl

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 23221 “Computational Geometry”. The seminar was held from May 29th to June 2nd, 2023, and 39 participants from various countries attended it, including two remote participants. Recent advances in computational geometry were presented and discussed, and new challenges were identified. This report collects the abstracts of the talks and the open problems presented at the seminar.

Seminar May 29 – June 2, 2023 – <https://www.dagstuhl.de/23221>

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Design and analysis of algorithms; Mathematics of computing → Discrete mathematics

Keywords and phrases Algorithms, Combinatorics, Geometric Computing, Reconfiguration, Uncertainty

Digital Object Identifier 10.4230/DagRep.13.5.165

1 Executive Summary

Siu-Wing Cheng (HKUST – Hong Kong, CN, scheng@cse.ust.hk)

Maarten Löffler (Utrecht University, NL & Tulane University – New Orleans, US, m.loffler@uu.nl)

Jeff M. Phillips (University of Utah – Salt Lake City, US, jeffp@cs.utah.edu)

License  Creative Commons BY 4.0 International license
© Siu-Wing Cheng, Maarten Löffler, and Jeff M. Phillips

This Dagstuhl Seminar constituted a biennial gathering of computational geometers at the Dagstuhl venue to share recent results, and further research on some of the most important problems of the time in that field. This year, the seminar focused on two of the most exciting sub-areas within computational geometry: (1) reconfiguration, and (2) processing and applications of uncertain and probabilistic geometric data. Within the reconfiguration topic, two overview talks focused on triangulation and graph reconfiguration, and on how reconfiguration plays a role in puzzle complexity. A highlight of the seminar were the set of three-dimensional reconfiguration puzzles brought by Ryuhei Uehara, which occupied attendees endlessly, and brought the challenge of modelling such puzzles to life. In the second theme on uncertainty, one overview talk covered uncertainty issues in spatial data, and another focused on how uncertainty connects to differential privacy in geometric settings.

* Editor / Organizer

† Editorial Assistant / Collector



Other results were shared by participants connecting these topics to diverse motivations ranging from robotics to data analysis to graph drawing. Exciting open problems were proposed, and were used to focus the discussion for the span of the seminar.

2 Table of Contents

Executive Summary

Siu-Wing Cheng, Maarten Löffler, and Jeff M. Phillips 165

Overview of Talks

Uncertain Points and Trajectories

Kevin Buchin 169

Recent Progress in Geometric Random Walks and Sampling

Ioannis Emiris 169

Geometric Reconfiguration: Triangulations, Spanning Trees, Graphs

Anna Lubiw 170

Oblivious Sketching for Sparse Linear Regression

Alexander Munteanu 170

Parameterized Algorithm for the Planar Disjoint Path Problem

Eunjin Oh 171

Homology of Reeb Spaces and the Borsuk–Ulam Theorem

Salman Parsa 171

Random Projections for Curves in High Dimensions

Ioannis Psarros 172

Using SAT Solvers in Combinatorics, Combinatorial Geometry, and Graph Drawing

Manfred Scheucher 172

Modular Robot Reconfiguration: Sliding Squares

Willem Sonke 173

Deep Neural Network Training Acceleration with Geometric Data Structures

Konstantinos Tsakalidis 173

Computational Complexity of Puzzles (In the Context of Reconfiguration)

Ryuhei Uehara 174

Combinatorial Reconfiguration via Triangle Flips

Birgit Vogtenhuber 174

Differential Privacy and Computational Geometry

Ke Yi 175

Open Problems

Largest Precise Subset (Making Points Precise)

Peyman Afshani 176

Partial Vertex Cover in Planar Bipartite Graphs

Peyman Afshani, Kevin Buchin, Fabian Klute, Willem Sonke, and Ryuhei Uehara 177

Flipping Spanning Trees

Anna Lubiw 177

Shortest Path Retaining k Obstacles

Subhash Suri 178

168 23221 – Computational Geometry

Why Are Polytime and NP-Hard Switched for Fréchet and Hausdorff? <i>Carola Wenk</i>	179
Participants	180
Remote Participants	181

3 Overview of Talks

This section contains short abstracts of the various talks that were given throughout the seminar. Four speakers (K. Buchin, A. Lubiw, R. Uehara, and K. Yi) were invited to give an overview of the state of the art and to highlight challenges in the two seminar themes, and helped set the stage for the rest of the seminar. In addition, nine participants gave shorter talks throughout the week, which helped exchange ideas and focus the discussions.

3.1 Uncertain Points and Trajectories

Kevin Buchin (TU Dortmund, DE, kevin.buchin@tu-dortmund.de)

License © Creative Commons BY 4.0 International license
© Kevin Buchin

Location data often comes with some error. For instance, the measurements might be imprecise, or the exact location might be obfuscated to preserve privacy. Nonetheless, most algorithms treat these data as if the exact locations are known. The computational geometry of uncertain points is about designing algorithms that deal with the uncertainty explicitly.

In my talk, I first present how uncertain points are commonly modelled and give an overview of algorithmic results on uncertain points for a variety of geometric problems. Then I focus on uncertain trajectories and discuss in particular the problem of computing the Fréchet distance between uncertain curves in detail. I finish my talk with a discussion of open challenges.

3.2 Recent Progress in Geometric Random Walks and Sampling

Ioannis Z. Emiris (Athena Research Center – Marousi, GR & National and Kapodistrian University of Athens – Zografou, GR, emiris@athenarc.gr)

License © Creative Commons BY 4.0 International license
© Ioannis Emiris

Joint work of Ioannis Emiris, Apostolos Chalkis, Vissarion Fisikopoulos

We overview recent advances in geometric random walks for producing a sample of points in convex polyhedra, and applications of such random samples. The main motivation comes from polynomial-time randomized approximation schemes for computing the volume of H-polytopes, i.e. presented as intersection of halfspaces; our methods tackle inputs whose dimension is in the thousands. We extend these algorithms to V-polytopes and zonotopes and present algorithmic results and implementations that show that in practice such polytopes can also be handled efficiently in dimension around 100. Here we have used the existing random walks but optimized the sequence of bodies employed in the standard multiphase Monte Carlo approach. The second part of the talk concentrates on non-linear bodies such as spectrahedra, and an application in systems biology. We conclude with open questions on employing these methods to optimization, and to devising efficient approximate polytope oracles.

3.3 Geometric Reconfiguration: Triangulations, Spanning Trees, Graphs

Anna Lubiw (University of Waterloo, CA, alubiw@uwaterloo.ca)

License  Creative Commons BY 4.0 International license
 © Anna Lubiw

I discuss recent results and open questions on three topics in geometric reconfiguration.


1. Reconfiguring triangulations of a set of n points in the plane. A flip, or reconfiguration step, replaces one edge by another to give a new triangulation. It is known that if some edges are *fixed* (i.e. required to be present in all the triangulations) then reconfiguration is still possible via the constrained Delaunay triangulation. I show some results about *forbidding* some edges. Reconfiguration is no longer always possible, i.e. the flip graph may become disconnected, even by forbidding a single edge. However, for points in convex position, we must forbid $n - 3$ edges in order to disconnect the flip graph.
2. Reconfiguring non-crossing spanning trees on a set of n points in the plane. A flip replaces one edge by a new edge to give a new non-crossing spanning tree. Reconfiguring one non-crossing spanning tree to another takes at most $2n$ flips [1] and (in the worst case) at least $1.5n$ flips, even for points in convex position. I show some improvements in the upper bound for points in convex position, and for some cases where one tree is a path. There are still gaps, and the complexity of finding the minimum flip distance (in P or NP-complete) is open.
3. Reconfiguring planar graph drawings by moving the points (aka morphing). One straight-line planar drawing of an n -vertex graph can be morphed to another (with the same combinatorial embedding) via $\mathcal{O}(n)$ *unidirectional morphs* that move vertices along parallel lines at uniform speeds. If the original drawings lie on a small grid, can the $\mathcal{O}(n)$ intermediate drawings be constrained to a small grid?

References

- 1 David Avis and Komei Fukuda. Reverse Search for Enumeration. *Discrete Applied Mathematics*, 65(1–3):21–46, 1996. doi:10.1016/0166-218X(95)00026-N.

3.4 Oblivious Sketching for Sparse Linear Regression

Alexander Munteanu (TU Dortmund, DE, alexander.munteanu@tu-dortmund.de)

License  Creative Commons BY 4.0 International license
 © Alexander Munteanu

Joint work of Alexander Munteanu, Tung Mai, Cameron Musco, Anup B. Rao, Chris Schwiegelshohn, David P. Woodruff

Main reference Tung Mai, Alexander Munteanu, Cameron Musco, Anup Rao, Chris Schwiegelshohn, David Woodruff: “Optimal Sketching Bounds for Sparse Linear Regression”, in Proc. of The 26th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, Vol. 206, pp. 11288–11316, PMLR, 2023.

URL <https://proceedings.mlr.press/v206/mai23a.html>

Oblivious sketching enables efficient algorithms for analysing data streams and distributed data by reducing the number of data points while preserving a $(1 + \varepsilon)$ -approximation for various regression loss functions such as l_p -norms, or logistic loss. For dense models, a sketching complexity of $\Omega(d)$ is immediate. For very high-dimensional data, model sparsity is a common assumption, where we do not regress on all, but only on at most $k \ll d$ dimensions. While the computational complexity for solving this problem becomes hard, the assumption allows us to sketch to a smaller $o(d)$ size. We show that reducing to $\frac{k \log(d/k)}{\varepsilon^2}$ is

essentially optimal for any sketching algorithm under several regression losses, combining high-dimensional probability and geometry, information-theoretic arguments, and metric embedding theory.

3.5 Parameterized Algorithm for the Planar Disjoint Path Problem

Eunjin Oh (POSTECH – Pohang, KR, eunjin.oh@postech.ac.kr)

License © Creative Commons BY 4.0 International license
© Eunjin Oh

Joint work of Eunjin Oh, Kyeongjin Cho, Seunghyeok Oh

Main reference Kyungjin Cho, Eunjin Oh, Seunghyeok Oh: “Parameterized Algorithm for the Disjoint Path Problem on Planar Graphs: Exponential in k^2 and Linear in n ”, in Proc. of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023, pp. 3734–3758, SIAM, 2023.

URL <https://doi.org/10.1137/1.9781611977554.ch144>

In this talk, I present a parameterized algorithm for the planar disjoint paths problem. Given a planar graph $G = (V, E)$ and a set $T = \{(s_1, t_1), \dots, (s_k, t_k)\}$ of terminal pairs, the goal is to compute a set of vertex-disjoint paths, each connecting s_i and t_i . This problem is NP-hard if we measure the running time as a function of the input size, but if we measure the running time as a function of n and k , we can achieve a non-trivial bound. In this talk, I give a sketch of this algorithm with running time of $2^{\mathcal{O}(k^2)}n$. This improves the two previously best known algorithms running in $2^{\mathcal{O}(k^2)}n^6$ and $2^{2^{\mathcal{O}(k)}}n$ time, respectively.

