DAGSTUHL
REPORTS

Volume 12, Issue 1, January 2022

Dagstuhl Reports, Vol. 12, Issue 1 ISSN 2192-5283

ISSN 2192-5283

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

Publication date
August, 2022

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`.

Digital Object Identifier: 10.4230/DagRep.12.1.i

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

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

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

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

Report from Dagstuhl Seminar 22021

Mobility Data Science

Mohamed Mokbel*[1], Mahmoud Sakr*[2], Li Xiong*[3], Andreas Züfle*[4], Jussara Almeida[5], Taylor Anderson[6], Walid Aref[7], Gennady Andrienko[8], Natalia Andrienko[9], Yang Cao[10], Sanjay Chawla[11], Reynold Cheng[12], Panos Chrysanthis[13], Xiqi Fei[14], Gabriel Ghinita[15], Anita Graser[16], Dimitrios Gunopulos[17], Christian Jensen[18], Joon-Sook Kim[19], Kyoung-Sook Kim[20], Peer Kröger[21], John Krumm[22], Johannes Lauer[23], Amr Magdy[24], Mario Nascimento[25], Siva Ravada[26], Matthias Renz[27], Dimitris Sacharidis[28], Cyrus Shahabi[29], Flora Salim[30], Mohamed Sarwat[31], Maxime Schoemans[32], Bettina Speckmann[33], Egemen Tanin[34], Yannis Theodoridis[35], Kristian Torp[36], Goce Trajcevski[37], Marc van Kreveld[38], Carola Wenk[39], Martin Werner[40], Raymond Wong[41], Song Wu[42], Jianqiu Xu[43], Moustafa Youssef[44], Demetris Zeinalipour[45], Mengxuan Zhang[46], and Esteban Zimányi[47]

1 University of Minnesota – Minneapolis, USA. mokbel@umn.edu
2 Université Libre de Bruxelles – Brussels, Belgium. mahmoud.sakr@ulb.be
3 Emory University – Atlanta, USA. lxiong@emory.edu
4 George Mason University – Fairfax, USA. azufle@gmu.edu
5 Federal University of Minas Gerais – Brazil. jussara@dcc.ufmg.br
6 George Mason University – Fairfax, USA. tander6@gmu.edu
7 Purdue University – West Lafayette, USA. aref@cs.purdue.edu
8 Fraunhofer IAIS – St. Augustin, Germany.
gennady.andrienko@iais.fraunhofer.de
9 Fraunhofer IAIS – St. Augustin, Germany.
natalia.andrienko@iais.fraunhofer.de
10 Kyoto University – Kyoto, Japan. yang@i.kyoto-u.ac.jp
11 Qatar Computing Research Institute – Doha, Qatar. schawla@hbku.edu.qa
12 University of Hong Kong – Hong Kong, China. ckcheng@cs.hku.hk
13 University of Pittsburgh – Pennsylvania, USA. panos@cs.pitt.edu
14 George Mason University – Fairfax, USA
15 University of Massachusetts at Boston – Boston, USA. Gabriel.Ghinita@umb.edu
16 Austrian Institute of Technology – Vienna, Austria. Anita.Graser@ait.ac.at
17 University of Athens – Greece. dg@di.uoa.gr
18 Aalborg University – Denmark. csj@cs.aau.dk
19 Pacific Northwest National Laboratory – USA. joonseok.kim@pnnl.gov
20 AIST – Tokyo Waterfront, Japan. ks.kim@aist.go.jp
21 University of Kiel – Germany. pkr@informatik.uni-kiel.de
22 Microsoft – Redmond, USA. jckrumm@microsoft.com
23 HERE Technologies – Germany. johannes.lauer@here.com
24 University of California – Riverside, USA. amr@cs.ucr.edu
25 University of Alberta – Edmonton, Canada. mario.nascimento@ualberta.ca
26 Oracle Corp. – Nashua, USA. siva.ravada@oracle.com
27 University of Kiel – Germany. mr@informatik.uni-kiel.de
28 Université Libre de Bruxelles – Brussels, Belgium. dimitris.sacharidis@ulb.be
29 University of Southern California – Log Angeles, USA. shahabi@usc.edu
30 University of New South Wales – Sydney, Australia. flora.salim@unsw.edu.au
31 Arizona State University – Tempe, USA. msarwat@asu.edu

32 Université Libre de Bruxelles – Brussels, Belgium. maxime.schoemans@ulb.be
33 TU Eindhoven – Netherlands. b.speckmann@tue.nl
34 University of Melbourne – Australia. etanin@unimelb.edu.au
35 University of Piraeus – Greece. ytheod@unipi.gr
36 Aalborg University – Denmark. torp@cs.aau.dk
37 Iowa State University – USA. gocet25@iastate.edu
38 Utrecht University – Netherlands. m.j.vankreveld@uu.nl
39 Tulane University – New Orleans, USA. cwenk@tulane.edu
40 Technical University of Munich – Munich, Germany. martin.werner@tum.de
41 Hong Kong Univ. of Science & Technology – Hong Kong, China. raywong@cse.ust.hk
42 Université Libre de Bruxelles – Brussels, Belgium. song.wu@ulb.be
43 Nanjing University of Aeronautics and Astronautics, China. jianqiu@nuaa.edu.cn
44 AUC and Alexandria University – Egypt. moustafa.youssef@gmail.com
45 University of Cyprus – Nicosia, Cyprus. dzeina@cs.ucy.ac.cy
46 Iowa State University – USA. mxzhang@iastate.edu
47 Université Libre de Bruxelles – Brussels, Belgium. esteban.zimanyi@ulb.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22021 "Mobility Data Science". This seminar was held January 9-14, 2022, including 47 participants from industry and academia. The goal of this Dagstuhl Seminar was to create a new research community of mobility data science in which *the whole is greater than the sum of its parts* by bringing together established leaders as well as promising young researchers from all fields related to mobility data science.

Specifically, this report summarizes the main results of the seminar by (1) defining Mobility Data Science as a research domain, (2) by sketching its agenda in the coming years, and by (3) building a mobility data science community. (1) Mobility data science is defined as spatiotemporal data that additionally captures the behavior of moving entities (human, vehicle, animal, etc.). To understand, explain, and predict behavior, we note that a strong collaboration with research in behavioral and social sciences is needed. (2) Future research directions for mobility data science described in this report include a) mobility data acquisition and privacy, b) mobility data management and analysis, and c) applications of mobility data science. (3) We identify opportunities towards building a mobility data science community, towards collaborations between academic and industry, and towards a mobility data science curriculum.

Seminar January 9–14, 2022 – http://www.dagstuhl.de/22021
2012 ACM Subject Classification Information systems → Information systems applications; Information systems → Data management systems
Keywords and phrases Spatio-temporal, Tracking, Privacy, Behavior, Data cleaning, Data management, Analytics
Digital Object Identifier 10.4230/DagRep.12.1.1

* Editor / Organizer

Mobility Data Science, *Dagstuhl Reports*, Vol. 12, Issue 1, pp. 2–34
Editors: Mohamed Mokbel, Mahmoud Sakr, Li Xiong, and Andreas Züfle, et al.
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Mobility data is typically available in the form of sequences of location points with time stamps, that are generated by location tracking devices. The use of mobility data has traditionally been linked to transportation industry. Nowadays, with the availability of GPS-equipped mobile devices and other inexpensive location tracking technologies, mobility data is collected and published ubiquitously, leading to large data sets of volunteered geographic information (VGI).

In general, mobility data science is the science of transforming mobility data into (actionable) knowledge. This knowledge is critical towards solutions for traffic management, disease pandemic mitigation, micro-mobility (e.g., shared bikes and scooters), health monitoring, logistics (e.g., delivery services), to mention a few.

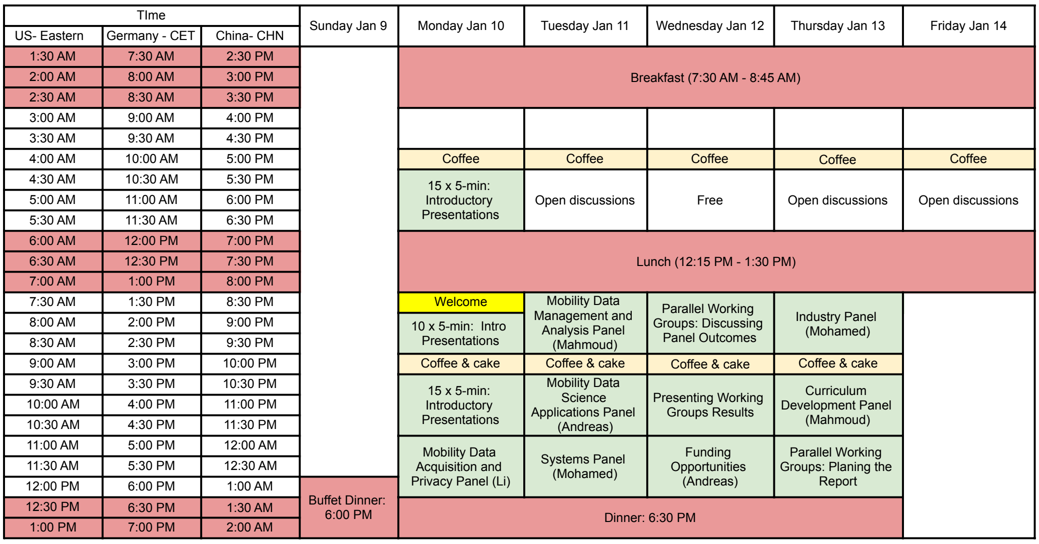
Despite the common goal of acquiring, managing, and generating insights from mobility data, the mobility data science community is largely fragmented, developing solutions in silos. It stems from a range of disciplines with expertise in moving object data storage and management, geographic information science, spatiotemporal data mining, ubiquitous computing, computational geometry and more. Furthermore, there is a disconnect in both industry and science between mobility data scientists and domain scientists or end users for which solutions are designed. Therefore, the goal of this Dagstuhl Seminar was to bring together and recognize the mobility data science community as an interdisciplinary research field, strengthen the definition of mobility data science, and together explore challenges and opportunities in the field. The seminar had two objectives: (1) to build a new research community of mobility data science as amalgamation of the several communities who have been looking at mobility data, and (2) to draft a research agenda for mobility data science. This dagstuhl seminar was the first towards these objectives. The consensus of the participants is that more events will be needed in the future to continue the community building effort.

Seminar Program

The seminar was held in the week of January 9 – 14 , 2022. It had 47 participants specialized in different topics: data management, mobility analysis, geography, privacy, urban computing, systems, simulation, indoors, visualization, information integration, and theory. Due to COVID-19, the seminar took place in hybrid mode, with 8 onsite, and 39 remote participants. Despite the challenge of different time zones of the participants, all sessions were attended by at least 37 participants.

The seminar program is given in Figure 1. In the first day, every participant gave a five-minutes self introduction, research interests, and position statement on mobility data science. The rest of the program consisted of panels, and open discussions. To work around the time zone challenge, the seminar activities were centered in the afternoon of Dagstuhl, which was still possible for the Eastern and CHN time zones. The open discussions slots were planned ad-hoc during the seminar. In particular, the slot on Tuesday was used to define what is mobility data science, or more precisely what is the scope of work of this community.

All working group and panel discussions were moderated to converge towards the seminar goals of defining a research agenda and building a community. The results are summarized in this report.



TIme			Sunday Jan 9	Monday Jan 10	Tuesday Jan 11	Wednesday Jan 12	Thursday Jan 13	Friday Jan 14
US- Eastern	Germany - CET	China- CHN						
1:30 AM	7:30 AM	2:30 PM		Breakfast (7:30 AM - 8:45 AM)				
2:00 AM	8:00 AM	3:00 PM						
2:30 AM	8:30 AM	3:30 PM						
3:00 AM	9:00 AM	4:00 PM						
3:30 AM	9:30 AM	4:30 PM						
4:00 AM	10:00 AM	5:00 PM		Coffee	Coffee	Coffee	Coffee	Coffee
4:30 AM	10:30 AM	5:30 PM		15 x 5-min: Introductory Presentations	Open discussions	Free	Open discussions	Open discussions
5:00 AM	11:00 AM	6:00 PM						
5:30 AM	11:30 AM	6:30 PM						
6:00 AM	12:00 PM	7:00 PM		Lunch (12:15 PM - 1:30 PM)				
6:30 AM	12:30 PM	7:30 PM						
7:00 AM	1:00 PM	8:00 PM						
7:30 AM	1:30 PM	8:30 PM		Welcome	Mobility Data Management and Analysis Panel (Mahmoud)	Parallel Working Groups: Discussing Panel Outcomes	Industry Panel (Mohamed)	
8:00 AM	2:00 PM	9:00 PM		10 x 5-min: Intro Presentations				
8:30 AM	2:30 PM	9:30 PM						
9:00 AM	3:00 PM	10:00 PM		Coffee & cake	Coffee & cake	Coffee & cake	Coffee & cake	
9:30 AM	3:30 PM	10:30 PM		15 x 5-min: Introductory Presentations	Mobility Data Science Applications Panel (Andreas)	Presenting Working Groups Results	Curriculum Development Panel (Mahmoud)	
10:00 AM	4:00 PM	11:00 PM						
10:30 AM	4:30 PM	11:30 PM						
11:00 AM	5:00 PM	12:00 AM		Mobility Data Acquisition and Privacy Panel (Li)	Systems Panel (Mohamed)	Funding Opportunities (Andreas)	Parallel Working Groups: Planing the Report	
11:30 AM	5:30 PM	12:30 AM						
12:00 PM	6:00 PM	1:00 AM	Buffet Dinner: 6:00 PM					
12:30 PM	6:30 PM	1:30 AM		Dinner: 6:30 PM				
1:00 PM	7:00 PM	2:00 AM						

Figure 1 Dagstuhl Seminar on Mobility Science – Program.

Organization and Panels

To accommodate for the hybrid mode and the time zone differences, we opted to let the participants choose to participate in one of the following three thematic working groups, each having 14-17 members and led by one of the seminar co-organizers:

- **Seminar co-organizers:** Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle
- **Working Group 1:** Mobility Data Acquisition and Privacy
  The scope includes the full cycle of obtaining and preparing mobility data for further processing. Examples include innovative ways of data collection, crowdsourcing, simulation, data uncertainty, data cleaning, and data visualization. It also includes innovative ways of ensuring mobile users privacy as a means of encouraging users to share their data. Results of Working Group 1 are found in Section 4
- **Working Group 2:** Mobility Data Management and Analysis
  This includes the full data pipeline from modelling, indexing, query processing/optimization, and data analysis. Existing solutions for mobility data management were discussed and a way forward for a next generation system for mobility data management was conceived. Results of Working Group 2 are found in Section 5.
- **Working Group 3** Mobility Data Science Applications
  This working group discussed the broader impacts of mobility data science to improve understanding of human behavior, urban sustainability, improving traffic conditions, health, and situational awareness. Specific applications towards these broader impacts including map making, contact tracing, pandemic preparedness, indoor navigation, and marine transportation were discussed. Results of Working Group 3 are found in Section 6.

For each working group a dedicated panel session was organized which was attended by all seminar participants. In addition, two parallel working group sessions were held for discussions and for planning the writing of this report. The working groups presented and further discussed their results with all participants on Wednesday. Four cross-cutting panels discussed the topics of systems, funding opportunities, industry involvement, and curriculum development. All panels started with presentations of panelists as listed below, seven minutes each, where they expressed their positions concerning questions given by the panel moderator. The rest of the panel time opened the discussion to all participants.

- Mobility Data Acquisition and Privacy Panel. Moderator: Li Xiong
  Panelists: Gennady (& Natalia) Andrienko, Kyoung-Sook Kim, John Krumm, Cyrus Shahabi
- Mobility Data Management and Analysis Panel. Moderator: Mahmoud Sakr
  Panelists: Walid Aref, Panos Chrysanthis, Christian Jensen, Yannis Theodoridis
- Mobility Data Science Applications Panel. Moderator: Andreas Züfle
  Panelists: Sanjay Chawla, Flora Salim, Moustafa Youssef, Demetris Zeinalipour
- Systems Panel. Moderator: Mohamed Mokbel
  Panelists: Walid Aref, Dimitrios Gunopulos, Cyrus Shahabi, Esteban Zimányi
- Funding Opportunities Panel. Moderator: Andreas Züfle
  Panelists: Johannes Lauer, Mario Nascimento, Matthias Renz, Carola Wenk
- Industry Panel. Moderator: Mohamed Mokbel
  Panelists: John Krumm, Johannes Lauer, Siva Ravada, Mohamed Sarwat
- Curriculum Development Panel. Moderator: Mahmoud Sakr
  Panelists: Anita Graser, Marc van Kreveld, Martin Werner, Esteban Zimányi

2 Table of Contents

3 What is Mobility Data Science?

During initial seminar discussions, a controversial statement was made that many problems that the spatiotemporal data community is working on do not have any real-world applications. An example was the problem of location prediction, where the number of visitors for point of interests are predicted over time. This controversial statement led to a vivid discussion. While there was disagreement about the broader impacts of the problem of location prediction and other spatiotemporal data science problems, it was widely agreed that a useful and open research challenge is to understanding the underlying behavior: Why does the number of visitor changes? How can explain the underlying human behavior? And how can we leverage this understanding to predict what-if scenarios to maximize future visitor numbers at a point of interest.

Thus, highlighting need to go beyond the question of *where* will users be, but also *why* they will be there. Understanding the underlying behavior of users would allow us to explain trends. For example, reduced popularity of a coffee shop could be explained by a nearby competitor attracting customers with a special offer. These explanation of trends can then be used to take actions (such as offering a similar or better offer) to not only predict future trends, but to change them.

But to answer "why" users visit places, we have to think beyond spatial and spatiotemporal data science to understanding the underlying human behavior. Adding a component (human) behavior allows us to transition from traditional spatiotemporal data to mobility data science.

Therefore, we formalize Mobility Data Science as follows.

▶ **Definition 1** (Mobility Data)**.** Spatiotemporal data capturing behavior of moving entities.

▶ **Definition 2** (Mobility Data Science)**.** The science of mobility data.

The additional consideration of (human) behavior in spatiotemporal data is challenging. As Physics nobel laureate Murray Gell-Mann once famously said: "Think how hard physics would be if particles could think". By adding a dimension of behavior, we're allowing the entities in our universe of discourse to think. The seminar agreed that experts in social sciences will be paramount to explain and model human behavior, to mine knowledge from mobility data, and to create a mobility data science community. Cross-disciplinary research is challenging because it requires collaboration between domain experts that may address research in very different manners. However, cross-disciplinary research ensures that the solutions provided by the mobility data science community are relevant for society. Another clear benefit from working with experts from other domains is that this creates new research ideas. The mobility data science community should encourage cross-disciplinary research by ensuring prestigious outlets publish industry papers and best-practice papers.

4 Future Research Agenda: Mobility Data Acquisition and Privacy

Results of the first working group on mobility data acquisition and privacy are presented in this section. Specifically, Section 4.1 discusses challenges related to data acquisition, Section 4.2 describes issues on mobility data quality, Section 4.3 discussed bias in mobility data, and Section 4.4 describes privacy threats for mobility data.

4.1 Data Acquisition

Acquisition in the context of this report refers to acquiring mobility data. This is often necessary for the research we want to do – as a way to train models, assess the quality of our algorithms, and analyze the data for patterns and relationships with other data.

Mobility data can take different forms. We normally derive mobility data from humans, but it can also come from animals. Inanimate objects can move as well (e.g., planets, soccer balls, and leaves), but we tend to think of mobility data as coming from something whose motion is intentional or at least guided with intention, such as someone in a bus or autonomous vehicle. The exact data included in mobility data can vary. Perhaps the most common example is timestamped coordinates, such as latitude/longitude. If the timestamps are missing, we would likely call it population data or occupancy data rather than mobility data. The coordinates could instead be replaced by higher level abstractions like points of interest (POI), activities, or other location-specific properties.

The main issue with acquisition of mobility data is how we can get mobility data for our research. Getting the data takes time and effort, especially if it is annotated with human-generated abstractions like activity. There have been efforts to gather and share mobility data, such as GeoLife from Microsoft Research Beijing which covers 182 different users. But there is not yet a massive, shared mobility dataset for human motion. One trend among general research papers over the past several years is to encourage or require authors to make their data publicly available. As an example, there is a repository of animal mobility data at `https://www.movebank.org/cms/movebank-main`. Besides encouraging sharing in individual papers, we should consider starting a publication venue for mobility datasets (or more generally spatial datasets), similar to the journal “Nature Scientific Data”, which is a “journal for descriptions of scientifically valuable datasets.”

Human mobility data is sensitive (considered “personally identifiable information” (PII) by some companies), so we cannot necessarily share all human mobility data. We note that part of this Dagstuhl Seminar also concentrated on privacy, which is at odds with data sharing. One approach to the privacy problem is to create a community-based open source project where we enroll people who are willing to log and share their location data, possibly with privacy guarantees for the contributors. However, such data would be biased toward people who are willing to share their location data. One solution for the privacy problem is for the data owner to accept code from outside, run it on their proprietary data, and then return the results.

Another way to get mobility data is to generate it synthetically. While numerous synthetic trajectory and movement data generators have been proposed in the past, there is not a generally accepted synthetic generator for mobility data, and no consensus on whether or not such data would be acceptable as the sole test set for a research project. Some have used a synthetic dataset as *one* of their test datasets for a publication. And sometimes a specialized simulation is used for testing an algorithm, e.g., flocking behavior or location-based social networks.

Our community could take a lesson from generative adversarial networks (GANs) in the deep learning community where the network inspects its artificially generated results for realism. However, we do not yet know how to measure the realism of mobility data. If synthetic mobility data is too realistic, it may invade someone’s privacy if it, for instance, shows where members of a given household actually visit.

In the end, the research community will have to agree on the suitability of synthetic mobility data for research. This agreement could be an explicit statement or could grow naturally from the pool of papers that are successfully published.

4.2 Data Quality

There exists different procedures for collecting mobility data:

- **Time-based**: positions of movers are recorded at regularly spaced time moments.
- **Change-based**: a record is made when a mover's position, speed, or movement direction differs from the previous one.
- **Location-based**: a record is made when a mover enters or comes close to a specific place, e.g., where a sensor is installed.
- **Event-based**: positions and times are recorded when certain events occur, in particular, when movers perform certain activities such as cellphone calls or taking photos.
- **Various combinations** of these basic approaches. In particular, GPS tracking devices may combine time-based and change-based re-cording: the positions are measured at regular time intervals but recorded only when a significant change of position, speed, or direction occurs.

To identify data errors, it is necessary to consider systematically the relevant properties of mobility data, including:

- Mover set properties:
  - number of movers: a single mover, a small number of movers, a large number of movers;
  - population coverage: whether there are data about all movers of interest for a given territory and time period or only for a sample of the movers;
  - representativeness: whether the sample of movers is representative, i.e., has the same distribution of properties as in the whole population, or biased towards individuals with particular properties.
- Temporal properties:
  - temporal resolution: the lengths of the time intervals between the position measurements;
  - temporal regularity: whether the length of time intervals between measurements is constant or variable;
  - temporal coverage: whether the measurements were made during the whole time span of the data or in a sample of time units, or there were intentional or unintentional breaks in the measurements;
  - time cycles coverage: whether all positions of relevant time cycles (daily, weekly, seasonal, etc.) are sufficiently represented in the data, or the data refer only to subsets of positions (e.g., only work days or only daytime), or there is a bias towards some positions.
- Spatial properties:
  - spatial resolution: the minimal change of position of an object that can be reflected in the data;
  - spatial precision: whether the positions are defined as points (by exact coordinates) or as locations having spatial extents (e.g., areas). For example, the position of a mobile phone call is typically a cell in a mobile phone network;
  - spatial coverage: are positions recorded everywhere or, if not, how are the locations where positions are recorded distributed over the studied territory (in terms of the spatial extent, uniformity, and density)?

Further sources of data errors and uncertainty relate to

- data collection procedure:
  - position exactness: How exactly could the positions be determined? Thus, a movement sensor may detect an object within its range but may not be able to determine the exact coordinates of the object within its detection area. In this case, the position of the sensor stands in for the object's true the position;

- positioning accuracy, or how much error may be in the measurements;
- missing positions: in some circumstances, object positions cannot be determined, leading to gaps in the data;
- meanings of the position absence: whether absence of positions corresponds to stops, or to conditions when measurements were impossible, or to device failure, or to private information that has been removed.

Irrespective of the collection method and device settings, there is also indispensable uncertainty in movement data (and, more generally, any time-related data) caused by their discreteness. Since time is continuous, the data cannot refer to every possible instant. For any two successive instants t1 and t2 referred to in the data there are moments in between for which there are no data. Therefore, one cannot know definitely what happened between t1 and t2. Movement data with fine temporal and spatial resolution give a possibility of interpolation, i.e., estimation of object positions between the measured positions. In this way, the continuous path of the mover can be approximately reconstructed by interpolation.

Movement data that do not allow valid interpolation may be called episodic. Episodic data are usually produced by location-based and event-based collection methods but may also be produced by time-based methods when the position measurements cannot be done sufficiently frequently, for example, due to the limited battery lives of the devices. Thus, when tracking movements of wild animals, ecologists have to reduce the frequency of measurements to be able to track the animals over longer time periods.

According to the structure and properties of mobility data, we can identify the following classes of errors in movement data:

- temporal properties, e.g., missing data during some time intervals;
- spatial properties, e.g., errors in positions, absence of data in some parts of territory;
- mover identity properties, e.g., misspelled or duplicate identifiers; or no identifiers at all;
- data collection properties, e.g., results of restricting positions to a given bounding rectangle.

For identifying data quality issues, it is necessary to consider all major components of the data structure, namely identities, spatial locations, times, and thematic attributes, and their combinations. Any unexpected regularity or irregularity of distributions requires attention and explanation, as such patterns may indicate sampling bias or other kinds of errors. Calculation of derived attributes (e.g., speed, direction) and aggregates over space, time, and categories of objects provides further distributions to be assessed.

4.3 Bias in Data

All datasets are biased, examples include:

- App data and mobile phone network data are biased against people who don't use smart phones or use prepaid plans
- Most traffic counting sensors are installed to count cars but do not count pedestrians, cyclists, wheelchairs, (e-)scooters, and similar
- Object-detection in video-based systems depends on labels in training data, which often don't include new mobility options such as (e-)scooters
- Surveys are biased towards people who are willing and able to fill surveys, i.e., interested, literate, sufficient free time, reachable (e.g., for phone surveying)

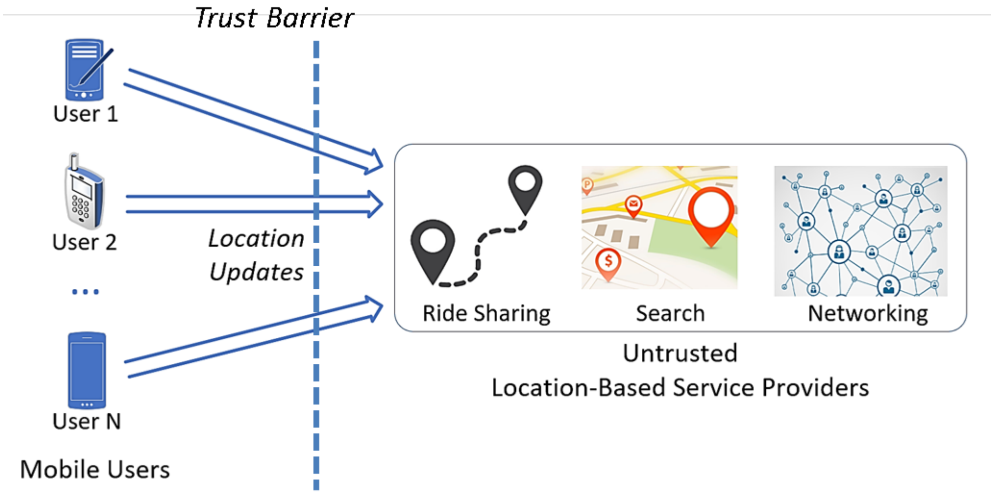


Figure 2 Location protection without trusted third party.

- Cells in mobile phone networks vary widely in size. Trips that stay within a single cell cannot be detected. This affects rural areas with larger cells more than urban areas.
- Volunteered tracking data is biased towards technically savvy people
- Sports tracking data is biased towards health conscious middle and upper class

It is important to acknowledge and quantify the bias in mobility data sets to ensure that actions and policies that are based on mobility data science results are equitable and fair and include vulnerable populations.

4.4 Privacy

We divide our discussion on privacy into two settings: local settings and central settings. In the local setting (as shown in Figure 2), the mobile service provider is not trusted. Hence, mobile users need to protect their location information (e.g., via perturbation) before uploading them to the service provider in exchange for location based services (LBS). The services include ride sharing, spatial crowdsourcing, and POI search. The aggregate locations from all users can be also used to support location based analytics. In the central setting (as shown in Figure 3), the mobile service provider is trusted and has access to users' mobility data. It acts as a trusted data curator and shares some form of the mobility data with a third party or the public which are untrusted. The shared data can be aggregate statistics, sanitized or anonymized version of the data, machine learning models trained from the data, or synthetic data generated based on the original data which can support a variety of applications using mobility data (as we discussed in the application section). The privacy goal is to protect against inference risks to the original data from the shared data.

4.4.1 Threat Models and Privacy Definitions

The first challenge for mobility data privacy protection is the need to understand the threat models and adopt or define proper criteria by which to enforce privacy. We need to define first what needs to be protected (i.e., the sensitive information). This may vary for different

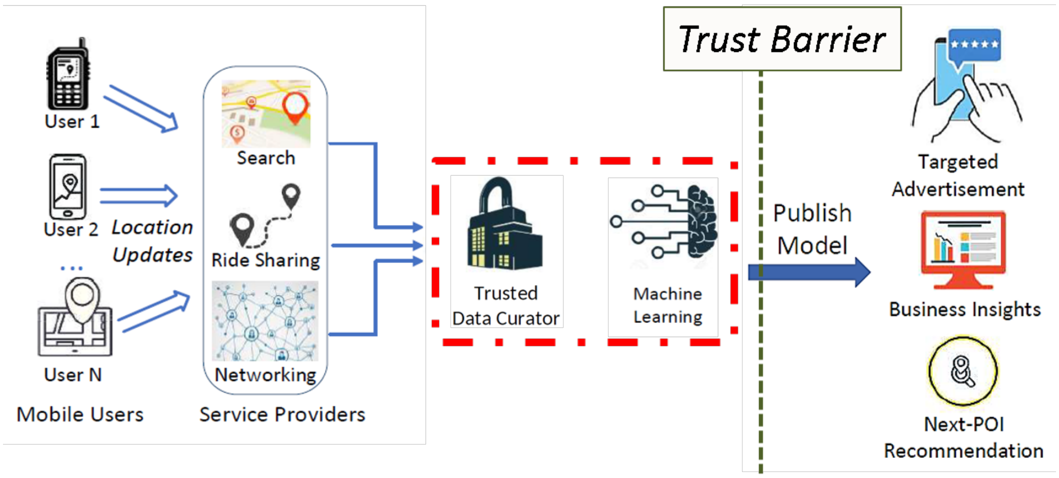


Figure 3 Location Protection with Trusted Third Party.

mobile users and applications. It may be the exact location coordinates, association of coordinate with a sensitive place, co-location of two users, or spatiotemporal activities of a user (e.g., stay at a place, or a trajectory). When defining privacy models and designing subsequent privacy mechanisms, there will (almost always) be an attack to a privacy model which is based on side channel information exploitation. A clear threat model needs to be defined. Relation to cybersecurity and cyber-defense could be interesting and provides a research direction.

Syntactic Privacy Approaches. Early work on privacy for location data focused on location-based services (LBS). Typical use-case scenarios are those where a user wants to retrieve points of interest in her proximity, e.g., nearest gas station. Query types are limited to range or nearest-neighbor queries. In this setting, the protection model considered was one where a user's whereabouts are hidden, or *cloaked*, among a set of $k-1$ other individuals, according to the concept of spatial k-anonymity. This is a syntactic model, which works well if the adversary does not have access to background knowledge, but fails to provide the required amount of protection when the adversary can use auxiliary information to eliminate some of the locations in a cloaking set. *Cryptographic Approaches.* Later on, formal protection guarantees in the LBS setting were achieved with the help of encryption approaches. Specifically, existing work encrypts the data and query domain, and answers queries such as nearest neighbors in the transformed space. Other work that provides encryption-strength protection employs Private Information Retrieval (PIR) in conjunction with Voronoi diagrams to answer privately nearest-neighbor queries. *Differential Privacy (DP).* The introduction of DP and its *Geo-Indistinguishability (GeoInd)* counterpart specifically designed for location data have opened the field for new approaches that provide formal protection with statistical guarantees. In the case of GeoInd which is designed for the local setting (as shown in Figure 1), one protects against pinpointing the exact location of a targeted user from the reported (perturbed) location. The level of protection is controlled with the help of a privacy parameter. Each user can perturb their own location before reporting it according to well-defined protection mechanisms that provide GeoInd. Later works extended geoInd to account for temporal correlations between consecutive locations of mobile users, protection of customizable spatiotemporal activities instead of raw locations or trajectories. Challenges remain in providing customizable and rigorous privacy notions and mechanisms that provide high utility for a variety of applications.