3.6 Homology of Reeb Spaces and the Borsuk–Ulam Theorem

Salman Parsa (DePaul University – Chicago, IL, US, s.parsa@depaul.edu)

License © Creative Commons BY 4.0 International license
© Salman Parsa

Joint work of Salman Parsa, Sarah Percival

In this talk, I prove an extension of the Borsuk–Ulam theorem for maps from 2-sphere into \mathbb{R} . The extension says that there are always two antipodal points $S^2 \ni x, -x$ such that $f(x) = f(-x)$ and the two points are connected in the preimage.

The proof uses the concept of the Reeb graph. We also consider the relationship between extra homology of the Reeb space of $f : S^n \rightarrow \mathbb{R}^{n-1}$ and the existence of the analogous extensions of the Borsuk–Ulam theorem.

3.7 Random Projections for Curves in High Dimensions

Ioannis Psarros (Athena Research Center – Marousi, GR, ipsarros@athenarc.gr)

License © Creative Commons BY 4.0 International license
© Ioannis Psarros

Joint work of Ioannis Psarros, Dennis Rohde

Main reference Ioannis Psarros, Dennis Rohde: “Random Projections for Curves in High Dimensions”, in Proc. of the 39th International Symposium on Computational Geometry, SoCG 2023, June 12-15, 2023, Dallas, Texas, USA, LIPIcs, Vol. 258, pp. 53:1–53:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPIcs.SocG.2023.53>

Modern time series analysis requires the ability to handle datasets that are inherently high-dimensional; examples include applications in climatology, where measurements from numerous sensors must be taken into account, or inventory tracking of large shops, where the dimension is defined by the number of tracked items. The standard way to mitigate computational issues arising from the high dimensionality of the data is by applying some dimension reduction technique that preserves the structural properties of the ambient space. The dissimilarity between two time series is often measured by “discrete” notions of distance, e.g. the dynamic time warping or the discrete Fréchet distance. Since all these distance functions are computed directly on the points of a time series, they are sensitive to different sampling rates or gaps. The continuous Fréchet distance offers a popular alternative which aims to alleviate this by taking into account all points on the polygonal curve obtained by linearly interpolating between any two consecutive points in a sequence.

We study the ability of random projections à la Johnson and Lindenstrauss to preserve the continuous Fréchet distance of polygonal curves by effectively reducing the dimension.

3.8 Using SAT Solvers in Combinatorics, Combinatorial Geometry, and Graph Drawing

Manfred Scheucher (TU Berlin, DE, scheucher@math.tu-berlin.de)

License © Creative Commons BY 4.0 International license
© Manfred Scheucher

In this talk, we discuss how modern SAT solvers can be used to tackle mathematical problems. We discuss various problems to give the audience a better understanding, which might be tackled in this fashion, and which might not. Besides the naïve SAT formulation further ideas are sometimes required to tackle problems. Additional constraints such as statements which hold “without loss of generality” might need to be added so that solvers terminate in reasonable time.

3.9 Modular Robot Reconfiguration: Sliding Squares

Willem Sonke (TU Eindhoven, NL, w.m.sonke@tue.nl)

License © Creative Commons BY 4.0 International license
© Willem Sonke

Joint work of Willem Sonke Hugo A. Akitaya, Erik D. Demaine, Matias Korman, Irina Kostitsyna, Irene Parada, Bettina Speckmann, Ryuhei Uehara, Jules Wolms

Main reference Hugo A. Akitaya, Erik D. Demaine, Matias Korman, Irina Kostitsyna, Irene Parada, Willem Sonke, Bettina Speckmann, Ryuhei Uehara, Jules Wolms: “Compacting Squares: Input-Sensitive In-Place Reconfiguration of Sliding Squares”, in Proc. of the 18th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2022, June 27-29, 2022, Tórshavn, Faroe Islands, LIPIcs, Vol. 227, pp. 4:1–4:19, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <https://doi.org/10.4230/LIPIcs.SWAT.2022.4>

Modular robots consist of a large number (say n) of small identical modules that can move along each other to change the overall shape of the robot. A well-established model for modular robots is called sliding squares. Here, each module is represented by a square in the 2D grid, which form an edge-connected configuration. Modules can perform two types of moves: slides and convex transitions.

The main question now is: given a source and target configuration, can we find a short move sequence to transform the source into the target? In this talk I show an algorithm named Gather&Compact that finds such a move sequence of length $\mathcal{O}(Pn)$ where P is the circumference of the configurations’ bounding box. Furthermore, I show that it is NP-hard to find a move sequence of minimum length.

3.10 Deep Neural Network Training Acceleration with Geometric Data Structures

Konstantinos Tsakalidis (University of Liverpool, GB, tsakalid@liverpool.ac.uk)

License © Creative Commons BY 4.0 International license
© Konstantinos Tsakalidis

The efficiency of deep learning applications deteriorates significantly as the sizes of the training data and of the neural networks grow larger. In this talk we identify beyond-state-of-the-art open problems in the intersection of deep learning with computational geometry. Motivated by the recent application of dynamic data structures for geometric halfspace range searching in the acceleration of deep neural networks’ training and preprocessing complexity, we revisit efficient algorithms for constructing geometric multi-dimensional data structures and maintaining them dynamically.

3.11 Computational Complexity of Puzzles (In the Context of Reconfiguration)

Ryuhei Uehara (JAIST – Nomi, Ishikawa, JP, uehara@jaist.ac.jp)

License © Creative Commons BY 4.0 International license
© Ryuhei Uehara

Main reference Ryuhei Uehara: “Computational Complexity of Puzzles and Related Topics”, *Interdisciplinary Information Sciences*, Vol. 29(2), pp. 119–140, 2023.

URL <https://doi.org/10.4036/iis.2022.R.06>

I first give a short history of computational complexity of puzzles and games, including combinatorial reconfiguration ones. In this context, there are three open problems.

1. The Rubik’s cube has a similar property to one of the $n^2 - 1$ puzzles. Then what is the counterpart of the sliding block puzzle?
 - Is there some PSPACE-complete problem in general?
 - What about rectangular faces?
2. Reconfiguration of triangulations of a simple polygon also has a similar property. Then, again, can we have a counterpart of the sliding block puzzle that leads us to PSPACE-complete variant in general?
 - How about non-simple polygon with holes?
 - Can the diameter of the configuration space be super-poly? (Otherwise, it is in NP since we have a poly-length witness.)
3. Computational complexity of *slide-and-pack* puzzles. By some observations, it seems to be PSPACE-complete in general. Do we have some tractable restrictions?

3.12 Combinatorial Reconfiguration via Triangle Flips

Birgit Vogtenhuber (TU Graz, AT, bvogt@ist.tugraz.at)

License © Creative Commons BY 4.0 International license
© Birgit Vogtenhuber

Joint work of Birgit Vogtenhuber, Oswin Aichholzer, Man-Kwun Chiu, Stefan Felsner, Hung P. Hoang, Michael Hoffmann, Yannic Maus, Johannes Obenaus, Sandro Roch, Manfred Scheucher, Alexandra Weinberger

In this talk we discuss the reconfiguration of arrangements of simple curves in the plane or on the sphere via triangle flips (a.k.a. Reidemeister moves of Type 3). This operation refers to the act of moving one edge of a triangular cell formed by three pairwise crossing curves over the opposite crossing of the cell, via a local transformation. We study two types of arrangements, namely, arrangements of pseudocircles in the plane and simple drawings of graphs on the sphere.

An arrangement of pseudocircles is a finite collection of simple closed curves in the plane such that every pair of curves is either disjoint or intersects in two crossing points. We show that triangle flips induce a connected flip graph on intersecting arrangements, i.e. on arrangements where every pair of pseudocircles intersects. To obtain this result, we first show that every intersecting arrangement can be flipped into some cylindrical arrangement, i.e. an arrangement where a single point stabs the interior of every pseudocircle. Then we show that the cylindrical arrangement can be flipped to a canonical arrangement, which also shows the connectivity of cylindrical intersecting arrangements of pseudocircles under triangle flips. With a careful analysis we obtain that the diameter of both flip graphs is cubic in

the number of pseudocircles. The construction of the two flipping sequences makes essential use of variants of the sweeping lemma for pseudocircle arrangements due to Snoeyink and Hershberger [1].

A simple drawing of a labelled graph is a drawing in which the vertices are distinct points and the edges are simple curves connecting their endpoints. Moreover, any two edges share at most one point, which is either a common endpoint or a crossing. The extended rotation system of such a drawing is the collection of the rotations of all vertices and crossings of the drawing. The rotation of a vertex or crossing is the cyclic order in which the edges emanate from it. Gioan's Theorem states that for any two simple drawings of the complete graph K_n with the same crossing edge pairs, one drawing can be transformed into the other by a sequence of triangle flips. We investigate to what extent Gioan-type theorems can be obtained for wider classes of graphs. A necessary (but in general not sufficient) condition for two drawings of a graph to be transformable into each other by a sequence of triangle flips is that they have the same ERS. We show that for the large class of complete multipartite graphs, this necessary condition is in fact also sufficient and give bounds on the diameter of the resulting flip graph. We further show that this result is essentially tight in the sense that there exist drawings of multipartite graphs plus one edge or minus two edges which cannot be transformed into each other via triangle flips.

References

- 1 Jack Scott Snoeyink and John E. Hershberger. Sweeping Arrangements of Curves. In *Proceedings of the 5th Annual Symposium on Computational Geometry (SCG 1989)*, pages 354–363, New York, NY, US, 1989. Association for Computing Machinery. doi:10.1145/73833.73872.

3.13 Differential Privacy and Computational Geometry

Ke Yi (HKUST – Hong Kong, CN, yike@cse.ust.hk)

License  Creative Commons BY 4.0 International license
© Ke Yi


Differential privacy has become the de facto standard for personal information privacy, and has been recently widely adopted in both governments and the industry. Roughly speaking, a differentially private algorithm should have indistinguishable output distributions on neighbouring instances, which differ by one individual's data. Such a definition also applies to geometric data, where one input point set has one more point than the other. In this talk, I give an overview of existing differentially private algorithms for geometric data, including range counting, mean estimation, and convex hull. I also discuss geometric privacy, a variant of differential privacy that is more suitable for certain geometric problems.

4 Open Problems

This section contains short summaries of specific relevant open questions that were proposed at the start of the seminar and discussed in more depth during the week. Partial progress was made towards resolving these questions, and we expect there will be further collaboration on these topics between seminar participants beyond the duration of the seminar.

4.1 Largest Precise Subset (Making Points Precise)

Peyman Afshani (Aarhus University, DK, peyman@cs.au.dk)

License  Creative Commons BY 4.0 International license
© Peyman Afshani

Imprecision through Precision

In geometric computation, it is useful to be able to answer geometric predicates, e.g. a sidedness test with respect to points in the plane. To model imprecision, it is common to replace points with disks in 2D.

First we fix our geometric predicate. If we replace every point with an imprecise point modelled as a circle, then the test can still be resolved on the points if there is no line that intersects three circles; let S be the set of input circles. We call S a *precise point set* (w.r.t. the sidedness test). The motivation is that any algorithm that can use only sidedness test can be run without any modifications on the set S .

Some Open Problems

Selecting a large precise set. One open question could be to select a large precise subset of an imprecise points. Let S be a set of circles such that any line intersects at most ℓ circles. What is the largest precise subset of $H \subset S$ one can select, so that no three circles in H should be stabbed by any line?