For the global setting, several types of aggregate computation can be performed privately using DP-compliant mechanisms. One can even train a machine learning model in a private way. Mobility data introduces unique challenges to applying standard DP techniques due to its spatiotemporal correlations which often results in increased privacy cost due to privacy composition and the difficulties in modeling the correlations.

4.4.2 Privacy-Utility Tradeoff and Emerging Applications

When designing privacy mechanisms, it is important to consider the utility of the intended applications. For LBS (as typical in the local setting), the utility needs to be measured by the precision or accuracy of the LBS queries such as range queries for POI search. For aggregate data analytics and machine learning applications using mobility data (in both local setting and central setting), the utility need to be measured by the accuracy of the statistics (e.g., frequency or density estimation), the trained model, or the fidelity of the synthetic data.

The emergence of novel computing paradigms like spatial crowdsourcing led to interesting technical approaches and challenges to achieving privacy. For instance, existing work investigates how one can perform assignment between spatial tasks and workers in crowdsourced systems, without compromising the location privacy of the participants involved. Specifically, the matching is performed based on a DP-compliant algorithm that ensures that an adversary cannot determine with significant probability if a particular individual took part in the crowdsourced system or not.

Another important paradigm shift in terms of type of supported queries occurred with the emergence of techniques for digital contact tracing, which are important in controlling the spread of pandemics, such as COVID-19. In this context, it is less important to capture exact user locations, and what is required is to establish co-location relationships. Determining whether two individuals have been in close proximity or not in a privacy preserving way is a challenging and important problem. For example, while geoInd may be used, the perturbation may introduce false positives and false negatives, a careful balance is needed if the result will be used to notify potential users at risk. In addition, privacy compositions may further increase privacy cost or degrade the utility when considering trajectories of users.

4.4.3 Societal Education

An important challenge of mobility data privacy is to improve explainability of privacy definitions and mechanisms. DP-compliant algorithms and DP-fashioned location privacy models (such as Geo-Ind) as described earlier use privacy parameters to control the tradeoff between privacy guarantee and the utility of the private outputs. However, there is a significant gap between the theory and practice of DP: we lack principles and guidelines for choosing privacy parameters when collecting or processing mobility data using DP techniques in the real world.

The parameter ϵ of DP is mathematically defined but not well-aligned with the stakeholders' beneficial interests. Theoretically, ϵ of DP represents the bound of marginal adversarial beliefs conditioned on the private output. It is known that the privacy parameters of DP are not absolute but rather relative measures of privacy. That is to say, even for the same ϵ, the privacy guarantees enforced by DP could be different based on the datasets and algorithms at hand. In addition, the ϵ is not always linked to a specific privacy risk for the users (such as "the probability that an attacker can correctly infer my data") or a precise utility level for data analyzers (such as "the accuracy of the DP-ML model"). It may arouse distrust towards DP technology without explaining why the specific privacy parameters are chosen and what the potential privacy risk is for the users.

To promote the acceptance of DP technology in practice, we should establish principles, design guidelines, and provide tools for explaining DP's protection and limitation from stakeholders' practical interests. For example, we can help data contributors understand the privacy risk (such as membership inference attacks or reconstruction attacks) under different privacy parameters given a concrete DP algorithm; we can also design efficient methods to visualize how data analyzers' utility metrics (such as MSE or model accuracy) may change along with different privacy parameters.

Towards a privacy and ethics review board, it was discussed that every research project needs to be evaluated in privacy/quality management. This may create a need for an ethics commissions in mobility data science.

5 Future Research Agenda: Mobility Data Management and Analysis

This section discusses the priority topic of research related to the full mobility data pipeline from modelling, indexing, query processing/optimization, and data analysis. It summarizes the discussions that took place within the Working Group of data management and analysis, and were then elaborated in a general discussion between all seminar participants.

5.1 The Architectural Challenges of Mobility Data Management

5.1.1 Hybrid Batch Real-time Data Management and Analytics

Today's requirements for mobility data management and analytics must combine both batch and real-time data. For example, a common requirement is to visualize the position of a fleet of vehicles in real time (which only requires to access the most recent positions of the vehicles), but at the same time to perform batch analytics on the full trajectory of these vehicles (for example to assess whether the trajectories exhibit ssome unexpected behavior).

Since the 1990s, the need to have both real-time and historical data has led to the development of the data warehouse domain, where operational databases cover the real-time Online Transaction processing (OLTP) while data warehouses cover the historical Online Analytical Processing (OLAP). However, having two different systems for the two kinds of workloads is very costly, and for this reason a new approach referred to as Hybrid Transactional and Analytical Processing (HTAP) has been recently proposed. This approach aims at removing the need for Extraction Transformation and Loading (ETL) process and enables real-time analytical queries on current data. The new generation of NewSQL database management systems aims at achieving this approach. However, research work is currently being done in order to solve all the challenges posed by the HTAP approach, especially in the context of big data.

The situation is similar for mobility data since the training part of ML requires offline processing of large training datasets, while we also need to do real-time prediction for numerous simultaneous moving objects (e.g., air traffic control). It is still an open question whether the recent developments in HTAP and in NewSQL can be extended for manipulating real-time stream and historical mobility data. Another line of research is how edge computing could be integrated into this context, since it may enable a continuum between these two extremes of historical and real-time data analytics.

5.1.2 Distributed Data Management and Processing for Mobility Data

The paradigms of distributed data management for mobility data include distributed RDBs, NoSQL systems, and data flow processing systems based on e.g., the Spark architecture. It has turned out that these architectures have their pros and cons. For instance,

- while distributed RDBs provide a rich repertoire of features and algorithms, scaling out is hard;
- NoSQL systems are really scalable but they are still immature for handling mobility data (limited queries, basic partitioning / indexing);
- systems following the Spark architecture, though scalable and with a rich repertoire of queries, they provide ad-hoc solutions for indexing and querying.

In the near future, the goal for the mobility data management community obviously includes technologies that will combine the best features of each, in fact, scalability together with rich (though not ad-hoc) indexing support and querying functionality. How can it be reached? Among the three above baselines and since high scalability of RDBs seems questionable, the most promising roadmaps are expected to be those exploiting on NoSQL and/or Spark-based systems. Towards this goal, a number of challenges arise. To name but a few:

- Regarding indexing, (a) how does modern architecture (e.g., SIMD instructions and MSMC architectures) affect the in-memory vs. disk-based processing balance? (b) will the *auto-generated index configurations* paradigm be the future?
- Regarding query processing and the convergence of DBMS and DSMS in a single architecture, which is the golden ratio between offline (archive) and online (stream) processing?
- Regarding analytical processes (from aggregations to forecasting), to what degree they should be considered as duty of the DBMS?
- Regarding the IoT environment, how smoothly can we deal with the edge/fog/cloud settings that appear there?

In either case, the target should be offering high scalability (not only with respect to volume but also to velocity of data) along with query processing methods – and indexing schemes as their background supporters – that will be based on strong foundations (could it e.g., be a multi-model, multi-granularity algebra?).

5.1.3 IoT Challenges as a Driver Use-case for Mobility Data Science

The Internet of Things (IoT) has recently received significant attention. An IoT device may possess an array of sensors that for example monitors the air temperature, carbon monoxide level, wifi signals, and sound intensity. IoT data is initially created on the device, then sent over to a central database system (e.g., the cloud) that organizes and prepares such data for the ongoing use by myriad applications, which include but are not limited to smart home, smart city, the industrial internet, connected cars, and connected health. Data generated by IoT devices is inherently mobile with spatial and temporal attributes. For instance, an audio signal represents the variation of the sound intensity (retrieved by a sound sensor) over the time dimension. Furthermore, IoT devices can be attached to moving objects such as a connected vehicle or a wearable device.

A promising research direction will incorporate IoT data awareness in state-of-the-art mobility data systems. That also requires that system developers design a middleware framework, which understands the IoT devices streaming data to the central data system on one side and the requirements of applications accessing such IoT data on the other side. The proposed middleware system will tune the mobility data system to adaptively decide whether or not to eagerly propagate data from the device, to the edge, or to the central system. Another important direction is to modify existing relational and spatial query processing algorithms to leverage the IoT device capabilities and handle the different rate and types of data generated by various IoT devices. To capture the interconnected nature of IoT data, a mobility data system must also provide a graph processing API in addition to the spatial / spatiotemporal API. Such a combination is already supported by existing graph data systems, however such systems treat the geospatial location attribute as a second class citizen, and hence cannot achieve real-time or near real-time performance. The community needs to craft efficient query operators that accelerate location-aware graph queries and also investigate new index structures that take into account network aspect of linked IoT data as well as the spatial and spatiotemporal aspects.

5.2 Next Generation Systems

5.2.1 Lessons Learned from Existing Systems

Existing systems are built for a variety of application purposes such as on-line updating and querying, and historical data querying and analytics. Next generation systems may be towards a general platform or cloud system into which individual systems can be integrated as modules. Developers who have already built ad-hoc systems could still further enhance their systems and then upload them into the platform.

Existing systems for mobility data have been already well equipped with a number of data models, structures, functions and operators. For example, there are historical data representation methods, different updating policies, a list of R-tree based index structures and grid index structures, and query algorithms. Those well-established techniques should be encapsulated into next generation systems. That is, one should not build the system from scratch.

One aspect receiving moderate attention in existing systems is to evaluate the data quality of raw mobility data and perform data cleaning. This is because application queries and analytics performed on low-quality data inhibit the system performance and even lead to incorrect results. A powerful, general and intelligent tool should be developed in the community to (i) effectively evaluate the quality of mobility data and (ii) efficiently perform the data cleaning task to provide as high-quality data as possible for the system. The tool should be open sourced and sustained and empowered by the community such that a wide range of researchers as well as those from other communities (e.g., GIS) can benefit from that. The system should be completely or almost completely open source rather than partially open source. In particular, the underlying implementation details should be visible to the community after paper publication such that the techniques can be utilized or even enhanced in a wide domain.

In relational database systems and data mining systems, a natural language interface has been provided such that users can express their queries in natural language. This benefits a wide range of users, especially, non-experts who cannot write structured query expressions. Mobility data systems should also provide a user-friendly interface for non-specialists from other domains. The task is to precisely transform queries in natural language into structure and executable languages in the system. An interactive interface is preferred as users may check the correctness and accuracy of transformed expressions.

Data visualization also plays a pivotal role for application users and developers as not only mobility data but also structures for storing and manipulating them should be displayed in a clear way. For application users, the system fluently displays the data in a clear and interactive way such that they understand the meaning and why such results appear. For developers, the system reports the structure statistics in an intuitive and customizable way such that they well understand the internal characteristics of underlying data structures, and then build the structures in an optimal way for application tasks.

5.2.2 Towards an Eco-system for Mobility Data Science

Location data has almost always been supported in data systems as an afterthought. Many systems, e.g., Postgres, Storm, Spark, and Hadoop, have not been originally designed with location data support in mind. What typically happens is that spatial data types get augmented into tuple-oriented systems to support the location data type. For example, a restaurant tuple that describes various attributes of a restaurant is augmented with the latitude and longitude of the location attribute of the restaurant to support location services. Spatial indexes are provided to speedup the access to these attributes, and some accompanying spatial operators are provided to operate on the location attributes to provide location services, e.g., range or k-nearest-neighbor searches. While this approach works, systems that result are not well-optimized for supporting location data. This is similar in analogy to trying to repurpose a car to be able to fly. While this is possible, it will not perform optimally and will never be the same as designing an airplane from scratch in contrast to starting from a car and modifying the car's design. Thus, the quest is to realize a data system where location is treated as a first class citizen from the start, i.e., the system will be optimized firsthand for location data and its corresponding services.

It is important to note that over 80% of the data is associated with location data. Thus, it is important that when we design a system that supports location as a first class data type that this system also supports the other forms of data that are associated with the location. A vision, termed Location+X, calls for treating location data as a first-class citizen in a location data system that at the same time can be extended to support other data types (the "X" in Location+X). The data types "X" can be keywords (e.g., to support spatial-keywords and tweets), can be graphs (e.g., to support road-network data), can be relational data (e.g., to support descriptions of spatial data objects), can be click streams (e.g., to support check-in data), can be document data (e.g., to support points of interest and documents that describe them), can be annotated trajectories (e.g., location + time + textual annotations), etc. Notice that in many location services, more than one data type X may need to be supported at the same time, e.g., a graph data type combined with a document or keyword data types, etc., which calls for a multi-model-like data system.

Following the above vision, this gives rise to an eco-system where location is at the core with some form of an extensible multi-model data system that supports the multitude of data types "X". However, current multi-model data system technology is lacking in several aspects. First, they do not support data streaming that is a cornerstone in location data systems due to the online streamed locations of moving objects. Second and as important, we do not want to fall into the trap of adopting existing multi-model technologies that may affect location being a first class citizen. However, the need for supporting multi-models in one seamlessly integrated location+X system remains a necessity.

In addition to supporting location data via a native location+X engine, an eco-system for mobility data would include many important utilities to facilitate a broad spectrum of location service applications. From the input data side, to help navigate the vast amounts of available location datasets, and discover the right data sets for a given task, a location dataset lake infrastructure and location dataset discovery, cleaning, and integration facilities are needed. From the presentation side, a comprehensive visualization suite is envisioned to support visualizations for combinations of spatial and temporal data analytics on top of location data.

5.2.3 Standards

It is important to have standard interfaces and data exchange formats for the development of an eco-system. The current landscape for mobility data standards has two main players: ISO and OGC (The Open Geospatial Consortium). ISO has one published standard ISO19141 which specifies an abstract data model of a moving geometry consisting of translation and/or rotation of a geographic feature. Based on it, OGC has published multiple standards, of a technical nature, defining data exchange formats and a data access API. Nevertheless the standardization work in mobility data is still in the infancy stage. Compared to spatial data, there already exist more than 80 published ISO standards as part of the ISO191xx suite. This is complemented by more than 160 OGC implementation standards, that are written for a more technical audience, and detail the interface structure between software components.

On the one hand, standards should address the needs of technology users and vendors. Therefore, standardization ideally should involve user communities and vendors. For example, this is important to ensuring that user and vendor needs are addressed; and it can provide assurances to vendors: if they invest in products that comply with standards, the investments are relatively safe. On the other hand, standards should be theoretically sound and should build on the most recent scientific advances. This will benefit all involved, and it calls for the involvement of scientists. To scientists, a key reward for involvement in standardizations is potential real-world impact.

Next, standardization faces a chicken-and-egg dilemma: The existence of high-quality standards can accelerate the exploitation of technology, which is an argument for developing standards early. However, developing standards early runs the risk of the standards being of low quality and hampering the development of compliant products, which in turn slows down exploitation. From this perspective, it may be best to not standardize technology until several products are available.

The "standardize late" route may lead to situations where vendors try to take over standardization with the objective of protecting their investments into their own products. Then standardization processes run the risk of becoming politicized, and scientific arguments (and user needs) are trumped by commercial interests, thus limiting the prospects for achieving science-based standards.

SQL standardization of temporal support provides some insights into potential difficulties. In 1994, a sizable committee of temporal database researchers released that TSQL2 language specification that they hoped could serve as a foundation for standardization. However, the subsequent standardization efforts ran into difficulties. First, it turned out to be very expensive to participate in ISO standardization, which involved meetings across the globe. Second, standardization was dominated by vendors with successful businesses built around SQL-based systems. Third, SQL was, and is, a dinosaur on clay feet: having evolved over many years, it is a complex language that is not the best example of clean language design. Extending such a language is challenging. It turned out that the TSQL2 design did not "scale," and a redesigned version of TSQL2 based on so-called statement modifiers was eventually proposed for standardization. Although the proposal was good science, it did not make it. A key lesson is that standardization is far from an academic exercise where the best proposal wins. Rather, standardization can involve strong political and commercial special interests. Scientists who pursue standardization need to be prepared for such aspects. They should try to understand the costs and reward structure clearly before they embark on standardization. They may consider forming alliances with vendors or user communities, or with other scientists.

Mobility data science, and data science general, has two main setbacks: (1) GDPR and relevant privacy regulations, and (2) data preprocessing is suspected to generate bias in the analysis results. Standards (and community best practices) could help overcome these two setbacks. Data scientists can then base their work on them, and be able to defend it. This creates requirements for developing standards and community best practices for GDPR compliant data management, and for mobility data science pipelines.