- Using standard techniques (random sample and refine), it is easy to show that we can take $H = \Omega(\sqrt{n}/\ell)$.
- There are also previous results on similar questions but when input is a set of points rather than circles. This line of work is studied under the name of *general position subset selection problem*. For example, it is known that given a set of n points such that no ℓ of them are on a line, one can select a subset of size $\Omega(\sqrt{n/\log \ell})$ that is in general position [1]. Observe that this bound is much better than the our trivial bound. However, to prove it, the authors use incidence bounds that we do not have for a set of circles.

Open question 1. Improve the bound for the circles.

Open question 2. Motivated by the techniques in the above mentioned paper, we can also ask the following question. Let S be a set of circles such that no ℓ of them are on a line. What is the maximum number of triples of circles (c_i, c_j, c_k) such that all three can be stabbed by a line? The trivial upper bound is $\mathcal{O}(n^2\ell^3)$, because we have at most $\mathcal{O}(n^2)$ combinatorially different lines and each line can stab ℓ , so we can select $\mathcal{O}(\ell^3)$ triples out of them. The trivial lower bound is $\Omega(\min(n^2, n\ell^2))$ obtained by either using a $(3 \times n/3)$ -grid or placing n/ℓ groups of ℓ circles on a line.

Open question 3. For the general position subset selection problem, there exists an $o(n)$ upper bound through Hales–Jewett theorem (referenced in the paper [1]). Can we improve the upper bound for the circles? Note that for this upper bound problem, having the possibility of replacing a point by a circle should make the upper bound problem easier.

References

- 1 Michael S. Payne and David R. Wood. On the General Position Subset Selection Problem. *SIAM Journal on Discrete Mathematics*, 27(4):1727–1733, 2013. doi:10.1137/120897493.

4.2 Partial Vertex Cover in Planar Bipartite Graphs

Peyman Afshani (Aarhus University, DK, peyman@cs.au.dk)

Kevin Buchin (TU Dortmund, DE, kevin.buchin@tu-dortmund.de)

Fabian Klute (UPC Barcelona Tech, ES, fmklute@gmail.com)

Willem Sonke (TU Eindhoven, NL, w.m.sonke@tue.nl)

Ryuhei Uehara (JAIST – Nomi, Ishikawa, JP, uehara@jaist.ac.jp)

License  Creative Commons BY 4.0 International license
 © Peyman Afshani, Kevin Buchin, Fabian Klute, Willem Sonke, and Ryuhei Uehara

Partial Vertex Cover asks whether we can cover at least t edges by selecting k vertices. In contrast to vertex cover, this problem is NP-hard in bipartite graphs [1]. How about planar bipartite graphs?

It seems that the paper by Caskurlu et al. [1] asks this as an open problem, although “bipartite” is missing in the problem statement.

This problem is inspired by the obstacle-deletion problem by Subhash Suri. If this problem is NP-hard, we could use this fact in the following way for the obstacle-deletion open problem. Since the graph is bipartite, we can find a path in the plane (not in the graph) that crosses every edge exactly once. Now we interpret every vertex as obstacles, and for every edge, we make the corresponding obstacles interlock where the path would like to pass. Now partial vertex cover corresponds to allowing to delete k obstacles while trying to get rid of as many interlockings as possible.

For the obstacle-deletion problem we can give weights to vertices, we can also give weights to edges. Therefore, even weighted versions of partial vertex cover in planar bipartite graphs would be of interest.

References

- 1 Bugra Caskurlu, Vahan Mkrtchyan, Ojas Parekh, and K. Subramani. Partial Vertex Cover and Budgeted Maximum Coverage in Bipartite Graphs. *SIAM Journal on Discrete Mathematics*, 31(3):2172–2184, 2017. doi:10.1137/15M1054328.

4.3 Flipping Spanning Trees

Anna Lubiw (University of Waterloo, CA, alubiw@uwaterloo.ca)

License  Creative Commons BY 4.0 International license
 © Anna Lubiw

During the seminar, the following question from the talk given by Anna Lubiw was investigated. Given a set P of n points in the plane and two non-crossing spanning trees T_r, T_b on P , what is their flip distance $\text{dist}(T_r, T_b)$ in terms of n in the worst case?

The following bounds are known:

1. lower bound of $\frac{3}{2} \cdot n - 5$ [3];
2. upper bound of $2n - 4$ [1];
3. upper bound of $2n - \Omega(\sqrt{n})$ if P is in convex position [2].

Can these bounds be improved? Are there interesting special cases with tighter bounds?

References

- 1 David Avis and Komei Fukuda. Reverse Search for Enumeration. *Discrete Applied Mathematics*, 65(1–3):21–46, 1996. doi:10.1016/0166-218X(95)00026-N.
- 2 Nicolas Bousquet, Valentin Gledel, Jonathan Narboni, and Théo Pierron. A Note on the Flip Distance between Non-crossing Spanning Trees. 2023. arXiv:2303.07710v1 [cs.CG].
- 3 M. Carmen Hernando, Ferrán Hurtado, Alberto Márquez, Mercè Mora, and Marc Noy. Geometric Tree Graphs of Points in Convex Position. *Discrete Applied Mathematics*, 93(1):51–66, 1999. doi:10.1016/S0166-218X(99)00006-2.

4.4 Shortest Path Retaining k Obstacles

Subhash Suri (University of California – Santa Barbara, US, suri@cs.ucsb.edu)

License  Creative Commons BY 4.0 International license
 © Subhash Suri

Suppose we are in polygonal domain in the plane with n vertices in total. The obstacles (holes) do not have to be convex. Given two points s and t , what is the Euclidean length of the shortest path between s and t that may ignore k obstacles? In other words, if we are allowed to remove k obstacles, what is the length of the shortest obstacle-free path between s and t ? Formulated differently, let \mathcal{B} be the set of m disjoint simple polygons that are the obstacles. Define $d_{\mathcal{B}'}(s, t)$ as the shortest path distance from s to t with only $\mathcal{B}' \subseteq \mathcal{B}$ present. For given k and s and t , we wish to find

$$d_k(s, t) = \min_{|\mathcal{B} \setminus \mathcal{B}'|=k} d_{\mathcal{B}'}(s, t).$$


It is known how to compute the solution in polynomial time for convex obstacles [2]. A similar problem with overlapping obstacles is known to be intractable even for very simple obstacle shapes [2]. The weighted problem, where each obstacle has a cost and we are given a budget, is NP-hard even for vertical line segments as obstacles [1]. We would like to find out if the problem presented here is hard.

References

- 1 Pankaj K. Agarwal, Neeraj Kumar, Stavros Sintos, and Subhash Suri. Computing Shortest Paths in the Plane with Removable Obstacles. In *Proceedings of the 16th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2018)*, number 101 in Leibniz International Proceedings in Informatics (LIPIcs), article 5, Dagstuhl, DE, 2018. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.SWAT.2018.5.
- 2 John E. Hershberger, Neeraj Kumar, and Subhash Suri. Shortest Paths in the Plane with Obstacle Violations. *Algorithmica*, 82:1813–1832, 2020. doi:10.1007/s00453-020-00673-y.

4.5 Why Are Polytime and NP-Hard Switched for Fréchet and Hausdorff?

Carola Wenk (Tulane University – New Orleans, US, cwenk@tulane.edu)

License  Creative Commons BY 4.0 International license
© Carola Wenk

Consider a common definition of an uncertain trajectory as a sequence of uncertain points, where each uncertain point is some compact connected region, most commonly a disk in \mathbb{R}^2 or an interval in \mathbb{R} . A true location must be inside the region, but we do not know where. In this setting, it is natural to generalise the polyline metrics to ask for the maximum and minimum possible values. In particular, we say a polyline realises an uncertain trajectory if it is formed by a sequence of precise points, taken in the correct order from the uncertainty regions. Then we can ask for the lower bound and the upper bound distance, that is, given two uncertain curves, we want to find the minimum and the maximum distance between them over all realisations. Within this framework, we can use different uncertainty models. We can also use different distance metrics, most commonly, the Hausdorff or the Fréchet distance (or their variants).

Some of these combinations have previously been studied [1, 2, 3]; all either have a relatively simple polynomial-time solution, or are NP-hard. We can ask the following questions.

1. For the directed Fréchet distance, the lower bound problem is in P, and the upper bound problem is NP-hard. For the directed Hausdorff distance, the situation is essentially reversed. Similar dichotomies exist for the weak (discrete) Fréchet distance, where different settings turn out to be NP-hard compared to the Fréchet distance. Is there something in common among the NP-hard variants? Why are the lower and upper bound problems reversed for NP-hardness?
2. Can we fill in the gaps by studying the remaining variations, both for the Hausdorff and the Fréchet distance, and conclude for each whether it is NP-hard or solvable in polynomial time?


References

- 1 Kevin Buchin, Chenglin Fan, Maarten Löffler, Aleksandr Popov, Benjamin Raichel, and Marcel Roeloffzen. Fréchet Distance for Uncertain Curves. *ACM Transactions on Algorithms*, 19(3), article 29, 2023. doi:10.1145/3597640.
- 2 Kevin Buchin, Maarten Löffler, Tim Ophelders, Aleksandr Popov, Jérôme Urhausen, and Kevin Verbeek. Computing the Fréchet Distance between Uncertain Curves in One Dimension. *Computational Geometry: Theory and Applications*, 109, article 101923, 2023. doi:10.1016/j.comgeo.2022.101923.
- 3 Christian Knauer, Maarten Löffler, Marc Scherfenberg, and Thomas Wolle. The Directed Hausdorff Distance between Imprecise Point Sets. *Theoretical Computer Science*, 412(32):4173–4186, 2011. doi:10.1016/j.tcs.2011.01.039.

Participants

- Peyman Afshani
Aarhus University, DK
- Hee-Kap Ahn
POSTECH – Pohang, KR
- Håvard Bakke Bjerkevik
TU Graz, AT
- Kevin Buchin
TU Dortmund, DE
- Maike Buchin
Ruhr-Universität Bochum, DE
- Siu-Wing Cheng
HKUST – Hong Kong, CN
- Guilherme D. da Fonseca
Aix-Marseille University, FR
- Vincent Despré
LORIA – Nancy, FR
- Ioannis Emiris
Athena Research Center –
Maroussi, GR
- Linda Kleist
TU Braunschweig, DE
- Fabian Klute
UPC Barcelona Tech, ES
- Maarten Löffler
Utrecht University, NL & Tulane
University – New Orleans, US
- Zuzana Masárová
IST Austria –
Klosterneuburg, AT
- Bojan Mohar
Simon Fraser University –
Burnaby, CA
- Alexander Munteanu
TU Dortmund, DE
- Martin Nöllenburg
TU Wien, AT
- Eunjin Oh
POSTECH – Pohang, KR
- Irene Parada
UPC Barcelona Tech, ES
- Salman Parsa
DePaul University – Chicago, US
- Zuzana Patáková
Charles University – Prague, CZ
- Jeff M. Phillips
University of Utah –
Salt Lake City, US
- Aleksandr Popov
TU Eindhoven, NL
- Ioannis Psarros
Athena Research Center –
Maroussi, GR
- Manfred Scheucher
TU Berlin, DE
- Raimund Seidel
Universität des Saarlandes –
Saarbrücken, DE
- Willem Sonke
TU Eindhoven, NL
- Subhash Suri
University of California –
Santa Barbara, US
- Monique Teillaud
INRIA Nancy Grand Est –
Villers-lès-Nancy, FR
- Konstantinos Tsakalidis
University of Liverpool, GB
- Torsten Ueckerdt
Karlsruher Institut für
Technologie, DE
- Ryuhei Uehara
JAIST – Nomi, Ishikawa, JP
- Birgit Vogtenhuber
TU Graz, AT
- Hubert Wagner
University of Florida –
Gainesville, US
- Emo Welzl
ETH Zürich, CH
- Carola Wenk
Tulane University –
New Orleans, US
- Sang Duk Yoon
Sungshin Women's University –
Seoul, KR
- Yelena Yuditsky
Université Libre de Bruxelles, BE



 **Remote Participants**

■ Anna Lubiw
University of Waterloo, CA

■ Ke Yi
HKUST – Hong Kong, CN

Novel Scenarios for the Wireless Internet of Things

Haitham Hassanieh^{*1}, Kyle Jamieson^{*2}, Luca Mottola^{*3}, Longfei Shangguan^{*4}, Xia Zhou^{*5}, and Marco Zimmerling^{*6}

- 1 EPFL – Lausanne, CH. haitham.alhassanieh@epfl.ch
- 2 Princeton University, US. kylej@princeton.edu
- 3 Polytechnic University of Milan, IT. luca.mottola@polimi.it
- 4 University of Pittsburgh, US. longfei@pitt.edu
- 5 Columbia University – New York, US. xia@cs.columbia.edu
- 6 TU Darmstadt, DE. marco.zimmerling@tu-darmstadt.de

Abstract

The Internet of Things (IoT) aims to network everything near and far in our ambient environment. Although the functional innovations for IoT are going full steam ahead, newly-emerging scenarios such as the Internet of Ocean and Implantable Things often come with limited power budgets, challenging deployment scenarios, and demanding computational resources, which fundamentally stress conventional IoT architecture, communications primitives, and sensing capabilities. The goal of this Dagstuhl Seminar was to bring together researchers from both academia and industry globally to *i*) review the capacity of existing IoT research from radical perspectives; *ii*) summarize fundamental challenges in modern IoT application scenarios that may then be investigated in joint research projects; and *iii*) discuss new types of hardware architecture, network stack, and communication primitives for these emerging IoT scenarios.

Seminar May 29 – June 1, 2023 – <https://www.dagstuhl.de/23222>

2012 ACM Subject Classification Hardware → Emerging technologies; Networks → Network architectures; Networks → Public Internet

Keywords and phrases Internet of Ocean Things, Internet of Medical Things, NextG Communication

Digital Object Identifier 10.4230/DagRep.13.5.182

1 Executive Summary

Longfei Shangguan

Xia Zhou

Marco Zimmerling

License © Creative Commons BY 4.0 International license
© Longfei Shangguan, Xia Zhou, and Marco Zimmerling

The past few years have witnessed progressive deployments of IoT devices in both personal and public space to facilitate smarter living. For instance, home gateways such as Amazon Echo can now pick up human speech, understand the semantics, and further loop in machines, appliances, and many others to react. Likewise, mobile IoT devices such as drones and robots are deployed for urban modeling, express delivery, and industrial inspection. Although the functional innovations builds upon existing IoT architecture are going full steam ahead, the newly emerging scenarios such as Internet of ocean and implantable things often come with limited power budget, form factor, and computation resources, which fundamentally challenge the conventional IoT architecture, communication primitives, and sensing capabilities.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Novel Scenarios for the Wireless Internet of Things, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 182–205

Editors: Haitham Hassanieh, Kyle Jamieson, Luca Mottola, Longfei Shangguan, Xia Zhou, and Marco Zimmerling



DAGSTUHL REPORTS

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

■ **Table 1** Agenda.

Day One		
09:00 — 09:15	Opening remarks	Longfei Shangguan, Xia Zhou, Marco Zimmerling
09:15 — 10:15	Keynote speech	Victor Bahl (Microsoft)
10:15 — 10:45	Break	
10:45 — 12:00	Speak-and-spark I	
12:00 — 13:30	Lunch	
13:30 — 15:00	Speak and spark II	
15:00 — 15:40	Coffee Break	
15:40 — 16:20	Invited Talk I (Prof. Olga Saukh)	Embedded Intelligence: The Way Forward
16:20 — 17:45	Speak and spark III	
Day Two		
09:00 — 09:45	Invited Talk II (Prof. Alberto Quattrini Li)	Current Snapshot and Opportunities on Underwater Sensing and Communication
09:45 — 10:30	Invited Talk III (Dr. Justin Chan)	Intelligent Mobile Systems for Equitable Healthcare
10:45 — 12:00	Breakout session I	
12:15 — 13:30	Lunch	
13:30 — 14:15	Invited Talk IV (Prof. Marco Zuniga)	Sunlight for Wireless Communication
14:15 — 15:00	Invited Talk V (Prof. Baccelli & Prof. Hahm)	IETF Protocols & Embedded Software for Low-Power IoT using RIOT
15:00 — 16:00	Coffee Break	
16:00 — 17:15	Breakout session II	
17:15 — 18:00	Sync on technical report writing	
Day Three		
09:00 — 10:00	Plenary discussion I: underwater & medical & airborne	
10:00 — 11:00	Break	
11:00 — 12:00	Plenary discussion II: energy-efficiency & AI & sensing and communication	
12:00 — 12:15	Closing remarks	Longfei Shangguan, Xia Zhou, Marco Zimmerling

The goal of this Dagstuhl Seminar was to bring together researchers from both academia and industry from around the world to *i)* review the existing IoT research directions from radical perspectives; *ii)* figure out fundamental challenges in modern IoT application scenarios that may then be investigated in joint research projects; and *iii)* discuss new types of hardware architecture, network stack, and communication primitives for these emerging IoT scenarios. An essential part of this seminar was to bridge the gap between the research momentum from academia and the practical needs from the industry and to actively engage a dialog between different communities. As IoT by nature is a multidisciplinary area involving innovations in hardware and architecture, algorithm and protocol, as well as applications, we sought researchers from different domains that are open to different research perspectives and welcome research of a different nature.

Seminar Schedule

The seminar commenced with a keynote speech by Dr. Victor Bahl, a Technical Fellow at Microsoft. This was followed by a pair of speak-and-spark sessions, during which each attendee delivered a brief presentation on their research, facilitating mutual familiarization of everyone's areas of interest. The inaugural day culminated with an invited talk given by Dr. Olga Saukh. Moving to the subsequent day, the program featured four invited talks along with two break-out sessions. During these break-out sessions, attendees were organized into smaller groups, each focusing on a specific IoT research theme. The final day was dedicated to wrapping up the event and engaging in discussions. Leaders of each break-out group presented summary reports of their respective discussions, effectively concluding the seminar. The detailed agenda is listed in Table 1.

Outcomes

The seminar centered around six comprehensive research topics within the realm of the Internet of Things (IoT). These topics encompassed a diverse spectrum, including underwater communication, implantable IoT devices, airborne communication systems, integrated com-



■ **Figure 1** Speak and spark session I.

■ **Figure 2** Speak and spark session II.

munication and sensing technologies, the intersection of artificial intelligence and IoT, and the optimization of energy efficiency in IoT applications. Through in-depth presentations and break-out discussions on these subjects, participants gained valuable insights into the cutting-edge advancements, challenges, and opportunities shaping the future landscape of IoT technology. The seminar fostered a collaborative environment where participants exchanged ideas and discussed potential research collaborations. Overall, the seminar was considered a success, building the next cornerstone for the wireless Internet of Things.

Furthermore, the extensive research discussions held during the break-out sessions have ignited a multitude of captivating concepts for the upcoming wave of IoT applications. This enthusiasm culminated in the creation of four technical reports, each focusing on distinct research realms:

- Energy Efficiency of IoT (§4.1)
- Integrated Communication and Sensing (§4.2)
- Advancements in Medical IoT (§4.3)
- Airborne Internet of Things (§4.4)
- Impact of AI on IoT (§4.5)

These reports delve into the challenges faced and chart potential trajectories for future progress in these specified areas.

2 Table of Contents

Executive Summary

Longfei Shangguan, Xia Zhou, and Marco Zimmerling 182

Overview of Talks

Resilient and Secure IoT

Arash Asadi 187

IETF Protocols & Embedded Software for Low-Power IoT using RIOT

Emmanuel Baccelli and Oliver Hahm 187

The Inevitable Unification of the Cloud and Telecommunications Infrastructures –
Science, Technologies, Strategy and Opportunities

Victor Bahl 188

Software-Defined Wireless Communication Systems

Bastian Bloessl 188

Intelligent Mobile Systems for Equitable Healthcare

Justin Chan 189

Low-Power Internet-of-Things on Earth and Space

Akshay Gadre 189

Energy Efficient Sensing for IoT

Junfeng Guan 190

UAV Systems and Networks

Karin Anna Hummel 190

Wireless Sensing and Interactions

Tianxing Li 191

Toward a Battery-less Internet of Things

Andrea Maioli 191

Space-IoT

RangaRao Venkatesha Prasad 192

Exploring the Aquatic World with an Internet of Underwater Robots and Sensors:
Current Snapshot and Opportunities on Underwater Sensing and Communication

Alberto Quattrini Li 192

Embedded Intelligence: A Way Forward

Olga Saukh 193

Low Latency Data Downlink for Low Earth Orbit Satellites

Deepak Vasisht 193

Towards Ubiquitous Mobile Connectivity

Chenren Xu 193

The Small Data Problem in Understanding Older Adult Mobility

Rong Zheng 194

Sunlight for Wireless Communication

Marco Antonio Zúñiga Zamalloa 194

Breakout Session Report

Energy Efficiency of IoT <i>Tianxing Li and Andrea Maioli</i>	195
Integrated Communication and Sensing <i>Bastian Bloessl and RangaRao Venkatesha Prasad</i>	197
The Future of Medical IoT Research <i>Justin Chan and Rong Zheng</i>	199
Airborne Internet-of-Things <i>Akshay Gadre and Deepak Vasisht</i>	201
Impact of AI on IoT <i>Olga Saukh and Marco Antonio Zúñiga Zamalloa</i>	202
Participants	205

3 Overview of Talks

3.1 Resilient and Secure IoT

Arash Asadi (TU Darmstadt, DE)

License © Creative Commons BY 4.0 International license
© Arash Asadi

Joint work of Arash Asadi, Andrea Ortiz

Main reference Amir Ashtari Gargari, Andrea Ortiz, Matteo Pagin, Anja Klein, Matthias Hollick, Michele Zorzi, Arash Asadi: “Safehaul: Risk-Averse Learning for Reliable mmWave Self-Backhauling in 6G Networks”, CoRR, Vol. abs/2301.03201, 2023.