5.3 Learning and Analysis

5.3.1 ML over Mobility Data

The rise of machine learning (ML) has impacted plenty of applications in different domains. Mobility data is no exception, as researchers have explored using ML in boosting various related applications and solving systems issues. For example, improving the accuracy of map building is a recurring application that is being explored through ML models. A major hurdle, and a research opportunity as well, is that existing ML and analytics tools, e.g., tensor flow, do not support location and mobility as base data types to reason about. So, even the basic analysis, such as clustering, classification, similarity, etc, need to be explored when mobility data is involved. These tasks, as well as higher-level analysis, can not be totally independent. Instead, common basic building blocks could have an impact on all or some of them. For example, exploring the effectiveness of feature vector components for mobility data analysis is a basic block that could impact different ML-based analysis tasks. This raises a fundamental and big question on what are the analysis primitives and common building blocks for applications that could shape a framework of ML-based mobility data analysis?

With all those open research questions and challenges, there are still more directions that are hardly touched, although being promising and open for ample exploration. One of these promising directions is exploring the ability of ML technology to help improve mobility data preparation and cleaning. This is an increasingly important task with increasing noisy signals in human-generated mobile data and the availability of a wide variety of heterogeneous data sources. It also includes the fact that mobility data is often very sparse.

Another major direction that currently hinders ML models on spatial and mobility data is scalability on big data. Although it is usually assumed that scalability is a step after supporting primitive blocks and supporting data preparation, practically, this assumption causes a significant lag in using newly developed mobility data techniques in real environments for several reasons. First, it leaves making use of non-spatial technologies, such as general-purpose ML models, as an always-future step, which lags the advancements in mobility data behind the non-mobile environments. Second, the lack of scalability makes the developed techniques practically not very useful for evolving applications that operate on big data. This makes the mobility data community lose its audience early in the ever-running game. Thus, thinking of scalability on big data and solving system-level issues for end-user friendliness and usability should be hand-in-hand with exploring all the new techniques, applications, and directions that have been highlighted above.

5.3.2 In-database Analytics – Analytics Algebra

The wide adoption of a tool or system depends to a large extent on the ease of use. One question therefore is where the functionality of mobility data analysis should reside. One option is to make it a stand-alone application with dedicated data types and methods.

Another option is to see it as an extension, or rather a specialization, of a more general purpose system, which would logically be a spatial database or GIS. Since several data types and methods needed for mobility data analysis are already present in GIS, the latter option appears natural.

The most important data types to be supported are movement data in all of its forms: trajectory data, video data, check-in data with ids (like cell towers) or without ids (like induction loops), and more. Other relevant data are contextual, which help to understand or explain mobility or the behavior that underlies it. These data can be (topographic) map data, weather data, elevation data, and data on events in the real world, like a big sports match taking place at a certain location at a certain time.

Since data integration is key to nearly all mobility data analysis tasks, a well-designed system must include this functionality for heterogeneous data. One option is to provide an algebra that can combine different data types, but this is convenient mostly for field data in raster form, and less so for object data in vector form. Since GPS tracks (trajectories) are typically in vector form, further research is needed to see in what ways algebras can be utilized to deal with heterogeneous mobility data.

Another big issue in data management and analysis is the level of detail that must be maintained to support the analysis. Here the situation is the same as for any spatiotemporal research domain: a high level of detail in space and time leads to huge data sets and slow analysis, but a low level of detail may not suffice to answer the research questions. The presence of heterogeneous data implies that a system needs to deal with data with very different spatial and/or temporal resolutions.

5.3.3 Visual Analytics

Visualization and exploratory analysis of mobility data has long been a hot topic in visual analytics. More recently, the trend turned to combining visualization with modeling and simulation to support decision making. This kind of research is by necessity application-oriented, while much less is done on developing more general ideas and approaches.

One general research problem that has only been slightly touched in VA but not systematically addressed is human involvement in real-time analysis of big mobility data. Is it possible to define realistic scenarios for involving human intelligence in big data analytics taking into account the cognitive limitations of human analysts with regard to the amount of information that can be perceived, speed of processing, and time required for analytical reasoning and contributing to the analysis process? It appears possible, in principle, that human experts work at their own pace on gradual improvement of a computational model operating automatically in real-time, provided that the context and the properties of the incoming data do not change too rapidly.

There is even a more general problem of finding effective approaches to combining computational methods of analysis, such as ML, with human expert knowledge and reasoning. While it is usual in VA to involve algorithmic methods in analytical workflows, it is mostly limited to two scenarios. One is the use of algorithmic methods to support human reasoning and generation of new knowledge in the mind of an analyst. The other is human-controlled development of ML models according to the standard data science pipeline. The involvement of human intelligence is limited to thoughtful data preparation, feature selection, parameter setting, and so on. It would be great to find ways to make more direct and effective use of human-possessed concepts and, particularly, knowledge of causal relationships. Possibilities for that could be explored in two directions: (1) transforming purely data-driven ML methods into hybrid approaches that can use human expert knowledge in the learning process and (2) enriching ML outcomes with expert knowledge, so that data-level features are lifted to domain concepts and statistical associations transformed to causal links.

While these research problems and directions sound quite general, there is a specific topic for research on mobility data analysis: from low-level movement data, such as trajectories of moving entities, derive understanding and models of mobility behaviors. This is where the human capability of abstracting, forming general concepts, and reasoning with the use of these concepts are especially valuable, as current algorithmic methods are limited to operating on the data level.

Hence, a grand research challenge for visual mobility analytics is to develop approaches to understanding and modeling mobility behaviors, i.e., how purpose-oriented movements and actions are being adapted to the contexts in which they take place.

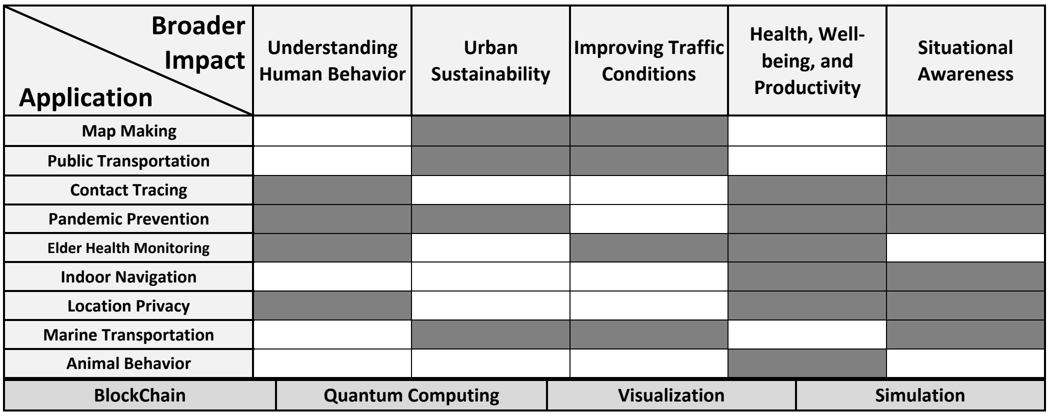
The research should include development of a conceptual framework for describing individual and collective behaviors of moving entities. It should encompass various kinds of collective behaviors, from individuals pursuing their own goals while adapting to behaviors of others to groups of cooperating entities having a common goal and, possibly, competing with other groups. The framework should define the set of potentially relevant aspects of a mobility behavior, such as visited places and characteristics of the visits (frequency, duration, regularity, typical times), dynamic properties of movements (speed, direction, mode), as well as affecting factors, including characteristics of the moving entities themselves and the context of their movement.

The next sub-problem to be solved is to find ways of transforming low level mobility data, i.e., trajectories of moving entities, into representations of mobility behaviors. It is the process requiring involvement of human abstractive perception and interpretation of the data in terms of high-level concepts. The human needs to understand how to transform the data and to tell this to the computer. Interactive visual interfaces should support the human both in gaining the understanding and in communicating necessary knowledge and instructions to the computer. In parallel, it is necessary to develop computer algorithms capable of incorporating human knowledge and adhering to the instructions.

The following research problem is how to analyze behaviors after they have been extracted from elementary movement data and represented by appropriate data structures. The conceptual framework should enable defining the types of conceivable patterns of movement behavior. This will provide orientation for developing visualization techniques facilitating visual discovery of behavioral patterns, as well as algorithmic methods for detection of specified types of patterns. These techniques and methods should be incorporated in systems and workflows for analyzing the contexts in which various patterns take place and developing models for describing and predicting mobility behaviors depending on the context.

6 Applications of Mobility Data Science

This section describes the broad impacts of mobility data science as well as specific applications that were identified by participants. Figure 4 provides an overview of this section. We first give an overview of broader impacts of mobility data science in Section 6.1 followed by specific applications described in Section 6.2. In addition, the working group also discussed algorithmic paradigm that may be applied to Mobility Data Science as described in Section 6.3.



Broader Impact Application	Understanding Human Behavior	Urban Sustainability	Improving Traffic Conditions	Health, Well-being, and Productivity	Situational Awareness
Map Making		■	■		■
Public Transportation		■	■		■
Contact Tracing	■			■	■
Pandemic Prevention	■	■		■	■
Elder Health Monitoring	■		■	■	
Indoor Navigation				■	■
Location Privacy	■			■	■
Marine Transportation		■	■		■
Animal Behavior				■	

BlockChain	Quantum Computing	Visualization	Simulation

Figure 4 Broader impacts (vertical), example applications (horizontal), and underlying algorithmic paradigms (bottom) of mobility data science.

6.1 Broader Impacts

This section describes broad areas that may benefit from a mobility data science research agenda. Specific applications which may fall under one or multiple of these areas are described in Section 6.2

6.1.1 Understanding Human Behavior

During the seminar, one participant made a controversial statement claiming that "many researchers focus on the problem of location prediction, yet no useful application for this problem exists". Indeed, it seems difficult to leverage knowledge such as "User X will visit Coffee Shop A next" or "32 users will visit Coffee Shop A today" for marketing or other applications. However, if we can begin to understand the underlying behavior, at the individual-, group-, or population-scale, that leads to such predictions, we could begin to understand *why* one coffee shop chain is increasing visitor rates (for example due to a movement towards organic coffee sold by the former coffee shop). Through inferring from the data about such behaviours, only then we can take corresponding actions not only to predict locations, but also to prescribe actions (such as offering more organic coffee) to improve visitor rates. Such understanding of (human) behavior will broadly affect applications using mobility data: While traditional spatiotemporal data science allows predictive analytics to predict the future, mobility data science enables prescriptive analytics by understanding the underlying human behavior to devise actions and policies that change the future in a desirable way.

6.1.2 Urban Sustainability

Rising temperatures on Earth, have led to the exploration of new technologies to help curb the severe effects of Climate Change. Highly urbanized environments are a focal point for the application of these new technologies as they introduce a variety of mobility modalities (e.g., EVs, bicycle, scooters with respective sharing programs) and smart spaces (e.g., smart IoT living spaces with renewable energy), where human spend 90% of their time. According to the European Environmental Agency, urban environmental sustainability encourages revitalization and transition of urban areas and cities to improve livability, promote innovation and reduce environmental impacts while maximizing economic and social co-benefits. In the epicenter of these challenges lay challenging spatiotemporal data-driven problems that require new operators, algorithms and infrastructures to schedule energy prosumer's (consumers-producers) activity to minimize the CO2 footprint while improving livability. Given that the computing field is anticipated to increase its CO2 footprint from 4% to 8% until 2025, compared to only 2% of the whole aviation industry, clearly demonstrate that not only computing requires to become more energy efficient, but also that the computing field has to compensate for this increase by reducing CO2 in a variety of other domains (i.e., smart-everything: transport, heating, industry, agriculture/land.)

By understanding how people move in cities, outer suburban, and regional areas, the demand for infrastructure and energy can be better understood. This, will not only help to ensure sustainable production and consumption, but also to reduce urban inequalities in cities. In addition, tools that use accurate models of people mobility to predict future transportation demands would be very useful to stakeholders to plan future developments, formulate evaluate alternative policies for improving city infrastructures, end evaluate the impact of various decisions.

6.1.3 Improving Traffic Conditions

Traffic is a problem of global scale, as recognized by transportation science over a decade ago. Drivers in the United States spend 6.9 billion of man-hours stuck in traffic and waste more than 11 billion liters of fuel per year according to INRIX. Measured per-capita, people in Russia and Thailand spend even more time in traffic, while Brazil, South Africa, the UK, and Germany are only slightly behind the United States. The SIGSPATIAL community had great success in exploiting mobility data for predicting traffic and using these predictions for traffic-aware routing. Leveraging mobility data science and understanding the underlying behavior of human participants concomitantly with different transportation modes, can enable more effective solutions to multiple problems at the heart of improving traffic management. One such example is devising accurate models for dynamic scheduling of public transportation. Another example is the context-aware optimization of traffic signals regime – i.e., incorporating the impact of additional flux of pedestrians in bus/train stations, to minimize the stop-and-go impacts for vehicles.

6.1.4 Health, Well-being, and Productivity

Mobility data has been found increasingly useful in novel applications in the social care domain. For example, contact-tracing solutions have been using human mobility data during the COVID-19 pandemic to notify individuals who may have been in close contact with an infectious individual. As another example, GPS-enabled smart-watch can be used to monitor the movement of elder users. In addition, given the increasingly blurred boundaries between work and private lives, especially since the disruptions led by the pandemic, understanding mobility patterns in conjunction with daily works can provide insights into work-life balance. Given that humans are creatures of habit, disruptions to the daily habits in work and private lives can also led to disrupted productivity. Mobility data can provide a deeper insight to assist users in maintaining health, wellbeing, and productivity, when incorporated to digital assistance.

6.1.5 Situational Awareness

Situational awareness, initially a term coined in defence applications, involve *perception* of the environmental states using the surrounding data, *comprehension* of the ingested data to understand the emerging situations, and *projection* of future states and/or events, which require predictive analytics. Mobility data provides critical components and insights into situational awareness in cities. When achieved, this is applicable not only to enabling robust critical infrastructures in cities, but also to protecting them from harm, such as forest fires, earthquakes, terrorist attacks, etc. Many researchers have used mobility data as an input to enable situational awareness in cities.

6.2 Specific Application Fields

This section described specific applications of mobility data science identified during the Dagstuhl Seminar.

6.2.1 Map Making

The spatiotemporal data science community has proposed many successful solutions to infer maps from trajectory data. Such data is important to identify changes in a road network which is paramount for autonomous driving. In addition to considering trajectories, understanding the underlying purpose of human mobility will allow to create maps tailored to different purposes. For example, by understanding which trajectories correspond to commuters, tourists, and joggers, we can devise better maps for such groups, rather than having a single map that is oblivious to the purpose of a trip and the underlying human behavior. It is important to note that accurate map making is an important enabler in all autonomous car applications, and at the same time data from automonous and semi-autonomous car operation (which includes GPS, video and other sensor readings) is a useful source for improving automatically generated maps.

6.2.2 Public Transportation and Traffic Management

Public transportation and mobility-as-a-service, particularly in urban areas, is a key application area for mobility data science. This also includes the recent ridesharing services. People need to move from A to B daily, either for work, leisure, or caring purpose. The means to achieve that vary, either a single modal trip with a private vehicle, or a multimodal trip involving driving a private car, parking, a subway ride, and a final Uber ride or an e-scooter hire. The latter, when computed as an integrated mobility and offered as a service, is called mobility-as-a-service. The challenges are broad, from understanding demands and volumes of traffic, rides, or shared bikes pick-ups, to allocating resources (vehicles as well as drivers) to regions with high demand. In peak travel times, and high holiday seasons, efficient and effective transportation and mobility services, and other application of smart cities (such as smart parking) are critical to enable sustainable operations in cities.

6.2.3 Contact Tracing

Contact Tracing refers to the process of tracking persons who may have come into spatial contact with an infected person, and subsequently collecting further information about these contacts. The feature-rich interaction, processing and localization/communication modalities of smartphone devices, have brought these to battle on the technological forefront and have curbed the fast spread of pandemics, like COVID-19. The community has to this date proposed a wide range of approaches, ranging from: opportunistic to participatory approaches, privacy-sensitive to no-privacy approaches, handheld-based (distributed) to cloud-based (centralized) approaches, proximity-based (e.g., BLE, sound) to location-based approaches (e.g., Wi-Fi, GPS), for only outdoor settings to indoor settings, using closed-source to open-source counterparts. However, a wide range of challenges remain unanswered, including methodologies to improve the penetration and adoption rates, alleviate privacy or expectation skepticism, ubiquitous availability on low-end terminals as well as technological/psychological adoption barriers, achieving cross-country interoperability with standard formations beyond recommendations, scalability/reliability and accuracy verification of engaged spatial technologies as well as lessons about effectiveness from real large-scale deployments.

6.2.4 Pandemic Prevention

Leveraging (human or animal) mobility data to understand the underlying behavior will allow us to gain an understanding of how a new or re-emerging disease spreads across space and time, provide accurate predictive models and devise actionable interventions and policies to prevent future epidemics and pandemics. Such understanding will require close collaboration with sociologists who are able to explain human behavior (such as social distancing behavior or vaccine uptake), epidemiologists to understand the ecology of infectious diseases, and experts in policy to help devise policies to mitigate diseases spread and to effectively communicate such policies to the public.

6.2.5 Elder Health Monitoring

GPS-enabled smart-watch technology can be used to monitor the movement of elder users. The trajectory data recorded by this smart-watch can reflect the mental health state of the user. In particular, if the user is showing early signs of dementia, her/his trajectories could show an abrupt change from her/his movement history. For instance, a user, who normally walks in a park then going to a restaurant, is found to only stay in the park for a substantial amount of time. Indoor sensors installed in the room can also be used to track whether an elder or a patient falls from the bed. Trajectory outlier analysis methods, together with gerontology knowledge, can be very useful for this kind of applications.

6.2.6 Indoor Localization/Navigation

Indoor localization is still an open research problem due to the non-existence of the indoor equivalent of GPS: a system that can provide the user location in any building worldwide. This is particularly important in many applications such as pandemic tracing, E911, indoor analytics, among others. There are a number of systems developed over the years to address this problem based on different data sources including WiFi signals strength and time of arrival, cellular signal, UWB, ultrasonic, magnetic tracking, inertial sensors, among others but also a new wave of infrastructure-free localization method using deep learning and computer vision.

6.2.7 Location Privacy

The incorporation of localization technologies creates new challenges with privacy, as localization processes might allow the service provider to know the location of a user at all times. Location tracking is unethical in many respects and can even be illegal if it is carried out without the explicit consent of a user. It can reveal the stores and products of interest in a mall we've visited, doctors we saw at a hospital, book shelves of interest in a library, artifacts observed in a museum and generally anything else that might publicize our preferences, beliefs and habits. Privacy-preserving localization has been an intensive subject of research in the past, yet the advent of new localization, tracking and contact tracing technologies brings new challenges to the topic that need to be investigated by the mobility data science commununity.

6.2.8 Marine Transportation

International shipping is $\sim$ 1000 million tonnes CO_2 one of the main emitters of Greenhouse Gases (GHG). This is equivalent to $\sim$ 3% of the global emissions worldwide, in the order of the total emissions of Germany. An optimization of ship routes could effectively lead to significant reductions of GHG emissions and contribute to the actions against anthropogenic global warming. The influence of ocean currents, waves and wind on the course and speed to ships are known for centuries. Mean currents, usually displayed in sea charts, can be used, for example to utilize the Gulf Stream on the way from North America to Europe and avoid it on the way back. However, ocean currents are highly variable and change their strength and directions on timescales of days to weeks. Used optimally, ocean currents lead to more efficient paths between two given ports. Though the spatial and spatiotemporal database and mobility community well studied how to optimize routes on street traffic networks taking many influencing parameters, like traffic jams, route preferences, etc. into consideration, problems of open-water routing for ships hasn't been touched by the community, yet. This application field opens up new interesting fields of research in the community, including studies on ocean current, wave and wind prediction, free-space routing, routing in highly dynamic environments, etc.

6.2.9 Animal Behavior

It has been long recognized that the motion trends of different wildlife species have a significant impact on ecological processes, spanning from the dynamics of biodiversity, through transmission of diseases, to feedback-like changes of the very animals behavior in terms of relocating breeding locations. Advances in sensing devices and communication have increased the observational capabilities – however, to date, there is no systematic exploitation of such data in a manner that would enable derivation of models for predicting the animal behavior and movement ecology broadly. In addition to the landscape type of interactions with terrestrial animals (e.g., pasture vs. forestation; impact of urbanization) that could benefit from mobility data science, there is significant potential of its application in the domain of aquatic animals where certain species (e.g., salmon) are well known to travel large distances between their adult habitat and the nesting one.

6.3 Algorithmic Paradigms

In addition to discussing how spatial computing can be applied to other fields, the Applications Working Group also discussed the flipside of which novel algorithmic paradigms may find application in spatial computing. This section summarizes these findings.

6.3.1 Blockchain Computing

A key component in future Web 3.0 scenarios using edge learning is the requirement to utilize a shared database that allows all participants to operate collaboratively with more functionality and transparency. The objective is to enable users execute updates and queries on the collaborative edge database while preserving a consistent view among all users maintaining the system consistency and transparency. Blockchain database architectures keep records on an immutable chain of blocks, so later on, nodes agree on the shared state across a network of untrusted participants. Thus, it forms the blockchain platform that can be viewed as a distributed (transaction-log or) database system. One basic obstacle in these are performance issues measured in terms of throughput and latency, because of the lengthy consensus protocols. As such, the Mobile Data Science community needs to identify new algorithms, structures and protocols to make blockchain databases for mobility data science practical and efficient.

6.3.2 Quantum Computing

Quantum computing provides a new way of designing algorithms that provides a number of advantages over their classical counterparts. For example, quantum parallelism can allow exponential gains in speedup and storage space compared to classical algorithms. Moreover, the no-cloning theory, which states that quantum bits (qubits) cannot be copied, gives an advantage to quantum algorithm in different security applications. This in turn opens the door for addressing different mobility data science problems in terms of enhancing the speed of the current algorithms as well as introducing new algorithms that benefit from the potential of quantum computing concepts. For instance, quantum computing allows to solve np-hard problems efficiently, which affects many spatial computing problems that can be reduced from the np-hard traveling salesman problem.

6.3.3 Visualization of Mobility Data

As discussed and defined in Section 3, mobility data science considers the human behavior that underlies the observed data. It requires expertise from behavioral and social sciences not only to understand this behavior, but also to leverage this understanding for informed decision and policy making. To make mobility data science results actionable for policy makers, such as local health departments, an important paradigm is effective visualization of mobility data to inform domain experts that may not be computer scientists. New solutions for mobility data visualization are required to go beyond understanding of spatiotemporal patterns towards understanding of the underlying behavioral patterns that drive them.

6.3.4 Modeling and Simulation

Understanding the human behavior that causes observed spatiotemporal data allows us to abstract behavior into models of human behavior that can be captured by simulation models. For example, the decision process of individuals to visit a certain point of interest like a coffee shop can be modeled depending on age and gender of individual agents in an agent-based model. Informed by observed human mobility data, such an (agent-based) model can than be used to simulate entire cities having millions of individuals. Such simulation can then be used to predict future “what-if”-scenarios, for example to predict the impact of opening a new coffee shop or of changing the menu to attract certain population groups, or to simulate the spread of an infectious disease across a city. Having the ability to run many such simulation, we can investigate optimal actions and policies achieve a desired future, for example to maximize the number of customers in a coffee shop, or to minimize the spread of an infectious disease.

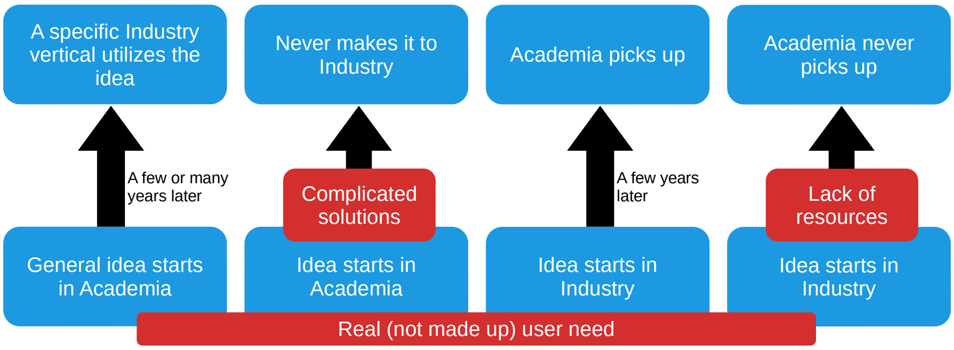


Figure 5 Industry/Academia Alignment.

7 Mobility Data Science and Industry

There is a consensus that we want to make our work practical, and the connection with industry is a good motivator toward this goal. But how to create such connections? Talking to engineers and product groups is more effective in establishing links than talking to industrial R&D groups. Engineers and product groups own the data and the products, and they have the problems. Visiting the company and giving presentations and demonstrations is a persuasive way to develop contacts. Industry is generally interested in getting exposed to good students for hiring. Thus, inviting industry to sponsor student activities such as programming contests and projects is another way to create links. Commercial companies also fund research directly, sometimes in the form of collaborations. Industry can also be found in 3rd party communities with common interest for research, like OGC, thematic meetups, hackathons, and tournaments.

University researchers can provide teaching and training opportunities for people in industry, which helps bridge the gap. Other than industry, research can generate impact by involving the future beneficiaries and users. Industry, on the other hand, could launch projects with a *lab culture*, where they prepare a sandbox where researchers can work and put their algorithms to have impact on an industry problem.

The efforts in academia and industry are not always aligned. Figure 5 outlines the different cases. The first case is where an idea starts in academia and gets adopted in industry. Good examples for this case are navigation, map matching, route planning, and map making. The three top problems with Mobile data that Microsoft is trying to solve are: (1) how to spend less on road map data, (2) how to identify where people visit, and (3) creating enterprise solutions for fleet management.

Spatial indexes have also proved to be very useful for practical applications, and they came from academia. However, after the very first spatial index proposals for R-tree and Quad-tree, very many indexing papers appeared proposing incremental changes. Improving a spatial index to perform 2% better is not an interesting problem to industry. It comes at a high cost of development, testing and deployment, which is not worth the time from the company's point of view. Integrating such solutions depends on the ROI of the proposed improvement. If this enables new products and applications or saves costs, it will be integrated – assuming it is maintainable.

Many inventions are getting traction when they are available for a broader community via open source. The speed may also result from industry actions like Amazon Scholar, where researchers can become affiliated to Amazon and get full employee access to data and resources to implement their ideas. Another way is to launch academic startups and take the ideas directly to industry.

The second case in Figure 5 is when an idea that starts in academia never makes it to industry. A main reason is that the problem is made up, and does not map to real user need. Location prediction has been around for a while, but never got into real use. Another example is the problem of understanding what people are doing from their mobility. A third example is COVID contact tracing. It became clear early that there is not enough data to make COVID tracking applications work. Nevertheless many papers continued to appear with COVID tracking solutions. Another reason for not reaching the industry is that the solution might be too complex to implement. If simpler solutions exist, they tend to be more adopted, being easy to integrate with the whole eco-system that the industry has. Some complex solutions found their way to commercial tools though, backed by strong customer need, e.g., RSA encryption.

Researchers might be in a position to think that the feedback from industry limits visionary thoughts. The problems of industry are no less interesting than the problems of open-ended research. Industry operate on constraints for price, for time, for infrastructure, and so forth. The ability of research to recognize as many of these constraints as possible, makes it effective.

The fourth case, and possibly the most recurring, is where a practical mobility problem is not picked up by academia. This is mostly because of lack of data. Data is a key for research in mobility, and the lack of data is a major obstacle. To strengthen the data sharing between industry and academia, both sides need to converge toward the ambitions of each other. The business ROI is different from research ROI. Researchers want publications, citations, keynotes, and research funds while industry wants profits, IP, new markets, and new products. Researchers tend to ignore practical/technical difficulties, such as the deployment of the solution. Something simple like the availability of a docker image can push forward an innovation to get used. This can also be addressed by starting further research/innovation projects in 3rd party founded programs. A connection of interesting and valuable research topics with market requests and opportunities can trigger a win-win-win situation, where researchers can publish results, companies can integrate the findings in their products and users do get improved solutions. Initiatives such as the EU's FAIR principle can ensure real data becomes available. However, the EU's GDPR rights must be fulfilled to ensure the privacy of individuals.

Synthetic data is still highly useful, e.g, to test the scalability of solutions. Such datasets should have similar properties with existing publicly available datasets, and will also foster the development and testing of new research by academia. To build such benchmarks we will need to leverage new optimization techniques and mobility models.

In early 2000s, spatial data used to be addressed as a special kind of data in database engines. This proved to be counter productive, as users rather prefer to have native support for their data in the database that they use. Applying this experience to the domain of mobility data, it has to be thought of from the beginning as a native data type.

8 Towards a Mobility Data Science Curriculum

When we initiate a curriculum, and consequently a degree, we have implicit commitment to the students that join this degree about the market opportunities. In the case of mobility data science, clear employers will be cities, transport authorities, and consultation companies. The current state in market is that there are CS graduates who approach every problem by code, and often try to re-invent the wheel. On the other hand there are geographers, planners, and architects who are often not able to code. This gap is currently closing by discrete initiatives for programs that link the two. With this argument, it is time to harmonize these initiatives and start collaborating on building curricula for mobility data science. Yet we need to take steps to evolve interest, and hence to evolve sufficient mass of students. Naturally this has to go bottom up, starting with a single course, then a specialization, then a study program, depending on the student enrollment.

A text book on mobility data science is missing. The seminar participants collaborated in creating a list of topics to be included in a curriculum.

- Basic principles of mobility data science
- Intro to time-series, ST and trajectory, spatial data handling
- Mobility data modeling, e.g., multi-modal traffic modeling
- Mobility databases and query processing
- Modelling, i.e., transform a real-world problem into a computational problem that is meaningful.
- Search and Optimisation, e.g., routing
- Movement data processing, and data warehousing
- Specific errors and problems in mobility data (data quality and data cleaning)
- Mobility data visualization
- ML techniques for mobility data
- Privacy of mobility data, GDPR, and data minimization
- Ethical, responsible / FATE in AI for mobility
- Hands-on mobility data science projects and case studies
- Big data mobility processing

A practical approach to developing a book is to start by a mobility data science problem, then develop the needed theory and techniques towards solving this problem. Imagine a company that wants to develop an application for eco-routing starting with CAN bus data. The process of developing the curriculum would then develop around the needed steps. The company will need to first collect the data, then clean, transform, correlate with weather data, etc. The curriculum shall then be developed in a way that teaches students the theory and practice of these tasks.

Building a complete program on mobility data science for CS students might be too specific at the moment. A pragmatic approach is to incorporate it as a special track in existing data science programs (1-2 semesters). It can also be offered as a specialized diploma/certification program. This argument is supported by the big overlap in knowledge, skills, and questions between data science and mobility data science. Both share the foundations of statistics, graph theory, linear algebra, measures and metrics, etc. They also share the techniques like clustering, classification, learning, heterogeneity and modeling. As such mobility data science will be more of an appealing flavor of data science, mainly to attract students, while teaching the same thing with a twist.

Yet if we expand the scope, which seems more proper, a mobility data scientist needs to acquire other knowledge and skills than CS. Domains like urban and transportation have went a long way in building curricula. A curriculum in mobility data science should not ignore and redo. Mobility is not entirely a CS topic. It connects to people mobility, and students must be trained on the human and policy aspects. A curriculum should cover at least the additional topics of: geography, urban/transport planning, and ethics. Students also need to learn the pitfalls of real-world data, which is a topic that is often missed in CS programs. They need to learn and appreciate the state of art in mobility and transport modeling. Finally they need to learn how to produce value with minimal data requirements, and how to deal with the human part of the problem. Building such a curriculum cannot be solely covered by CS departments. It has to be conceived in an inter-faculty structure. A mobility data science curriculum should not adopt on a toolbox view. It should rather train students for analyzing and explaining the results.

Promoting a disciplined approach to mobility data science goes through open science and open software. A basic toolset, developed through proper scientific software development, that every can use helps aligning the state of practice and the sharing of experience among practitioners. Successful examples in the spatial domain are the geopandas, and geos libraries. It also helps to promote this discipline in scientific and in practitioner events, and to possibly evolve communication platforms and events.

9 Closing Notes

The vivid discussions during the seminar demonstrated a consensus that it is timely to identify mobility data science as a multidisciplinary research topic, and to start building a community. This report presented a clear definition of the scope of work, and the key research problems for this community. Many of the participants indicated the necessity to have follow up events, e.g., another Dagstuhl Seminar, and community building tools, e.g., mailing list. Specially that the majority of participants in this seminar were remote, a physical gathering should be planned as soon as the pandemic situation allows.

The majority of participants in this seminar stem from CS domains. In follow-up events, it will be needed to increase the participation of other domains such as social and behavioral sciences and epidemiology whose expertise is paramount to understand the human behavior that causes mobility data.

In the time of COVID-19, organizing a scientific gathering is challenged by the uncertainty of physical participation. The organizer would like to thank the Dagstuhl team, which allowed the switching from purely onsite into hybrid only a few weeks before the seminar due to the Omicron variant of COVID-19. This flexibility, which is not matched by hotels and conference centers, allowed the organizers to focus on the program and the participation, rather than focusing on the venue logistics. This Dagstuhl Seminar allowed the field of mobility data science to take a great step forward. All participants greatly appreciate the unique opportunity that Dagstuhl provides for supporting the science.