URL <https://doi.org/10.48550/arXiv.2301.03201>

In this talk, I will delve into the endeavors of our research team aimed at enhancing mmWave self-backhauling. Our focus has been on addressing challenges related to beamforming and route selection within self-backhauled networks. We’ve explored solutions using reinforcement learning as well as traditional optimization methods. Additionally, I will elaborate on our strategies for countering adversarial WiFi sensing and mitigating user tracking through the application of radio fingerprinting techniques. Furthermore, I will provide a concise overview of our advancements in Reflective Intelligent Surfaces (RISs) development, encompassing both sub-6GHz implementations utilizing RF-switches and 60GHz designs employing Liquid Crystals.

3.2 IETF Protocols & Embedded Software for Low-Power IoT using RIOT


Emmanuel Baccelli (FU Berlin, DE) and Oliver Hahm (Frankfurt University of Applied Sciences, DE)

License © Creative Commons BY 4.0 International license
© Emmanuel Baccelli and Oliver Hahm

The Internet of Things (IoT) is a term generally used for a (too) wide variety of hardware, platforms and protocols. In this talk we focus less on cloud/edge IoT aspects and hardware based on CPU/GPU, and more on things operated “beyond” edge computing: we focus on ultra-low power hardware, software and protocols running on microcontrollers. This is a category of devices the IoT relies heavily upon. To be more concrete, RFC7228 distinguishes several classes of such devices, for example Class 1 gathers devices providing memory budgets of 10 kB RAM and 100 kB flash. With such a target in mind, we quickly overview the key aspects of relevant network protocols (distributed algorithms, standard specifications, interoperable implementations) and we flip through the list of relevant working groups that are currently active at the Internet Engineering Task Force (IETF/IRTF) standardizing IP protocol adaptations for ultra-low power IoT. In parallel, we overview the embedded software platform developments that take place, in order to provide open source interoperable implementations of these open standards. We overview developments based on RIOT, a small-footprint operating system that runs on a wide variety of microcontroller boards and aggregates various libraries and network stacks.

3.3 The Inevitable Unification of the Cloud and Telecommunications Infrastructures – Science, Technologies, Strategy and Opportunities

Victor Bahl (Microsoft Corporation – Redmond, US)

License  Creative Commons BY 4.0 International license
© Victor Bahl

We are at the beginning of an unprecedented opportunity for information technology startups and the established cloud industry to become a part of the next generation telecommunications infrastructure. Together, we can radically change it through softwarization, AI and edge computing. I will describe the scientific advances and business needs for bringing us to where we are today and then cast an eye to the future in sharing with the audience a vision for where things are going with telecommunications, including key enablers and potential surprises on the horizon. This will set the context for describing the opportunity ahead for the engineering and research community in the next several years, and beyond, as we stay at the forefront of the modernization that will enable ubiquitous computing via telecommunication networks propelled by innovations in AI and edge computing. I will next move into near-term strategy with Microsoft standing up of a new business division called Azure for Operators (AFO). I will describe the vision that led to the creation of AFO, its mission, and products, and the significant technical and scientific challenges, which we are pursuing. When overcome they will lead to the inevitable convergence of two massive industries that will open up a new decade of opportunities for universities, research institutes, established companies, and startups.

3.4 Software-Defined Wireless Communication Systems

Bastian Bloessl (TU Darmstadt, DE)

License  Creative Commons BY 4.0 International license
© Bastian Bloessl

Future wireless communication systems will be increasingly software-defined, with AI-native components to optimize performance. With current Software Defined Radios, there are powerful general-purpose hardware platforms that could drive such systems. What we are missing are software frameworks that can make efficient use of the heterogeneous compute resources available on those platforms. In this talk, I will briefly discuss my research effort on filling this gap by designing and implementing novel real-time signal processing systems that provide full control of the data flow through custom schedulers and native support for hardware accelerators like FPGAs and GPUs. This can serve as the enabler for a new generation of wireless networks that use AI not just for multi-objective optimizations but as the base for a new system design that goes beyond traditional protocol structures and complex error handling to create more robust and more resilient systems.

3.5 Intelligent Mobile Systems for Equitable Healthcare

Justin Chan (University of Washington – Seattle, US)

License  Creative Commons BY 4.0 International license
© Justin Chan

Access to even basic medical resources is greatly influenced by factors like an individual's birth country and zip code. In this talk, I will present my work on designing intelligent mobile systems for equitable healthcare. I will showcase three systems that are not only interesting from a computational standpoint but are also having real-world medical impact. The first system can detect ear infections using only a smartphone and a paper cone. The second system enables low-cost newborn hearing screening using inexpensive earphones. Lastly, I will present an ambient sensing system that employs smart devices to detect emergent and life-threatening medical events such as cardiac arrest. Through these examples, I will demonstrate how new computational and sensing techniques that generalize across hardware and work in real-world environments can help to address pressing societal problems.

3.6 Low-Power Internet-of-Things on Earth and Space

Akshay Gadre (University of Washington – Seattle, US)

License  Creative Commons BY 4.0 International license
© Akshay Gadre

As the number of connected devices increases, the stress on the wireless bandwidth available keeps increasing. Researchers across the area of wireless systems are building novel communication and sensing systems to overcome these limitations by addressing deployment domain-specific problems such as, Space-Based Internet, Low-Power Low-Bandwidth IoT Sensors for Agriculture, and Low-cost Sensing Solutions for the supply chain. Yet, many of these research prototypes are considered as research artifacts which provide basic connectivity and sensing, yet lacking much of the improved utility (as typically mentioned in many research contributions).

This talk will focus on the important wireless problems at the frontier of these domains to improve end-user acceptance and utilization of these impressive technologies. This talk will present several important questions that a typical user faces today and how academics and industry researchers can collaborate on these problems to alleviate these issues. Finally, this talk will provide the base for interesting discussion throughout the day of meetings with researchers on brainstorming these ideas and building collaborations.

3.7 Energy Efficient Sensing for IoT

Junfeng Guan (EPFL Lausanne, CH)

License © Creative Commons BY 4.0 International license
© Junfeng Guan

Main reference Junfeng Guan, Jitian Zhang, Ruochen Lu, Hyungjoo Seo, Jin Zhou, Songbin Gong, Haitham Hassanieh: “Efficient Wideband Spectrum Sensing Using MEMS Acoustic Resonators”, in Proc. of the 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), pp. 809–825, USENIX Association, 2021.

URL <https://www.usenix.org/conference/nsdi21/presentation/guan>

I will explore the integration of Micro-Electro-Mechanical-Systems (MEMS) filters with innovative sparse recovery algorithms, enabling energy-efficient IoT devices to effectively capture wideband spectra. I will showcase the practical implementation of this approach in two distinct applications. Firstly, I will introduce an energy-conscious spectrum sensing approach that empowers low-power IoT devices to detect wideband channel occupancies. This capability enables these devices to opportunistically utilize unoccupied channels, facilitating dynamic spectrum sharing. Secondly, I will unveil a precise self-localization technique for IoT devices. This method harnesses existing ambient 5G communication signals without necessitating any alterations to the 5G infrastructure. This innovation offers accurate self-localization capabilities without requiring modifications to the underlying infrastructure.

3.8 UAV Systems and Networks

Karin Anna Hummel (JKU Linz, AT)

License © Creative Commons BY 4.0 International license
© Karin Anna Hummel

We aim to develop efficient mobile networked systems that include machines, vehicles, and humans. Hereby we propose novel architectures of such networked systems, analyze the performance, and experimentally evaluate the overall capabilities of the wireless network and often context-sensitive algorithms for the wireless link or overall networked system, including predictive approaches. Two recent application scenarios for which we set-up small wireless networked testbeds to work on novel solutions target systems for human-drone teams in indoor scenarios, where the drone’s camera is the primary sensor, used to navigate and interact with the human and the environment. We find that a distributed control architecture leveraging direct Wi-Fi links is feasible and beneficial for such systems having an AI or machine learning component, yet common Wi-Fi cannot provide time guarantees which would be necessary for many real-time applications. However, this architecture enables the successful deployment of sophisticated applications such as drone learning from human emotions, which is one of our current research projects. In this talk, I will briefly introduce some of our current research results on these research threads.

3.9 Wireless Sensing and Interactions

Tianxing Li (Michigan State University – East Lansing, US)

License © Creative Commons BY 4.0 International license
© Tianxing Li

In this talk, I will present three distinct advancements in the realm of sensing and interaction technologies. Firstly, I will introduce an innovative approach to inaudible attacks on smart earbuds, overcoming limitations of existing methods by leveraging both direct and reflective paths for attacking signals, thereby enhancing signal-to-noise ratios. Secondly, I will present a 3D Hand Posture Reconstruction system utilizing 2D Rolling Fingertips, which capitalizes on active optical labeling and exploits rolling shutter effects in smartphone cameras to achieve improved 3D tracking for complex scenarios like underwater environments and specialized applications like virtual writing. Lastly, I will present a low-power system for precise face touch detection, integrating wrist inertial and novel finger vibration sensors with a cascading classification model for efficient gesture filtering. This system achieves a high F-1 score of 93.5% for detecting face touch events, while maintaining minimal power consumption, making it suitable for prolonged usage with battery-powered devices. Each advancement is supported by practical prototypes and evaluations, showcasing their potential across a range of real-world scenarios.

3.10 Toward a Battery-less Internet of Things

Andrea Maioli (Polytechnic University of Milan, IT)

License © Creative Commons BY 4.0 International license
© Andrea Maioli

Joint work of Andrea Maioli, Luca Mottola, Mikhail Afanasov, Naveed Bhatti, Dennis Campagna, Giacomo Caslini, Fabio Massimo Centonze, Koustabh Dolui, Erica Barone, Muhammad Hamad Alizai, Junaid Haroon Siddiqui

Main reference Andrea Maioli, Luca Mottola: “ALFRED: Virtual Memory for Intermittent Computing”, in Proc. of the SenSys '21: The 19th ACM Conference on Embedded Networked Sensor Systems, Coimbra, Portugal, November 15 – 17, 2021, pp. 261–273, ACM, 2021.

URL <https://doi.org/10.1145/3485730.3485949>

Battery-less devices revolutionize conventional battery-dependent setups by integrating ambient energy harvesting technologies, thereby enabling a sustainable Internet of Things (IoT) paradigm with minimal maintenance requirements. However, the intermittent nature of ambient energy sources presents challenges. These devices frequently encounter energy shortages, causing intermittent computation where active phases alternate with inactive ones due to waiting for adequate energy. Given their limited energy reservoir, optimizing battery-less device operation is crucial to maximize computational output from ambient energy. This presentation introduces our research effort addressing these concerns. We begin with the introduction of ALFRED, a virtual memory abstraction and compilation pipeline tailored for mixed-volatile platforms. ALFRED autonomously identifies optimal program state mappings between volatile and non-volatile memory, enhancing efficiency. Our focus also extends to ensuring consistent efficiency in battery-less devices. We’ve developed a system design to efficiently regulate supply voltage and clock frequency, even in hardware-limited devices lacking dynamic voltage and frequency scaling support. Implementing two hardware/software co-designs capturing these aspects, our approaches reduce battery-less device energy consumption by up to 170% and significantly decrease workload completion time, enhancing their overall performance. Through these advancements, we contribute to the realization of robust and optimized IoT systems powered by ambient energy sources.

3.11 Space-IoT

RangaRao Venkatesha Prasad (TU Delft, NL)


License  Creative Commons BY 4.0 International license
© RangaRao Venkatesha Prasad

Joint work of RangaRao Venkatesha Prasad, Sujay Narayana

The Internet of Things (IoT) is making a great impact on our lives and changing the way we interact with others, the environment, and even machines. Now there are two considerations for IoT space. (a) The IoT devices deployed on the ground need a reliable connectivity backbone. (b) Space technology is seeing unprecedented growth since space subsystems are becoming smarter, more reliable wireless links, and further miniaturized – all these without the need for radiation-hardened components in Low Earth Orbit (LEO). The above two aspects lead us to a new domain of IoT called the Space Internet of Things (Space-IoT). In recent times, there has been a huge interest and investment in Space-IoT. Space can be a suitable platform to solve many of the current and future problems in IoT, and the possible solutions are yet to be explored in depth in the space environment. In this talk, I will discuss my group's research effort in this research domain.

3.12 Exploring the Aquatic World with an Internet of Underwater Robots and Sensors: Current Snapshot and Opportunities on Underwater Sensing and Communication

Alberto Quattrini Li (Dartmouth College – Hanover, US)

License  Creative Commons BY 4.0 International license
© Alberto Quattrini Li

The aquatic world, with more than 70% of the Earth covered by water, plays a critical role in our society and the economy, as recognized by many organizations including the United Nations and the World Wildlife Fund, which estimated the economy value related to the aquatic world – also called Blue Economy – to be at least US\$24 trillion. Yet, currently, more than 80% of the ocean is unmapped. This talk delves into the challenges and opportunities of deploying an Internet of Underwater Robots and Sensors (or more generally, Things) to explore the underwater world and support the Blue Economy, particularly focusing on underwater sensing and communication. Supported by lessons learned from actual field experiments, I will present first the general requirements for such underwater robots and sensors to operate in such environments; second, the technologies used in the current deployments, highlighting gaps and open problems; and finally, current research that shows research opportunities in the space of Ocean Internet of Things to achieve the long-term goal of ocean IoT, to support large scale aquatic applications, such as environmental monitoring or archaeological exploration.

3.13 Embedded Intelligence: A Way Forward

Olga Saukh (TU Graz, AT)

License © Creative Commons BY 4.0 International license
© Olga Saukh

AI is utilized in various IoT applications, ranging from environmental monitoring to home automation and production processes. However, the most concerning applications of AI are in safety-critical systems, such as transportation, medicine, and control, because incorrect use of AI can have devastating consequences. A significant amount of current research effort is focused on increasing the robustness of deep models or implementing safe and low-resource post-deployment domain adaptation techniques to manage domain shifts. In this talk, I will provide a concise overview of the state-of-the-art in the field and present the latest research results from my team. We discover essential properties of the loss landscape geometry and leverage these to improve AI safety across a broad range of applications, particularly in scenarios where resource constraints are at stake.

3.14 Low Latency Data Downlink for Low Earth Orbit Satellites

Deepak Vasisht (University of Illinois – Urbana-Champaign, US)

License © Creative Commons BY 4.0 International license
© Deepak Vasisht

Joint work of Deepak Vasisht, Jayanth Shenoy, Ranveer Chandra

Main reference Deepak Vasisht, Jayanth Shenoy, Ranveer Chandra: “L2D2: low latency distributed downlink for LEO satellites”, in Proc. of the ACM SIGCOMM 2021 Conference, Virtual Event, USA, August 23-27, 2021, pp. 151–164, ACM, 2021.

URL <https://doi.org/10.1145/3452296.3472932>

Large constellations of Low Earth Orbit satellites promise to provide near real-time high-resolution Earth imagery. Yet, getting this large amount of data back to Earth is challenging because of their low orbits and fast motion through space. Centralized architectures with few multi-million dollar ground stations incur large hour- level data download latency and are hard to scale. We propose a geographically distributed ground station design that uses low-cost commodity hardware to offer low latency robust downlink. We also discuss intersections of compute and networks in this emerging context. We believe our work engineers a paradigm shift similar to the one from super-computers to data centers, where software architectures can extract high performance from commodity hardware.

3.15 Towards Ubiquitous Mobile Connectivity

Chenren Xu (Peking University, CN)

License © Creative Commons BY 4.0 International license
© Chenren Xu

Mobility is the essential norm in human society. As today’s wireless and mobile networking technology already brings 99% usable connectivity between our personal devices and (edge) cloud services, the remaining fragmented 1% connectivity scenarios, including but not limited to extremely lower power, high mobility, and massive access, which still face domain-specific networking challenges. More importantly, these minority cases are likely to turn over into

the majority tomorrow as our global society is consciously evolving in the direction of energy and production efficiency improvement. This talk will introduce our recent efforts towards the vision of “Ubiquitous Mobile Connectivity”. Specifically, we will present the design, implementation and deployment experience of our mobile RFID, VILD and multipath networking system for improving scalability, availability and reliability in logistics, Vehicular-to-X and high-speed railway networks.

3.16 The Small Data Problem in Understanding Older Adult Mobility

Rong Zheng (McMaster University – Hamilton, CA)

License © Creative Commons BY 4.0 International license

© Rong Zheng

Joint work of Rong Zheng, Yujiao Hao, Boyu Wang

The proportion of older adults (aged 65 and older) worldwide has been increasing steadily over the past 40 years. Among older adults, mobility is a crucial indicator of functional status, and a predictor of quality of life and longevity; hence, it is often called the sixth vital sign. Mobility encompasses not only the physical activities of older adults, and the performance of specific maneuvers such as sit-to-stand, walking or climbing stairs, but also participation in society (e.g., the ability to drive, accessibility to public transportation). Recent advancements in mobile technologies, artificial intelligence, embedded and sensing devices create exciting opportunities to address mobility challenges faced by the aging population. However, the shortage of quality, labeled, free-living data from target populations remains to be a main barrier in developing and applying effective solutions for continuous monitoring, analysis and assessment.

My research work aims to mitigate the data scarcity problem in characterizing old adult activities and answer “what”, “when” and “how well” questions regarding their functional mobility. We take a multi-pronged approach. Directly addressing data availability, we developed ChromoSim, a cross-mobility IMU data synthesizer, CROMOSim, that can transform abundant data from vision, motion capture systems to IMU data. At the modelling level, we devised efficient invariant feature learning framework that can easily adapt to novel subjects or devices and can handle noisy labeled data collected in the wild.

3.17 Sunlight for Wireless Communication

Marco Antonio Zúñiga Zamalloa (TU Delft, NL)

License © Creative Commons BY 4.0 International license

© Marco Antonio Zúñiga Zamalloa

Joint work of Seyed Keyarash Ghiasi, Marco Antonio Zúñiga Zamalloa, Koen Langendoen

Main reference Seyed Keyarash Ghiasi, Marco Antonio Zúñiga Zamalloa, Koen Langendoen: “A principled design for passive light communication”, in Proc. of the ACM MobiCom ’21: The 27th Annual International Conference on Mobile Computing and Networking, New Orleans, Louisiana, USA, October 25-29, 2021, pp. 121–133, ACM, 2021.

URL <https://doi.org/10.1145/3447993.3448629>

We, humans, already use 50% more energy moving information than moving airplanes around the world. Communication is central to our societies but it is taking a toll on the earth. We want to use a free, abundant, and natural resource for wireless communication: sunlight. Similar to the way you can use a mirror to communicate by reflecting light, our aim is to



■ **Figure 3** Plenary report and discussion I.

■ **Figure 4** Plenary report and discussion II.

exploit optical devices that can change their reflective and transmissive properties to send information, but without you noticing any flicker. In this manner, objects will be able to talk to each other using daylight, an eco-friendly solution. In this talk, I will present some of our recent work in this area and its potential applications.

4 Breakout Session Report

4.1 Energy Efficiency of IoT

Tianxing Li (Michigan State University – East Lansing, US), Andrea Maioli (Politecnico di Milano – Milano, IT)

License © Creative Commons BY 4.0 International license
© Tianxing Li and Andrea Maioli

4.1.1 Relevance of energy-efficiency improvements

Better energy efficiency leads to performance improvement. However, such performance improvement may be irrelevant, depending on the application scenario and the device's lifetime. For example, in agriculture applications where deployments last for an entire season, a couple of days of improvement in the battery life is not significant enough. Conversely, in other application scenarios, such as battery-less devices deployments that use harvested energy, a slight improvement (i.e., 10%) in energy efficiency may lead to a lower number of energy failures or to energy/power neutrality (i.e., the system consumes the same amount of harvested energy, enabling perpetual and unattended operations). Therefore, the relevance of performance increase due to higher energy efficiency is application-specific.

4.1.2 Energy efficiency improvements target

IoT devices usually sense, compute, interact with the environment, and communicate sampled data. These four types of actions have different energy consumptions, which change depending on the IoT device and the application scenario. Unthinkingly improving energy efficiency may not result in a performance increase. For example, improving the energy efficiency of computation in underwater devices and drones does not lead to a significant performance increase, as their energy consumption is primarily due to movements. Similarly, devices that periodically sense the environment and enter deep sleep after each measurement consume the most energy while inactive. In such a scenario, improving the energy efficiency of sensing, processing, or communication leads to a lower performance improvement than improving deep-sleep energy efficiency. For these reasons, researchers should focus their resources on improving the energy efficiency of the most energy-hungry components of their applications.

4.1.3 Design space exploration

There are multiple platforms, techniques, and components to design IoT devices. These elements affect devices' energy efficiency. However, as we argued before, the most efficient components/techniques (i.e., custom designing an SoC) does not necessarily translate to the most efficient choice due to higher costs, higher design time, or irrelevant performance improvements. Therefore, when designing IoT devices, we must consider every trade-off that comes with using specific hardware and software components and select the one that suits the application requirements.

4.1.4 Energy harvesting

Ambient energy harvesting unlocks new scenarios where devices use the energy available in the environment to operate and recharge their batteries, extending battery life and reducing maintenance efforts. Energy harvesting can be used to improve devices' energy efficiency. During the seminar, we identified two complementary options. First, devices can harvest energy produced during normal operations. For example, we need to sense the airflow to stabilize drone's movement. In this case, we can harvest the energy from the movement of the drones to build a battery-less airflow sensing system. Second, devices can correlate sensing operations with energy harvesting sources. For example, we can harvest solar energy from ambient light intensity sensing to enable a battery-less gesture recognition system. In general, improving energy harvesting techniques was considered the "to-go" option during the seminar, with a particular interest in recovering the energy wasted due to normal operations (e.g., harvesting wind energy from drone movements).

4.1.5 Dynamic/runtime parameter adjustments

The increase in design complexity and the possibility of energy harvesting unlocks new optimization techniques. Devices should adapt their duty cycles and their application workload depending on available energy and the power source they are currently using (e.g., ambient energy or battery). Moreover, applications should dynamically adapt their parameters to always operate in the most efficient setting. Existing systems vary device operating settings, such as voltage and frequency. However, devices can also adapt other parameters, such as their sampling rate, the computation resolution, and the number of algorithm inputs (e.g., machine learning models). This requires designing new techniques to identify the parameters to tune and new energy-aware hardware/software designs that provide the required capabilities, such as power source recognition and real-time device energy consumption measurement.