Seminar Hybrid-Photo

Participants

- Taylor Anderson
George Mason Univ. –
Fairfax, US
- Amr Magdy
University of California –
Riverside, US
- Mahmoud Sakr
ULB – Brussels, BE
- Flora Salim
University of New South Wales –
Sydney, AU
- Maxime Schoemans
ULB – Brussels, BE
- Bettina Speckmann
TU Eindhoven, NL
- Marc van Kreveld
Utrecht University, NL
- Andreas Züfle
George Mason Univ. –
Fairfax, US

Remote Participants

- Jussara Almeida
Federal University of Minas
Gerais-Belo Horizonte, BR
- Gennady Andrienko
Fraunhofer IAIS –
Sankt Augustin, DE
- Natalia V. Andrienko
Fraunhofer IAIS –
Sankt Augustin, DE
- Walid Aref
Purdue University –
West Lafayette, US
- Eric Auquiere
STIB / MIVB – Brussels, BE
- Yang Cao
Kyoto University, JP
- Sanjay Chawla
QCRI – Doha, QA
- Reynold Cheng
University of Hong Kong, HK
- Panos Kypros Chrysanthis
University of Pittsburgh, US
- Xiqi Fei
George Mason Univ. –
Fairfax, US
- Gabriel Ghinita
University of Massachusetts –
Boston, US
- Anita Graser
AIT – Austrian Institute of
Technology – Wien, AT
- Dimitrios Gunopulos
University of Athens, GR
- Joon-Seok Kim
Pacific Northwest National Lab. –
Richland, US
- Kyoung-Sook Kim
AIST – Tokyo Waterfront, JP
- Peer Kröger
Universität Kiel, DE
- John Krumm
Microsoft Corporation –
Redmond, US
- Johannes Lauer
HERE – Schwalbach am Taunus,
DE
- Mohamed Mokbel
University of Minnesota –
Minneapolis, US
- Mario A. Nascimento
University of Alberta –
Edmonton, CA
- Siva Ravada
Oracle Corp. – Nashua, US
- Matthias Renz
Universität Kiel, DE
- Dimitris Sacharidis
ULB – Brussels, BE
- Mohamed Sarwat
Arizona State University –
Tempe, US
- Cyrus Shahabi
USC – Los Angeles, US
- Egemen Tanin
The University of Melbourne, AU

- Yannis Theodoridis
University of Piraeus, GR
- Kristian Torp
Aalborg University, DK
- Carola Wenk
Tulane University –
New Orleans, US
- Martin Werner
TU München – Ottobrunn, DE
- Song Wu
ULB – Brussels, BE
- Li Xiong
Emory University – Atlanta, US
- Jianqiu Xu
Nanjing University of Aeronautics and Astronautics, CN
- Moustafa Youssef
Alexandria University, EG
- Demetris Zeinalipour
University of Cyprus –
Nicosia, CY
- Esteban Zimanyi
ULB – Brussels, BE
- Dimitris Zissis
University of the Aegean –
Ermoupolis, GR

Report from Dagstuhl Seminar 22022

Mobility Data Mining: from Technical to Ethical

Bettina Berendt*[1], Stan Matwin*[2], Chiara Renso*[3], Fran Meissner[4], Francesca Pratesi[5], Alessandra Raffaetà[6], and Geoffrey Rockwell[7]

1 **TU Berlin, Germany & Weizenbaum Institute – Berlin, Germany & KU Leuven, Belgium.** berendt@tu-berlin.de
2 **Dalhousie University – Halifax, Canada.** stan@dal.ca
3 **Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy – Pisa, Italy.** chiara.renso@isti.cnr.it
4 **University of Twente, The Netherlands.** f.meissner@utwente.nl
5 **Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy – Pisa, Italy.** francesca.pratesi@isti.cnr.it
6 **University Cá Foscari of Venice, Italy.** raffaeta@unive.it
7 **University of Alberta – Edmonton, Canada.** grockwel@ualberta.ca

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22022 "Mobility Data Analysis: from Technical to Ethical" that took place fully remote and hosted by Schloss Dagstuhl from 10–12 January 2022. An interdisciplinary team of 23 researchers from Europe, the Americas and Asia in the fields of computer science, ethics and mobility analysis discussed interactions between their topics and fields to bridge the gap between the more technical aspects to the ethics with the objective of laying the foundations of a new Mobility Data Ethics research field.

Seminar January 10–12, 2022 – http://www.dagstuhl.de/22022

2012 ACM Subject Classification Computing methodologies → Machine learning; Computing methodologies → Artificial intelligence; Security and privacy → Human and societal aspects of security and privacy; Human-centered computing → Interaction design; Social and professional topics → Computing / technology policy; Security and privacy → Database and storage security; Security and privacy → Software and application security

Keywords and phrases Dagstuhl Report, Mobility Data Mining: from Technical to Ethical

Digital Object Identifier 10.4230/DagRep.12.1.35

1 Executive Summary

Bettina Berendt
Stan Matwin
Chiara Renso
Fran Meissner
Francesca Pratesi
Alessandra Raffaetà
Geoffrey Rockwell

Mobility data is one of the fastest growing types of data, thanks to the increasing number of mobile devices approaching the population of the globe. The collection, storage and analysis of spatio-temporal data representing trajectories of moving objects is one of the topics that

* Editor / Organizer

Mobility Data Analysis: from Technical to Ethical, *Dagstuhl Reports*, Vol. 12, Issue 1, pp. 35–66
Editors: Bettina Berendt, Stan Matwin, Chiara Renso, Fran Meissner, Francesca Pratesi, Alessandra Raffaetà, and Geoffrey Rockwell
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

received major attention in the field of data analytics. The more semantic information is collected from various sources, the richer is movement data. This enriched mobility data is typically referred to as semantic trajectories. The analysis of such trajectories can produce powerful results in domains such as transportation, security, tourism, health, environment and even policy design. The recent COVID-19 outbreak shed a light on the importance of collecting mobility data for public health. However, at the same time, the more mobility data is enriched with semantics, the larger the risks of violating the privacy of users and of possible unethical uses of these data analysis results. Aspects of Computational Ethics include privacy, but they go beyond this, towards a more general vision of ethical gathering, processing, uses of data and the results of data analyses. How ethics interrelates with mobility data analysis is an emerging issue.

The objective of this Dagstuhl Seminar was therefore to start a deep interacting discussion between Mobility Data Analysis researchers and Ethics experts to link these two fields with the objective of creating the foundations of a new Mobility Data Ethics research field.

This Dagstuhl Seminar, organised by Chiara Renso, Bettina Berendt and Stan Matwin as an activity from and beyond the MASTER project [1], aimed at bringing together researchers from different disciplines from Computer Science, Mobility Analysis and Ethics to trace the path from a technical vision of mobility Analysis to an also ethics-based approach to the field.

The three-day seminar was structured into three main modules: (1) round-table presentations in which each participant presented him/her self with a question about Mobility and Ethics that represents his/her interest and an object to visualise this interest or serve as a starting point for further discussion; (2) three tutorial on “technical”, “ethical” and “legal” aspects of mobility data; (3) the working groups to discuss the main topics of interest that emerged during phases (1) and (2).

As a result of the group discussions on participants’ interests and the issues raised in the tutorials, we formed five main working groups:

- What is/are the trade-off(s) between data privacy and data utility?
- Mobility Data Anonymity (Can location data be really anonymous?)
- Ethics of Mobility Data: What is unique? Which guidelines?
- Mobility Data Analysis Ethics beyond the data
- Mobility Data Analysis Ethics beyond humans only: Tracking animals and moral agency

The tutorials and each of the working groups are described in a chapter of this report. Like other Dagstuhl Seminar reports, these chapters aim at makign the scientific results re-usable and extendable by others. In addition, we also want to help others profit from our experiences with the videoconferencing and other media technologies that we employed and the interaction-design choices that we made. This last chapter is a reflection also on ethical aspects of the precluded and the newly added forms of mobility of scientists (and others) in meetings during and after COVID-19.

References

1 Chiara Renso, Vania Bogorny, Konstantinos Tserpes, Stan Matwin, and José Antônio Fernandes de Macêdo. Multiple Aspect Analysis of semantic trajectories (MASTER). *Int. J. Geogr. Inf. Sci.*, 35(4):763–766, 2021.

2 Table of Contents

Executive Summary

Tutorials

Topic-based Working Groups

Online Interactions in a COVID-19-era Dagstuhl Seminar: Design, Experiences, and Reflections

3 Tutorials

3.1 Location privacy: an overview

Sébastien Gambs (Université du Québec à Montréal (UQAM), Canada)

In the introduction of his tutorial, Sébastien highlighted the fact that location is personal data that is collected and used in many different contexts such as location-based services, geolocated advertising, augmented reality, mobile game, collaborative traffic monitoring, call details records, physical analytics, smart cities as well as electronic payments, just to name a few. Moreover in situation of crisis, such as the COVID pandemic, there was a huge pressure in many countries by public health agencies for the access to mobility data (*e.g.*, Call Details Records collected by telecom operators) to use it to understand how the movements of persons affect the spread of the disease as well as whether the population was respecting the confinement rules. While location privacy can be preserved from an external attacker through classical security mechanisms (*e.g.*, the use of TLS (Transport Layer Security) to secure communications between the client and service provider), one of the inherent risks is its possible abuse by this service provider due to rich inference potential of location data.

In the second part of his talk, Sébastien Gambs reviewed inference attacks against mobility data whose main objective is to quantify the risks in terms of privacy related to the collection and disclosure of mobility data. After defining the goal of location privacy as preventing an undesired entity from learning the past, present and future location of an individual [1], he discussed why pseudonymization offers a very low level of privacy protection for mobility data due to the possibility of using the pair home-work as quasi-identifiers if the spatial granularity is too much fine-grained [2] or the observations that the combination of 4 four random locations visited by a user is usually unique in the population as previous works have shown.

Afterwards, Sébastien Gambs presented some of his own work on inference attacks against mobility data such as the identification of points of interests based on a clustering algorithm [3], the prediction of mobility patterns using a mobility Markov chain [4] or a de-anonymization attack against anonymized mobility traces in which the mobility model is used as auxiliary knowledge by the adversary to re-identify mobility traces [5]. Other inference attacks briefly reviewed include profiling, exploiting the co-location of users to predict their location [6], as well as performing membership inference [7] or reconstructing trajectories from aggregated data [8].

The third part of the tutorial was dedicated to privacy-preserving methods for mobility data publishing. The objective of these methods is to sanitize the data to preserve location privacy, in particular by preventing some of the attacks described in the previous part. When sanitizing data, there is an inherent trade-off between the desired level of privacy and the utility of the sanitized data. Here, utility can be defined with respect to global properties of the data or be dependent on the application considered. First, simple sanitization mechanisms were introduced such as the geographical masks, which protect privacy through aggregation or perturbation, or approaches based on sampling or removing records that are too atypical. Afterwards, the spatio-temporal version of the k-anonymity privacy model, called spatial cloaking, was presented [9]. Methods that address privacy by limiting the possibility for the adversary to link together the mobility traces belonging to the same identity, such as mix-zones and Swapmob [10], were also discussed. Finally, the differential privacy model [11] as well as its applications for privacy-preserving mobility analytics [12] as well as privacy-preserving trajectory synthesis [13] were also briefly mentioned.

In his conclusion, the speaker mentioned that while there has been a huge scientific literature in the last 20 years on location privacy, the fundamental question of defining and quantifying location privacy in a formal yet actionable manner is still partially open. More precisely, what does it mean to have "good" location privacy?

- To be hidden inside a crowd gathered in a small area?
- To be alone in a desert?
- To have a behavior indistinguishable from those of a non-negligible number of other individuals?
- To be unlinkable between different positions?

Sébastien Gambs emphasized that whatever the approach chosen, it should prevent the inference of sensitive information from the location data revealed, rather than focusing on hiding only the location itself. Finally, he listed several open challenges for research on location privacy: (1) the lack of a set of (preferably few) privacy and utility metrics on which there is consensus in the community, (2) the absence of large-scale reference datasets that could be used to benchmark algorithms (mainly due to confidentiality and privacy reasons) and (3) the fact that often the source code of algorithms is not published. All these three challenges limit the possibility of comparing different sanitization algorithms and hinder reproducible science.

References

1 Alastair R. Beresford and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive Comput.*, 2(1):46–55, 2003.

2 Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe, editors, *Pervasive Computing, 7th International Conference, Pervasive 2009, Nara, Japan, May 11-14, 2009. Proceedings*, volume 5538 of *Lecture Notes in Computer Science*, pages 390–397. Springer, 2009.

3 Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and I will tell you who you are. *Trans. Data Priv.*, 4(2):103–126, 2011.

4 Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In Hamed Haddadi and Eiko Yoneki, editors, *Proceedings of the First Workshop on Measurement, Privacy, and Mobility, MPM '12, Bern, Switzerland, April 10, 2012*, pages 3:1–3:6. ACM, 2012.

5 Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *J. Comput. Syst. Sci.*, 80(8):1597–1614, 2014.

6 Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. Quantifying interdependent privacy risks with location data. *IEEE Trans. Mob. Comput.*, 16(3):829–842, 2017.

7 Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Measuring membership privacy on aggregate location time-series. In Edmund Yeh, Athina Markopoulou, and Y. C. Tay, editors, *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems, Boston, MA, USA, June, 8-12, 2020*, pages 73–74. ACM, 2020.

8 Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1241–1250. ACM, 2017.

9 Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In Daniel P. Siewiorek, Mary Baker, and Robert T. Morris, editors, *Proceedings of the First International Conference on Mobile Systems, Applications, and Services, MobiSys 2003, San Francisco, CA, USA, May 5-8, 2003*, pages 31–42. USENIX, 2003.

10 Julián Salas, David Megías, and Vicenç Torra. Swapmob: Swapping trajectories for mobility anonymization. In Josep Domingo-Ferrer and Francisco Montes, editors, *Privacy in Statistical Databases – UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26-28, 2018, Proceedings*, volume 11126 of *Lecture Notes in Computer Science*, pages 331–346. Springer, 2018.

11 Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

12 Mohammad Alaggan, Mathieu Cunche, and Sébastien Gambs. Privacy-preserving wi-fi analytics. *Proc. Priv. Enhancing Technol.*, 2018(2):4–26, 2018.

13 Takao Murakami, Koki Hamada, Yusuke Kawamoto, and Takuma Hatano. Privacy-preserving multiple tensor factorization for synthesizing large-scale location traces with cluster-specific features. *Proc. Priv. Enhancing Technol.*, 2021(2):5–26, 2021.

3.2 Mobility data analysis: ethical issues

Geoffrey Rockwell (University of Alberta, Canada)

Geoffrey Rockwell's presentation started with an example from the early days of the pandemic when Tectonix, a geospatial analysis company, tweeted a video of a visualization showing where people active during the March 2020 spring break on a beach in Ft. Lauderdale returned to after their break. Tectonix was promoting how their analysis and visualization technology combined with mobility data from X-Mode could help efforts to control the pandemic by tracking the irresponsible youth who come to Florida to party on the beach instead of social distancing.

This technology demonstration on the one had appalled some on Twitter who were surprised by how their privacy wasn't respected, but it also coincided with stories that the Trump administration was in conversation with various companies to see if location data could be used in anonymized form to track the novel coronavirus. An official quoted in The Washington Post article that broke the story said that mobility data could, "help public health officials, researchers, and scientists improve their understanding of the spread of COVID-19 and transmission of the disease" [1]. The point of this introductory case study was that there are significant ethical issues around how mobility data is gathered, used, aggregated, and shared. The variety of ways mobility data can be gathered, the inferences possible, and the number of players who are commercializing such data raise difficult problems around principles, privacy, the application of ethical frameworks and how to develop a culture of ethical use. In the rest of the talk, Rockwell surveyed these issues in order to provide a common ground for discussion.

Why bother with ethics? Ethics is often seen as a drag on innovation. Rockwell reviewed some reasons why addressing ethics is becoming important in both commercial and academic research and development [2].

What ethical principles apply? A common starting point in data ethics is to identify common principles that should guide researchers, developers and users. Rockwell surveyed some applicable sets of big data principles like the 2017 Big Data Guidelines from the Information and Privacy Commissioner of Ontario and principles from the Harvard Business School [3]. Researchers, however, often come to different conclusions, even if they start from similar principles. For example, on the issue of using mobility data for public health we found positions like: Ultimately, if residual risk is acceptable, analysis of mobility data can be justified if it can yield actionable insights that benefit public health [4].

From these discussions spurred by the current pandemic, *three major location data issues* emerged: (1) anonymized data sometimes are not anonymous, (2) location data are often not representative and can exacerbate inequality, and (3) location data are a key part of the extension of the surveillance state.

Rockwell concluded the case study by asking whether this moment of visibility for location data collection could provide an opportunity to push for new media literacies [5].

Privacy. Rockwell reviewed some of the definitions of privacy and how they might apply to mobility data [6, 7, 8].