4.1.6 Edge/local computation

The advances in AI applications increase the computational complexity of the processing executed on IoT devices, with a consequent increase in energy consumption and resource utilization. Local device computation may no longer be beneficial or possible due to tight energy budgets or insufficient resources (e.g., ML models may require more memory than the one available). Therefore, energy consumption due to data communication and sensing may no longer dominate. Data offloading may become an option on highly resource-constrained IoT devices, despite communication costs that usually represent the highest energy consumption of applications. Therefore, system designers should investigate the trade-offs between local and edge computation. In both these cases, data compression techniques may improve

devices' energy consumption, as they can decrease the size of communication data or the size of ML models, potentially leading to lower energy consumption. However, the challenge is whether we can design a general framework to optimize the compression parameters on the fly so that the system can adapt to complex and dynamic scenarios.

4.2 Integrated Communication and Sensing

Bastian Bloessl (TU Darmstadt, DE), RangaRao Venkatesha Prasad (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Bastian Bloessl and RangaRao Venkatesha Prasad

4.2.1 The Role of Interference Management

Today's wireless technologies like LoRa and NB-IoT scale surprisingly well with the number of nodes. One of the reasons for this is the RF environment, which is in most cases relatively static, given the fact that gateways are deployed at exposed locations, making it less likely that significant proportions of the Fresnel zone are frequently obstructed. This raises the question, whether interference management is still of prime interest. We believe that the topic should not drift out of focus for two reasons: (1) Even though LoRa and NB-IoT are popular choices today, it is not clear what we will use in the future. The design space for wireless communication technologies is large. With better energy harvesting and more power-efficient nodes, it seems likely that we will see technologies with more capable nodes that can cater to higher traffic demands. In these cases, interference management will play a more important role again. (2) Future IoT networks will become three dimensional, integrating airborne gateways mounted on UAVs, balloons, or satellites. These exposed gateways will cover larger geographic areas and, therefore, a much higher number of nodes that compete for spectrum.

4.2.2 The Role of Intelligent Surfaces

Reconfigurable Intelligent Surfaces (RISs) are a technology that could help to shape the RF environment to increase coverage and manage interference, as they allow steering reflections on a surface to guide signal propagation. Apart from diversity gains from reflected signals, this is particularly interesting for NLoS scenarios, where RISs can increase coverage by steering reflected beams to a target. While this technology promises great advantages, it is, at the moment, mainly discussed for mmWave technologies and indoor scenarios, where the expected benefits are particularly large. Whether the adoption in IoT deployments is technologically feasible and economically attractive has yet to be seen. Both the lower frequency bands and the different deployment scenarios might be problematic: It is unclear how well RISs can be adapted for sub-GHz bands and whether these surfaces are still cost effective, especially considering the extent of typical IoT networks. Another angle towards RISs is their potential use to enhance the privacy of users. Controlling signal propagation, they cannot only influence where signals go but also where they do not go. A signal that an eavesdropper never receives or overhears presents certainly the best protection to the user's privacy. Thinking further into the future, it would be interesting to extend the idea of RISs from reflection to penetration, working on materials that can change their attenuation coefficient. This could either be used to increase indoor reception or shield RF inside a room or building to protect from eavesdroppers.

4.2.3 (Joint) Communication and Sensing

The possibilities and accuracy of RF sensing is impressive, especially considering that most works exploit signals from ubiquitous technologies like WLAN that are not specifically designed for sensing. The fact that it is possible to see through walls, detect people, emotions, and even vital signs shows the potential. Future technologies will likely be designed with sensing in mind and waveforms will be adapted accordingly. One of the main questions that arises in this context is whether there should be a waveform for joint communication and sensing or if there are dedicated waveforms for communication and sensing. The former will likely be a compromise between the two applications, while the latter provides more flexibility to support sporadic sensing or increasing the accuracy by dedicating more spectrum to sensing. Besides technologies with native support for sensing, the most disruptive change in the field results from advancements in machine learning algorithms. Yet, we often reuse existing models for RF sensing. Prime examples are image networks that are used to process RF spectrograms. We believe that the next step for sensing is the development of foundational models, dedicated to the application. A network that processes IQ samples, i.e., the time domain signal, does not require calculating the FFT and includes phase information. Another question is how we train and feed these networks. Today's sensors are not designed to provide the input for ML algorithms. An audio device might, for example, do preprocessing with a psychoacoustic model of humans, filtering out information that might be relevant for ML. Or videos and pictures will be preprocessed to make them look nicer. This goes as far as replacing objects with reference pictures that were available during training (e.g. Samsung smartphones replace low-quality photos of the moon with high-resolution pictures). Future sensors might, therefore, have an option to output raw data or do preprocessing specifically for ML algorithms.

4.2.4 Privacy Implications and Ethics

Today, inventing new technologies comes with more responsibilities. Given the potential scale of IoT applications, privacy should not be an afterthought. Instead, we have to consider from the start what would happen, if the technology is adopted by thousands, millions, or billions. A recent negative example are Apple AirTags and similar trackers from other vendors. The potential for misusing the technology was obvious, yet, it was released on a global scale. Just now, Apple and Google work together on a standard to make misuse like stalking harder. Also more and more funding agencies expect considerations with the impact on privacy, society, or the environment. As researchers, we have a responsibility to educate the general public about the privacy implications of technologies, help them to make informed decisions, and ideally provide tools to protect their personal information. Considering the increasing complexity and ubiquity of IoT applications, the next big breakthrough could be a standard or technology that enables users to stay in control of their data.

4.3 The Future of Medical IoT Research

Justin Chan (University of Washington – Seattle, US), Rong Zheng (McMaster University – Hamilton, CA)

License © Creative Commons BY 4.0 International license
© Justin Chan and Rong Zheng

4.3.1 Challenges and recommendations

Community datasets. There is a need for more high quality biomedical datasets like PhysioNet and CheXNet to enable the community to benchmark technical progress in a similar fashion to the ImageNet Large Scale Visual Recognition Challenge.

Below we outline several challenges for high-quality dataset collection and curation:

Dataset diversity. To ensure that an IoMT system can scale, it is essential that any datasets on which it relies on is diverse across at least three different dimensions: a) patient demographics – These populations can involve demographics related to age, race and socioeconomic income. The demographics in the dataset should be representative of the patient population in which the system would eventually be deployed in. b) deployment environments – The data should be collected from a mixture of both controlled and uncontrolled environments in the wild. Controlled environments can include labs and clinics, while uncontrolled environments can include home environments. c) device heterogeneity – The data should be collected across different hardware models (e.g. different models of smartphones) and a clear description of how the data was collected should be described.

Dataset imbalance. Often in biomedical datasets, it is significantly easier to obtain data from a control population than from populations with a particular disease. To overcome issues of dataset imbalance, we believe there are opportunities to develop generative AI systems or physical simulations to create synthetic datasets that complement data collected from the real world.

Data normalization and standardized study protocols. IoMT devices are diverse in their form factors and physical properties. Placements and means of attachment also affect the characteristics of data obtained. For datasets to be exchangeable, ideally, standardized study protocols should be followed and some form of normalization need to be done. Minimally, the measurement procedure and detailed device specification should be provided along with the collected data.

Data science. Below we describe several areas that should be considered when analyzing biomedical datasets:

Clinical implications of system errors. Different medical applications can have different evaluation metrics. Even for false positive and false negative rates, what is considered acceptable is heavily dependent on the medical context. Analysis on biomedical datasets should describe the clinical implications of the metrics adopted, as well as how the error rates compare to standards set by regulatory and standards such as the FDA, ISO, or the performance obtained by similar medical technologies.

Subgroup analysis. Analysis of biomedical data should provide subgroup analysis to evaluate system performance across different dimensions including patient demographics, deployment environments, and system configurations. This will provide a more detailed picture of how the system performs instead of relying solely on aggregate performance numbers.

Explainable models. The development of explainable biomedical models for IoMT can help increase confidence in the system's decisions and pave the way to adoption. Specifically, explainable models can help to verify that the system is not making a decision based on artifacts in the data and is instead relying on clinically sound features.

Clinical studies

Establishing partnerships. Clinical studies often require access to clinical sites and collaborations with interested clinicians, which may not be easily accessible for everyone. We believe that there are opportunities beyond traditional healthcare sites where subjects can be recruited: 1) Partnering with public health authorities can be another path to collecting data for IoMT systems 2) Collaborations with private wearable or medical devices companies can enable collection of large-scale datasets 3) NGOs and global health organizations with existing clinical connections can enable clinical studies particularly in low and middle income countries 4) Crowdsourcing platforms can be a way to efficiently collect data from a diversity of environments and smart devices.

4.3.2 Opportunities

Personalized medicine. IoMT systems can enable longitudinal monitoring of an individual's health conditions, and adapt its sensing or detection capabilities to the individual through online or lifelong machine learning with private data. For example, there are opportunities for smart speakers and voice assistants to track health biomarkers to predict COPD relapses, dementia and Alzheimers.

Tracking of rare medical events. IoMT systems are able to capture significantly more data outside clinical settings especially in the home and can potentially track difficult to predict events like cardiac arrests and seizures. This can enable pre-screening and early diagnosis of medical disorders that would otherwise be challenging to track.

Remote diagnostics for telemedicine. With the increased adoption of telemedicine, there is an opportunity to create IoMT systems that can be performed remotely in conjunction with a physician or health care professional. These systems can involve repurposing the sensors on existing smart devices, or low-cost attachments that can be distributed at scale.

Ambient sensing systems. Ambient sensing systems can enable passive collection of medical data in domestic and public settings. Opportunities are abundant in the development of a) novel sensing methods leveraging chemical, acoustic, radio, or optical modalities in existing infrastructure and b) new privacy-preserving systems can enable large-scale epidemiological systems for tracking disease outbreaks.

Smart materials for wearable biofluid sensors. There are lots of opportunities to bring chemical testing out of the lab and create wearable and flexible chemical sensors that can continuously monitor biofluids like sweat, saliva, and tears to track vital signs and diseases.

Closed-loop systems. Beyond sensing, IoMT can be combined with actuators for medical intervention. Examples are, wearable sensors with auto-injectors that can detect and reverse events like opioid overdoses or anaphylaxis caused by allergic reactions; and prosthetics controlled by electromyography. The rapid progress in brain-computer interface (BCI) technology recently has opened up new possibilities in augmented or virtual reality applications with human-in-loop.

4.4 Airborne Internet-of-Things

Akshay Gadre (University of Washington – Seattle, US), Deepak Vasisht (University of Illinois – Urbana-Champaign, US)

License © Creative Commons BY 4.0 International license

© Akshay Gadre and Deepak Vasisht

Joint work of all members of the Airborne Internet-of-Things session

Internet-of-things is also unlocking the potential to augment devices not typically deployed statically as a sensor but instead a mobile device that can move around and sense different environments using the same sensor. One such direction of research exploration is understanding and augmenting the opportunity of airborne internet-of-things where devices with batteries/solar panels spend several seconds (insects), minutes (UAVs/drones/blimps), hours (airplanes) and even days (satellites) away from power.