Ethical Approaches. Finally, Rockwell reviewed some of the common ethical approaches that can be used to think through situations:

- Duty-Based Ethics where rules and guidelines are used to establish the morality of actions rather than their consequences;
- Consequentialist Ethics such as Utilitarianism which judge the morality of actions based on their consequences;
- Ethics of Care which emphasizes the relationships in actions and care for others.

Rockwell concluded by asking how we can develop a culture of ethics in data disciplines. Some general resources mentioned in the talk include also [9]. [1]

References

1 Tony Romm, Elizabeth Dwoskin, and Craig Timberg. U.S. government, tech industry discussing ways to use smartphone location data to combat coronavirus. *The Washington Post*, 2020. https://www.washingtonpost.com/technology/2020/03/17/white-house-location-data-coronavirus/.

2 Anne-Laure Thieullent et al. AI and the ethical conundrum: How organizations can build ethically robust ai systems and gain trust. Technical report, Capgemini Research Institute, 2020. https://www.capgemini.com/research/ai-and-the-ethical-conundrum/.

3 Catherine Cote. 5 Principles of Data Ethics for Business, 2021. Harvard Business School online, https://online.hbs.edu/blog/post/data-ethics.

4 Bouke C. de Jong et al. Ethical considerations for movement mapping to identify disease transmission hotspots. *Emerging Infectious Diseases*, 25(7):e181421, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6590736/.

5 Jordan Frith and Michael Saker. It is all about location: Smartphones and tracking the spread of COVID-19. *Social Media + Society*, 6(3), 2020. https://journals.sagepub.com/doi/10.1177/2056305120948257.

6 James Moor. The ethics of privacy protection. *Library Trends*, 39, 1990.

7 Kate Raynes-Goldie. Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook. *First Monday*, 15(1–4), 2010. https://doi.org/10.5210/fm.v15i1.2775.

[1] The slides of this tutorial are available at: http://www.master-project-h2020.eu/dagstuhl-materials/.

8 Danah Boyd. Debating privacy in a networked world for the WSJ, 2011. http://www.zephoria.org/thoughts/archives/2011/11/20/debating-privacy-in-a-networked-world-for-the-wsj.html.

9 Josh Lauer. Sarah E. Igo. The Known Citizen: A History of Privacy in Modern America. *The American Historical Review*, 124(3):1019–1021, 2019.

3.3 Connected vehicles and mobility data – work done by the EDPB

Peter Kraus (European Data Protection Board, Brussels, Belgium)

Peter Kraus gave a short presentation of how the European Data Protection Board (EDPB) works as a European Body, composed of Members from the different Member States of the European Economic Area and the European Data Protection Supervisor. The EDPB, being tasked with ensuring the consistent application of Europe's General Data Protection Regulation (GDPR), has issued Opinions, Binding Decisions, Guidelines, Best Practices and Recommendations. He also gave a short overview of the EDPB Strategy and the transnational enforcement done by the Member State authorities.

Concerning the topic of mobility, Mr Kraus highlighted the work done by the EDPB regarding contact tracing during the COVID-19 pandemic and presented the EDPB's Guidelines on connected vehicles [1]. These Guidelines analyse the different legal instruments that are relevant from a data protection perspective and their interaction: the GDPR and the e-Privacy Directive. The e-Privacy Directive, as a so called *lex specialis*, provides additional protection to electronic communications, whereas the GDPR sets the general framework.

Following the presentation by Mr Rockwell and the mentioning of possibly applicable ethical principles, Mr Kraus highlighted that the GDPR is known as a principles-based legislation, which in its Article 5 contains the legal principles of data protection, which in many respects are similar to the ethical principles mentioned for ethical use of mobility data. This includes, in particular, the principles of fairness, transparency and data minimisation. These principles are highlighted and operationalised by Article 25 of the GDPR – Data Protection by Design and by Default.

One of the possible legal bases under the GDPR to process and to use personal data, as well as one of the requirements of the e-Privacy directive to access personal data on an end device, is consent. However, consent has a list of properties for it to be considered valid consent in the legal context. It needs to be informed, unambiguous, specific and freely given. This relates back to the principles of data protection. Further, the GDPR requires that consent can be withdrawn and that this withdrawing must be as easy to do as it is to give consent. Therefore, when using consent as a legal basis, consideration should be given to what happens when the consent is withdrawn.

One way to avoid being in scope of data protection law altogether is to effectively anonymise personal data. However, as already presented by Mr Gambs, anonymisation is hard to do in practice. Encryption and hashing are frequently mistaken for anonymisation processes, whereas from a legal data-protection perspective, they merely amount to pseudonymisation. One has to be aware of re-identification attacks that may turn a dataset previously deemed to be anonymised into one that is not anonymised at all. Lastly, it was highlighted, that anonymisation is particularly difficult for location data due to the additional context that can be derived from a specific location.

Finally, a few examples that are provided in the Guidelines on connected vehicles were given, which showcase how the legal framework can be applied to use cases such as contact tracing or usage based insurance services.

References

1 European Data Protection Board. Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications, 2020. `https://edpb.europa.eu/our-work-tools/documents/public-consultations/2020/guidelines-12020-processing-personal-data_en`.

4 Topic-based Working Groups

Inspired by the round-table presentations of attendees and the tutorials and discussions, we identified the following main topics as of interest of the audience to build the new mobility data ethics "from the technical to the ethical": (1) What is/are the trade-off(s) between data privacy and data utility?, (2) Mobility data anonymity – asking whether such data can ever be really anonymous, and (3) Ethics on mobility data: what is unique? Which guidelines? While these three groups retained the focus on data of the seminar title and the focus on humans often implicit in the term "ethics", two further groups explored the need to go beyond this: (4) Mobility Data Analysis Ethics beyond the data and (5) Mobility data analysis beyond humans only: Tracking animals and moral agency.

4.1 What is/are the trade-off(s) between data privacy and data utility?

WG scribe and other members: Francesca Pratesi, Bettina Berendt, Thierry Chevallier, Josep Domingo-Ferrer, Ioannis Kontopoulos, Jeanna Matthews, Anna Monreale

To understand the bond between data privacy and data utility, in a first step one needs to consider or develop definitions and measures for both these dimensions. What is *utility*? How to measure it? How to do a utility vs. privacy (but also fairness and other ethical dimensions) analysis? How do we actually and effectively define and measure privacy?

As starting points, from the background and the past experience of participants regarding the processing of mobility data, the best way to act is to reduce granularity. Unfortunately, even if many papers in the literature prove that this approach can really work [1, 3, 2], there are two main problems: (1) sometimes it is hard to apply data minimization (indeed, it is hard to define what we need before knowing what you want to do); (2) there is a huge difference between one-shot collection vs. the continuous anonymization of trajectories (i.e., data streams). Regarding the first problem, purpose specification can help to overcome this, especially for researchers; moreover, the General Data Protection Regulation (GDPR) defines the "public interest" purpose and allows derogation for research, cf. [4].

However, connected to this latter point, there is the *secondary-use problem.* Although it is generally considered good practice for administration or companies to *reuse* data (even for potentially good purposes) without the need to perform potentially expensive and time consuming new data collection (e.g., establish new surveys), we must be aware that this practice violates the GDPR principles. An example that was cited are instances of (mis?)use of a centralized Corona app in Germany[2], an app that was quasi-obligatory for the use of several services (such as going to the restaurant), so this implied a huge coverage of the system. Therefore, special caution must be taken and ensure that the collected data will not used for other purposes during or after the emergency in the context of which they were originally collected.

Unlike the data-protection principle of purpose limitation, principles of open science allow and encourage secondary use; the drawback is that stored datasets can often be de-anonymized quite easily. This is one reason why access to open datasets is conditioned on some constraints, such as the definition of a specific research project. As an example, in the Netherlands, it is possible, for researchers working with a university or other institution associated with the national statistics office, to access the comprehensive and highly detailed Dutch Microdata. To get access, researchers have to submit their project idea. Once approved, they can access the data in a secure environment provided by the statistics office. Payment is per month plus per export (because each export is checked for meeting common privacy standards). [3]

A solution to this problem is to build and use synthetic data.

During the working group discussion, we focused on concrete use cases, such as local buses management (in terms of both real-time position and capacity) and journey planners. Such planners strongly depend on the starting point and destination of the user, on the means of transport (which can also be suggested by the service), and on the time of the day. You can also consider "mobility as a service application", where you can also have other utility integration (e.g., buying tickets) that involve other stakeholders (e.g., bus companies), and possible second uses of data. This helped us highlight different contrasting dimensions:

Tracing people vs. not tracing people. Both tracing and not tracing can be approaches to solving a problem, but these problems are different in that they have different needs and challenges. Some services can be obtained even without the tracking of people in a privacy-preserving way, so not tracing people could still be a good option. In the case of buses, for example, sensors can be installed on buses that can give the real-time aggregated occupancy, a value that can be exploited to avoid further data gathering. Clearly, when we do not track the location of buses or people inside it, we do not know the actual route. For this kind of service, avoiding the tracking we are actually applying data minimization: we do not need to track people and we will not do it. The difference in terms of privacy in the different scenarios is due to the need to know an immediate location characteristic (e.g., a bus position or whether a person is on a bus) versus the need to know the actual movements and, thus, to actually trace people (of course, only in this case we have to deal with personal data).

Historical data vs. real-time data. Historical data are data collected in the past and generally used to build some models; these models can be used at a later time for several uses. Here, already existing methodologies can be applied, such as randomization or k-anonymity. Once the model has been computed on historical data, the model will likely be used. Indeed, one must share his/her position to take advantage of it, resulting in a real-time scenario. While hiding persons in the crowd during the building of the model is relatively simple, when

[2] This was the app "Luca", not to be confused with the "main" German tracing app CWA, which is decentralised. Regarding the adoption rate of such measures in general, and the importance to provide incentives, see [5].

[3] `https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research`

the system is in use you have more constraints, because the environment is both dynamic and distributed. There is probably the need to define a new concept of privacy is this case. Indeed, in the real-time context, it is more difficult to compute the privacy risk since, in general, we do not have the general overview of the situation. The most promising approaches in this context seem to be *randomization* or *federated learning in location data.* The limitation of the first solution is that when you are the only person in an area, even when when noise is consistently being added, you will always be re-identifiable; thus, you could have no privacy and no utility at all. Federated learning approaches are similarly limited: when the peers have very different location data, there are no guarantees of protection.

Service quality vs. data quality. To assess the utility, we usually rely on service quality, measuring whether we are able to perform similar analyses with or without privacy. This represents what the end user loses with a privacy-preserving service. In the bus scenario, an example could be to suggest a longer route to reach a destination in order to preserve privacy: the privacy-preserving route could be more expensive in terms of more time but also more money for gas or the bus ticket. Therefore, we must be able to measure the difference in performance (e.g., accuracy

or cost-sensitive performance metrics) between the model with and without privacy (i.e., with private data or raw data).

The "kind of risk". This dimension is related to the risk you are taking into consideration. Examples are the re-identification risk or the disclosure of sensitive attributes (e.g., profiling). We can also consider if the risk is data-centric or real-life-centric. Regarding the risk of re-identification, there are some practical tools[4], based on the work presented in [6, 7].

Multi-criteria utility function. Utility is often equated with accuracy of the model, but accuracy is not the only objective. For example, there are the *fairness* trade-off or the *transparency* trade-off. We call these options *simple model vs. complex model*: one can prefer an explainable model over one with perfect accuracy. While on the one hand, we need to protect privacy, on the other hand, transparency is a valuable aspect that could be in conflict with privacy. How to achieve both *and* maintain utility? In this sense, we could start from the work done in [8], where some contrasting dimensions are considered (even if here the utility is missing), and [9], where it is highlighted that there are values that help each other and there are values that harm each other; it also depends on the choice you made.

Participation vs. non-participation. This is another trade-off that needs to be considered. We should be able to measure the trade-off between benefits and opportunity costs, consider disincentives, and be sure that subjects really understand the risk of participation. Moreover, it happens very often that a certain service cannot be pursued if the adoption rate does not reach a certain threshold. Thus, we are led to thinking that there is a strong need to collect large quantities of data. However, this thinking may be the result of assuming *false trade-offs*: usually, a lot of data means a high utility, and this leads to a high risk; on the contrary, no data means no utility but also no risk. However, this need not be the case, and with some transformed (e.g., partially or badly anonymized) data one may obtain something in the middle: no utility and full risk [10].

Stakeholder diversity. During the discussion, we agreed on the existence of three different stakeholders for both privacy and utility, aiming at different kinds of utility:

[4] `https://github.com/scikit-mobility/scikit-mobility`

- *subject*, who provides the data or uses the service. Regarding the privacy aspects, subjects who provide data have the specific need that their data cannot be re-identified. At the same time we want to able to maintain some data utility. In general, we can refer to *individual utility* that is how good is the service or *collective utility* meaning that a user can share his/her data if there are benefits for the society, such as reducing traffic in a city. Regarding the final user of the service, we can use other metrics. For example, as said before, a user can take more time by using a path different from the shortest with the purpose of preserving privacy.
- *data owner*, who is usually the provider owning the data, and who possibly permits to third parties to access its data. This subject has interest in keeping the ownership of the data and prevent uses by third parties without their consent; their utility is usually the profit. We can consider the utility from the perspective of data collector: you should check the difference of performances (e.g., accuracy) of the model with or without privacy (i.e., with private preserved data or raw data).
- *data analyst*, who is in charge of analyzing data and providing the service. This subject wants the confidentiality of results. Its utility can be recognized in profits if it is related to a private company or in benefits for the society if it corresponds to a public entity. More information about these three kinds of privacy can be found in [11]. Moreover, these different kinds of perspectives can be easily related: for example, when data is privacy protected, this is good not only for the subjects but also for the owner of the service. It is important to consider in this analysis also the economic concept of externalities [12].

To conclude, the last part of the discussion focused on the accountability and the responsibility, especially at the societal level, of the actual implementation and on the possible implications of the utility computation. In particular, we discussed the use of checklists, where the outcome is not a specific value but rather the starting point for the evaluation. These checklists come in the form of a list of questions to the user, such as "Did you consider this aspect?".

Finally, about the application of this data utility mechanism: even supposing that we have the perfect formula for utility computation, who is going to implement it?

References

1 George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. Privacy and data protection by design-from policy to engineering, 2014. ENISA Report. `https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design`.

2 Anna Monreale, Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Transactions on Data Privacy*, 3, 2010.

3 Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, and Dino Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(10), 2014. `https://doi.org/10.1140/epjds/s13688-014-0010-4`.

4 Michael Beauvais. GA4GH GDPR Brief: The public interest and the GDPR, 2021. `https://www.ga4gh.org/news/ga4gh-gdpr-brief-the-public-interest-and-the-gdpr-february-2021/`.

5 Mirco Nanni et al. Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. *Trans. Data Priv.*, 13(1):61–66, 2020.

6 Luca Pappalardo, Filippo Simini, Gianni Barlacchi, and Roberto Pellungrini. Scikit-mobility: a python library for the analysis, generation and risk assessment of mobility data, 2019. `https://arxiv.org/abs/1907.07062`.

7 Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, and Tadashi Yanagihara. Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. *Trans. Data Priv.*, 11(2):139–167, 2018.

8 Marit Hansen, Meiko Jensen, and Martin Rost. Protection goals for privacy engineering. *2015 IEEE Security and Privacy Workshops*, pages 159–166, 2015.

9 Josep Domingo-Ferrer and Alberto Blanco-Justicia. Ethical value-centric cybersecurity: A methodology based on a value graph. *Sci. Eng. Ethics*, 26(3):1267–1285, 2020.

10 Bettina Berendt. Better data protection by design through multicriteria decision making: On false tradeoffs between privacy and utility. In Erich Schweighofer, Herbert Leitold, Andreas Mitrakas, and Kai Rannenberg, editors, *Privacy Technologies and Policy – 5th Annual Privacy Forum, APF 2017, Vienna, Austria, June 7-8, 2017, Revised Selected Papers*, volume 10518 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2017.

11 Josep Domingo-Ferrer. A three-dimensional conceptual framework for database privacy. In Willem Jonker and Milan Petkovic, editors, *Secure Data Management, 4th VLDB Workshop, SDM 2007, Vienna, Austria, September 23-24, 2007, Proceedings*, volume 4721 of *Lecture Notes in Computer Science*, pages 193–202. Springer, 2007. https://crises-deim.urv.cat/web/docs/publications/lncs/457.pdf.

12 Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda F. Gürses. Pots: protective optimization technologies. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 177–188. ACM, 2020.

4.2 Mobility data anonymity

WG scribe and other members: Francesca Pratesi, Jeanna Matthews, Anna Monreale, Florence Chee, Ioannis Kontopoulos, Karine Zeitouni

In this working group, the discussion focused on multiple aspects of data anonymity, which is extremely hard to reach [3] for mobility data [1], especially considering the fact that mobility patterns are usually highly predictable. As experts, can we recommend (and develop) ways to use technology still but with some evasions of location tracking (e.g., download offline maps and navigate with GPS), and do our best to increase people's awareness, especially considering how we are exposed and easily manipulable [2].