There are tremendous opportunities that can be enabled by effectively capitalizing on the ability of these sensors to move around and provide connectivity. For instance, satellites and drones can act as data grabbers from terrestrial sensors with poor connectivity. Satellites and aircrafts can provide ocean coverage for various capabilities (localization, tracking, SOS, communication) that we traditionally take advantage of in well-connected environments. There has also been demonstration of using satellites for cellular backhails for tracking global shipments by Telefonica. Further, there are new opportunities in exploring new capabilities such as edge computing and mesh networking in space. On the UAV side, enabling better autonomy for coordination of drone swarms for various sensing, transportation and entertainment applications remain important. Finally, planes can act as an intermediate zone of innovation between the drones which are battery constrained and the satellites which are communication constrained to assist the other two in enabling the same applications. It is also critical to not just build these services but also specializing them for commercial and environmental applications in real-world industries, such as mining, oil and gas, and agriculture. There are also upcoming deployments that cannot be classified into the above three categories but can enable new applications such as Insect IoT and Balloons IoT.

These opportunities can only be enabled by solving real world system challenges across the cyber-physical system stack from embedded and communication systems to better software applications that maximize the utilization of the limited resources at the sensors. One such challenge for satellites that is gaining increased exposure is the problem of reliability provided by satellite constellation-based internet. Indeed, a recent story of deployment in Ukraine highlighted this issue at a recent tutorial at MobiCom. Further, coordination of satellite communication directly to the sensors on the ground remains an open problem. Another important problem for satellite deployment is sustainability of deployment and debris reduction of existing satellites. In fact, recent efforts have taken place to use the 4th stage rockets lost in the orbit for satellite applications. Further, it was also described that most LEO satellite deployments deplane in orbit and get melted in the atmosphere. However, these are preliminary solutions and a more robust solution for sustainability at scale needs to be developed. Another important problem is standardization of communication protocols across the planet like cellular protocols where the constellations can effectively compete for the users without unnecessary augmenting overlapping infrastructure. On the UAV front, there are several traditional challenges of battery powered devices taken to the extreme – power, computation, storage, mobility and communication in both rural and urban environments.

As we move towards a more data-driven world, the role of AI becomes more inevitable. Especially with such mobile cyber-physical systems, it is critical to think about the guard rails and data-resilient technologies that remain safe for usage in human-coexistent environments. Further, parallel research in data sciences need to build tools and solutions that are trustworthy, reliable, and perhaps provable. An important aspect of that provable aspect will be the explainability of the model to describe what are the key features that had the most impact on the decision it took. As we learnt during a live talk by Olga Saukh, “the robustness of models to data drifts and sensors capability drifts will be critical to such sensors’ and data-driven algorithms’ ability to sustain equitable AI for cyber-physical systems.” It will also be critical to build solutions that understand and evaluate the data for data biases and build automated tools to understand the biases of black-box AI models. Further, human-in-the-loop models should create more accessible interfaces for the human to explain the deviation in the decision-making process that it is introducing and the logic behind it needs to be learnt by AI algorithms on-the-line. Many of the more energy hungry AI models will need energy-elastic capabilities to continuously manage the energy-accuracy tradeoff during live applications. Finally the continuous verification of these AI systems and detection of maleficent intent is critical for a safer future.

4.5 Impact of AI on IoT

Olga Saukh (TU Graz, AT), Marco Antonio Zúñiga Zamalloa (TU Delft, NL)

License  Creative Commons BY 4.0 International license
© Olga Saukh and Marco Antonio Zúñiga Zamalloa

The impact of AI on IoT research has been significant and transformative. Integration of ML models into IoT systems has led to numerous advancements across various domains, including medical, underwater and space IoT applications discussed at the seminar.

As increasingly more data is generated by IoT sensors, ML algorithms running on edge devices extract valuable insights, identify patterns, and make predictions based on the collected data. This facilitates the development of new diagnostic tools and real-time decision-making. The approach is particularly valuable in manufacturing, but also builds the core of early warning systems in medicine, precision agriculture and cattle farming.

AI-powered algorithms can help to optimize network management and communication protocols in IoT systems. This improves connectivity, enhances reliability, and reduces latency issues. AI can play a crucial role in enhancing the privacy and security of IoT systems. ML algorithms can analyze network traffic patterns and detect anomalies that indicate potential cyber threats. They can help to anonymize and encrypt sensitive IoT data, ensuring privacy protection.

Natural language interfaces and vision-based tools can enable seamless interactions between humans and IoT devices, creating more intuitive and user-friendly interfaces and experiences.

4.5.1 Challenges

Designing safe AI-based systems. AI-based systems in the IoT domain pose challenges such as distribution shift, where the real-world data encountered during deployment differs from the training data, robustness against adversarial attacks and noisy sensor data, and adaptation to dynamic environments. Addressing these challenges requires techniques such as

continuous monitoring of data distributions, domain adaptation, adversarial training, robust optimization, and online learning to ensure the reliability, robustness, and adaptability of AI models in IoT systems.

Tackling scalability and resource constraints. Scaling AI-based IoT systems to handle a large number of connected devices and manage massive data streams can be challenging. AI algorithms may require significant computational resources, which can strain the limited resources of IoT devices. Optimizing AI algorithms for resource-constrained environments becomes necessary. Since integrating intelligence into IoT devices may become an unsolvable challenge, offloading parts of data processing to the cloud is a viable option, yet has to carefully take into account limited bandwidth and increased unbounded latencies.

Guaranteeing data quality and privacy. IoT systems generate massive amounts of data, but ensuring data quality and privacy can be challenging. AI models heavily rely on high-quality data for accurate results, and ensuring data integrity, security, and privacy protection becomes crucial. Addressing these challenges requires robust data governance and security measures on edge devices.

Enabling trustworthy AI-based tools. Explainability and interpretability of AI models in IoT devices pose a significant obstacle. Service providers and regulatory bodies require transparency to trust the decisions made by AI models. However, many advanced AI techniques are inherently complex and black-box in nature, making it challenging to provide understandable explanations for their outputs. The interpretability challenge is amplified in IoT devices where computational resources are at stake, limiting the feasibility of deploying complex interpretable models. Striking a balance between accuracy and interpretability is crucial in ensuring the safe and ethical deployment of AI.

Expanding skills and expertise. Developing AI-based IoT systems requires a mix of skills and expertise in AI, IoT, data engineering, and domain knowledge. Finding professionals with interdisciplinary skills is challenging. Promoting cross-domain collaboration becomes necessary.

4.5.2 Opportunities

Domain-specific AI solutions. The IoT community plays a crucial role in advancing domain-specific AI solutions by providing insights into challenges and requirements unique to various fields. By sharing anonymized and appropriately labeled data sets collected from diverse IoT sensors, the community can enable AI researchers to develop tailored models and benchmark their performance effectively. However, while data is crucial, there is often an imbalance in focus, with more emphasis on model development rather than data collection. To address this, a culture of sharing and reusing datasets needs to be cultivated, encouraging the IoT community to contribute to open data repositories and facilitating the exchange of valuable data for the advancement of AI in IoT applications. Data reuse is often difficult in the IoT community, since in many cases shared data tightly reflects peculiarities of a specific setup. Therefore sharing large rigorous and well-curated datasets for selected applications and updating these data over time can ultimately benefit the entire IoT ecosystem.

Sensors and AI-based edge systems. There is a need for a stronger hardware-model co-design to optimize sensors and enhance overall AI-based system performance. In the IoT domain, where multimodal sensing is prevalent, we call for exploring the appropriate modeling and representation learning techniques for different modalities. The current emphasis in AI

research on images and text data should be expanded to encompass the multimodal nature of IoT data, which includes a wide variety of sensors. Additionally, the IoT community should question the applicability of conventional machine learning practices inherited from other domains. The latest results in sparsity suggest that the best performing sparsity methods for computer vision fail on large language models. It is essential for the IoT community to focus on identifying foundational models specific to IoT data that can capture the underlying dynamics and characteristics of the IoT environments.

Real-world use cases and knowledge sharing. Collaboration between the IoT community and AI researchers opens up opportunities for innovation and the development of tailored solutions that meet the specific demands of diverse IoT applications. The IoT community can share their real-world use cases and challenges with AI researchers, providing insights into specific domain requirements and practical scenarios. This sharing can take various forms, including data, generative models, simulations, deployment traces and lessons learned. By collaborating with AI researchers, the IoT community can help identify and address the challenges associated with IoT-related issues such as data silos, distribution shifts, domain adaptation and the need for continual learning. Furthermore, integrating simulators, game engines, and digital twins into IoT systems can enhance the development and testing of AI algorithms in realistic and open-world environments. The combination of physics and machine learning models also holds promise for advancing AI in the IoT domain, enabling more accurate and interpretable solutions.

Performance evaluation of AI-based systems. AI algorithms need to perform efficiently and effectively in IoT environments. The IoT community can actively participate in evaluating AI model compression techniques for their suitability for resource-constrained IoT devices and various hardware architectures, considering factors like energy consumption, computational requirements, and latency. This feedback can guide AI researchers in optimizing their algorithms and optimization techniques for practical settings. In addition, the verification of edge AI, running in parallel to current efforts, will emerge as a distinct field that addresses the unique requirements and considerations of AI algorithms deployed at the edge. The focus on performance evaluation and verification will contribute to the development of robust and efficient AI-based systems for IoT applications.

Ethical considerations. Deploying AI in IoT systems raises ethical concerns, such as privacy, security, bias, and transparency. The decision-making capabilities of AI algorithms should align with ethical and legal frameworks. Ensuring fairness, accountability, and transparency in AI-based IoT systems requires careful design, monitoring, and governance. The IoT community plays a crucial role in establishing standards and protocols for IoT devices and systems. Similarly, the AI community can contribute to defining standards for AI-enabled IoT devices, ensuring interoperability, data exchange, and security. Collaborative efforts can lead to the development of standardized interfaces and frameworks that facilitate the integration of AI and IoT technologies.

Participants

- Arash Asadi
TU Darmstadt, DE
- Emmanuel Baccelli
FU Berlin, DE
- Victor Bahl
Microsoft Corporation –
Redmond, US
- Bastian Bloessl
TU Darmstadt, DE
- Justin Chan
University of Washington –
Seattle, US
- Akshay Gadre
University of Washington –
Seattle, US
- Junfeng Guan
EPFL Lausanne, CH
- Oliver Hahm
Frankfurt University of Applied
Sciences, DE
- Richard Han
Macquarie University –
Sydney, AU
- Karin Anna Hummel
Johannes Kepler Universität
Linz, AT
- Tianxing Li
Michigan State University –
East Lansing, US
- Andra Lutu
Telefónica – Madrid, ES
- Andrea Maioli
Polytechnic University of
Milan, IT
- RangaRao Venkatesha Prasad
TU Delft, NL
- Alberto Quattrini Li
Dartmouth College –
Hanover, US
- Olga Saukh
TU Graz, AT
- Longfei Shangguan
University of Pittsburgh, US
- Deepak Vasisht
University of Illinois –
Urbana-Champaign, US
- Chenren Xu
Peking University, CN
- Rong Zheng
McMaster University –
Hamilton, CA
- Xia Zhou
Columbia University –
New York, US
- Marco Zimmerling
TU Darmstadt, DE
- Marco Antonio Zúñiga
Zamalloa
TU Delft, NL