Moreover, there is a need to consider the danger of predatory data collection. Sometimes, the choices are false choices (you "must" download an app due to peer pressure or to participate in ordinary social life). Therefore, we should encourage the definition of different levels of consent. Another possibility is to implement tools to help people to analyze the permissions required from mobile phone apps[5] or from web pages[6].

However, as said in Section 4.1, it is worth noting that the choice to participate in a project or to share something affects the final goal: the goodness of the model depends on the number of individuals that participate. And this has an impact on the individuals

[5] https://privacyflag.eu

[6] https://github.com/chatziko/location-guard

themselves. In the literature a scenario that evaluates this trade-off is missing: what is the impact (in terms of privacy loss) if you participate? And what is the impact (in terms of service quality) if you do not participate? Another challenge is that this must be computed and evaluated in real-time. Moreover, in the mobility context, it is not enough to have a certain percentage of the population, but we need to evaluate also the geographical area; it seems that the only possibility is to simulate this computation.

The discussion ended up with a possible trend of personal data management, providing decentralized solutions. There are solutions actually existing, such as personal data cloud, where you locally store data, but to be usable, they need suitable infrastructures and ad hoc protocols. This is not implementable today because there is no pressure neither from governments and surely not from companies. We need to ask ourselves if we have the power to influence choices, both at individual and at companies levels.

A follow-up article on the concerns of surveillance and the best practice we can adopt to contrast them has already been published [4].

References

1 Hui Zang and Jean Bolot. Anonymization of location data does not work: a large-scale measurement study. In Parmesh Ramanathan, Thyaga Nandagopal, and Brian Neil Levine, editors, *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MOBICOM 2011, Las Vegas, Nevada, USA, September 19-23, 2011*, pages 145–156. ACM, 2011.

2 Jeanna Matthews. How fake accounts constantly manipulate what you see on social media – and what you can do about it. *The Conversation*, 2020, Jun 24. `https://theconversation.com/how-fake-accounts-constantly-manipulate-what-you-see-on-social-media-and-what-you-can-do-about-it-139610`.

3 Alex Hern. “anonymised” data can never be totally anonymous, says study. *The Guardin*, 2019, Jul 23. `https://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds`.

4 Geoffrey Rockwell, Bettina Berendt, Florence M. Chee, Jeanna Matthews, Sébastien Gambs, and Chiara Renso. Ottawa’s use of our location data raises big surveillance and privacy concerns. *The Conversation*, 2022, Jan 27. `https://theconversation.com/ottawas-use-of-our-location-data-raises-big-surveillance-and-privacy-concerns-175316`.

4.3 Ethics on mobility data: what is unique? Which guidelines?

WG scribe and other members: Geoffrey Rockwell, Christine Ahrend, Florence Chee, Thierry Chevallier, Maria Luisa Damiani, Peter Kraus, Fen Lin, Fran Meissner, Alessandra Raffaetà, Chiara Renso, Paula Reyero Lobo, Yannis Theodoridis, Karine Zeitouni

The question of data ethics and technology ethics has been a hot topic for some time. Over the past decade following an explosion of the availability of various and voluminous data sets often referred to as big data, a common laminate has been that technology is often developing faster than the regulations that might keep the implications of those technologies and data uses at bay in step with the innovations in the field. The mobility data analysis field is not excluded and efforts are underway to engage with and “update” ethics thinking to new data and technology realities. Those efforts raise a number of questions. In particular, we discussed the questions described in the following Sections 4.3.1–4.3.3. As a result of the discussion, we discussed existing guidelines and drew up a tentative set of guidelines tailored to the needs of a graduate student doing a mobility data project, provided here in Section 4.3.4.

4.3.1 What is unique about Mobility Data from an ethics perspective?

In many ways the field of mobility data aligns with many other fields that have moved from relative data scarcity to a relative data abundance. So is there anything unique about the type of research? We here argue that while there are many parallels and there is a lot to learn from the ongoing wider debates, our argument here is that there is indeed something, if not unique, at least specific about the mobility data if viewed from an ethics perspective. Much of that uniqueness derives from the type of data needed for the mobility analysis.

One of the peculiarities of mobility data is that we can deduce/infer knowledge. Mobility data is first of all characterized by spatial-temporal information and this allows us to relate this data to the context, by considering the environment where the moving object is traveling. In this way, we can detect the points of interest that the moving object is crossing, the time spent in each place, their semantic categories. Based on the time and frequency of the places, we can infer the workplace and/or home. Moreover, we have the possibility to understand the interaction among the moving objects, if they are moving in a group, who they are in a meeting and to have an idea of the social network.

Some features of mobility data merit highlights.

- Scale: the amount of information from smartphones/sensors is massive. Such data expand the notion of temporal and spatial scales that used to be developed from the "small-and-static" data. Moreover, the granularity of the spatial and temporal information can vary significantly (e.g. days vs. seconds, km vs. cm).
- Associated Metadata: depending on the provider, mobility data can often come with associated metadata about the user or about the application.
- Auto-correlation: the data are not independent and they have incredible details correlated in time and space. The temporal component, especially, seems to add a new dimension to the ethics.
- Relational: the mobility data is beyond categorical information, it also reveals the relational behaviour – who were close to, who were you likely talking to, what might you have been doing.
- Behavioural prediction: due to the temporal and spatial detail, Mobility Data can be used to infer everyday behaviour in ways that many other sources of data don't. If ethics is about guidelines for behaviour, then having such detail about day-to-day behaviour raises questions about how behaviour can be surveilled, managed, and manipulated in new ways.

As a result, mobility data is much more than "simple" geographic data, and these data are incredibly valuable both for fine-grained analysis but also in economic terms. This value of mobility data brings with it an array of different interests in how and for what purposes mobility data ought to be used. At the same time it is also hard to anonymize full mobile datasets, and this creates challenges when it comes to sharing data for open science purposes.

Such features of mobility data imply that there is an array of normative questions that mobility researchers are faced with today. These questions go beyond, but also accompany, more traditional concerns such as how to preserve privacy of individuals in the data, as elaborated on in Section 4.1. This also means that due to the nature of mobility data is engaging with, articulating and addressing those questions is exceptionally important. The myriad of ethical considerations that need to enter the mobility data field have gained attention that is too fragmented to create clarity, for researchers, about applicable ethics guidelines and how to use ethical considerations as a means to engage in responsible research. The following takes the uniqueness of mobility data we here postulate as a starting point to engage with some important questions that are often not very clearly spelled out.

4.3.2 What is the role of context and culture in Mobility Data Ethics?

We recognize that ethical practices vary across contexts and cultures. For example, the MIT moral machine experiments that examined moral decisions of self-driving cars across 42 countries revealed a cultural divide in ethics [1]. The use and regulation of mobility data might therefore need to integrate contextual situations and normative ethical principles. At the same time it may be necessary to recognize that some technical solutions that work in one context might have adverse effects if applied in another context, making awareness for the context of analysis and application crucial in assessing ethics concerns.

Even though big data technology in general, and technology using mobility data specifically, enables researchers to efficiently understand the people's behaviour, which might offer insightful suggestions for policy-makers to develop better governance, it inevitably invites the conundrum of balancing utility and personal privacy. How to make such a balance might also vary by contextual situation (thus adding to the many other factors to take into account that are discussed in Section 4.1) and across cultures. The contextualized ethical practices might also be manifest in various stages of operationalization of mobility data: data collection, data analysis, data publication, data sharing and data storage. Such contexts might also bring issues that are related to ethics but go beyond ethics. For example, when a private citizen attends a public event, to what extent do they own their behaviour and trajectory data in the public event?

We recognize that there are communities that do not want to be surveilled and, in fact, want to remain hidden from the research gaze. Likewise, we recognize that researchers are not necessarily neutral or objective observers who have only the good of the community in mind. It is the responsibility of the researcher to engage communities that they are going to surveil in a dialogue and be willing to *not* gather data in some cases or to redesign their research in others. Further, there are now contexts in which it would be an "appropriation of voice" to do research which is not co-designed with the community. Indigenous communities in particular have made it clear that no research should be conducted without their involvement and approval.

On these debates much can be learned from ongoing debates about data justice. For example, Taylor [2] argues that one useful approach might be to take a capabilities approach – meaning to think through how individual and group capability is hindered or facilitated – to think through data justice. This requires thinking through at least three conundrums. (1) what sort of visibility is being made possible with the data and can access to representation as well as access to information privacy be maintained. (2) how is an engagement with the technology made possible that allows both to share the benefits of data and allow for an autonomy in technology choices; and (3) how can a principle of non-discrimination be adhered to in a way that allow for challenging bias and preventing discrimination. Facilitating capabilities is also an relevant aspect of debates that link data technologies to the questions of ethics.

4.3.3 What are the benefits and risks of ethics (and too much ethics)?

There is a balance between research utility and addressing various ethical concerns. In this sense, ethics should be part of the research process as a dialogue to ensure there is a minimum risk of harming people and this goes beyond complying with legal regulations.

Several factors involved with the use of mobility data advocate for the need for an ethics assessment to be in place before and during the whole lifetime of the project. Privacy preservation is a fundamental aspect, but not the only one. While there are questions about data manipulation and control which relate to both ethics and privacy, we discuss other important aspects of data ethics that differentiate these two concepts. The impact of the project needs to be understood to be able to make decisions such as which data to use or how to use it, in order for it to have the best possible outcomes. This is an intricate task, as the impact may not always be straightforward, for example, if the data being used may not be directly linked to an individual. However, is it important that researchers have a legitimate interest and data subjects are aware that their activities are being recorded and the purpose for which they are used?

Such question invokes conversation that were circled around the tensions between a) the value that we, as researchers, put on openness and the sharing of data and b) the privacy rights of individuals or groups whose data are collected. This is particularly a problem in mobility data research due to the difficulty involved in anonymizing datasets. In fact, sharing the data and the methods used in a research project is essential in open science. It allows the community to check and reproduce the results that support the findings. It is the basis of evaluating and benchmarking different approaches.

Some of the approaches discussed that could mitigate this tension include:

- Minimize your data to that which is needed for replication.
- Researching and applying new practices in anonymization. We recognize that this is itself a research field where approaches are changing.
- Embargo your data for a number of years until the privacy challenge is lessened.
- Ask people to apply for access to your data and provide it with clear terms of use.
- Provide synthetic data instead of the real data.

4.3.4 A Tentative Guideline

On the second day we discussed guidelines. We were inspired by the Locus Charter [3]. We agreed that the guidelines could be organized by audience and different sets of guidelines might be proposed at different stages of processing data, such as data collection, data analysis, data storage and data sharing. We proposed a tentative guideline for a graduate student doing a mobility data project.

Guidelines for a Graduate Student doing a Mobility Data project

Policy. Inform yourself about the research ethics and data privacy policies of your university and research lab. If there is a research ethics board to which you need to apply to proceed with your project then you should engage that board. Don't think of the process as an obstacle, but as a chance for dialogue.

Data. Identify where you are getting your data from. Are you gathering the data yourself, or are you getting data from a mobility provider, or have you inherited a data set? Make sure you fully document the provenance and structure of the data even if you have inherited it.

Transparency and Consent. If you are gathering data you should consider how you could get consent or at least be transparent about what you are gathering and why. You might, for example, have a notice at the location where you gather data or you might publish a web site that provides information on the project. As part of your transparency you should describe what you are gathering, what you will do with it, and how someone might hold you accountable.

Security. Develop a data security plan. Check it with your supervisor and data security colleagues. This is especially important if you are working with data that can identify individuals or is sensitive in some fashion.

Check for Bias. Ask yourself if your dataset is biased. Who is represented? Who is not? Does this make a difference to the research? Familiarise yourself with the literature on bias so you understand what might be reasons for bias. Test your dataset for bias if you can.

Research Goal. Ask yourself what you are trying to achieve with this research and whether the research project itself is ethical. Are you trying to do good, or could there be unintended harms from misuse of your data or research? Is the new knowledge generated helpful, useful, or welcome?

(a) A good starting point is to identify your values, including the often unexpressed values embedded in Western modes of research. What do you think is good such that your research will make a difference.
(b) It is a good idea to identify the stakeholders in your research, both those whose data is gathered and others (including the research team). Engage the stakeholders in a dialogue about the design of the research.
(c) Consider how they might have different values. They may not value the research you are doing or value it differently.
(d) Make sure you are transparent about the goals of your research so people understand the purpose for the data gathering, the analysis, and publication.
(e) Consider the ethics of the methods and analytics used in the project. Could your methods be unethically applied in a different context, or with different data?
(f) Audit the research. Share your research goals and plans with other people, starting with your supervisor and colleagues in order to get input on the ethics of the means and the goals. If there is an ethics board, or people with a designated ethics role, you should get their feedback early and often.

Data Management Plan. Develop a long-term DMP that describes how you will process your data to minimize it, what you will deposit in the university research data repository for future use, whether there should be an expiry date to the data, and whether it should be embargoed for a while. Consult with research data librarians about the plan. Ask yourself and others how your data could be used for good or misused. Document your wishes for the data (and include this with the data deposited) so you are clear how you hope it will or will not be used.

Ask yourself about possible second use cases. Even if your own research may adhere to best research standards you should audit your data and project. Could your data, the methods you develop or the findings you plan to publish lead to adverse effects? If so, what are those effects and can you think of and actively implement mitigation strategies that will prevent adverse effects and promote socially responsible reuse of your work?

Care and Repair. Maintain and repair your data, methods and research publications.

References

1 Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337, 2020.

2 Linnet Taylor. What is data justice? the case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2):2053951717736335, 2017. https://doi.org/10.1177/2053951717736335.

3 EthicalGEO. Lotus charter, 2021. https://ethicalgeo.org/locus-charter/.

4.4 Mobility Data Analysis Ethics beyond the data

WG scribe and other members: Fran Meissner, Fen Lin, Florence Chee, Chiara Renso, Paula Reyero Lobo, Yannis Theodoridis

Debates about the ethics of mobility analysis tend to focus on the data and what it reveals. They are often framed in light of privacy and the problems with guaranteeing privacy given the specificity of spatio-temporal data compared to other categorical data. To take a step back from this perspective, we explore what kinds of ethics questions surface if we think about the ethics of mobility data analysis beyond the data. Engaging in such an exercise is helpful to consider mobility data and its analysis as part ofmobility information infrastructures. These infrastructures include broader institutional infrastructures that make mobility research possible (think servers, data specialists, multi-origin data sources) and political and economic infrastructures that influence research priorities.

As with many other fields, mobility data analysis used to grapple with relative data scarcity. New data collection and processing capabilities have transformed how mobility research is done. More and more actors interested in technological innovation and monetising data are entering the field. Such changes bring opportunities and concerns. These require a broader engagement with ethics – with what it means to do good. The lure of innovation itself can be a concern as it often trumps a real need for innovation. Use cases are frequently found after the fact. Given economic interests that undergird the field, it is not immune to a mentality of "move fast and break things" that leaves little room for considering ethics and long-term implications of new mobility data ecosystems. The concerns raised in our discussion centred around questions of consent, equitable access, reproduction of power asymmetries, and fundamental questions about who should benefit from mobility research how.

Our discussions involved thinking about issues linked to a digital divide, both in terms of data produced and the mobility tools that data is used to create. New ideas about mobility as a service, for example, are frequently thought about in terms of creating sustainability and efficiency. Still, social aspects require considering a diversity of stakeholders who might be disadvantaged by hyper-efficient and environmentally sustainable systems. City building always tends to happen both as a byproduct of natural growth and the engineering of cities – cities are arenas where there is an interplay between policy and new knowledge. At the level of ethics, this raises the question of what kind of city we want to live in. It seems complicated to imagine a truly ethically 'optimised' mobility data application without concerted efforts engaging with that question.

The question of third-use scenarios and the pivoting of services that might have been conceptualised with good intentions is another concern that looms large in the field. Beyond this, there is an increasing growth of hybrid applications that take advantage of mobility data information infrastructures. One prominent example is how developers might draw on mobility choice data for identifying investment areas – accelerating problems with already tight housing markets in many urban areas across the globe.

Questions of access and equity also matter for how mobility data becomes bound up in platforms and projections of how mobility should work in the future. For example, while visions for developing so-called EU "mobility data spaces" are framed as openly accessible platforms, de facto who registers to those spaces is limited, as technology knowledge and resources required to access such platforms are pretty high. This might perpetuate inequities as it consolidates the prominent role of large corporations and creates dependencies that may not necessarily be just or desirable.

There are different technological solutions envisaged to bridge these different access inequalities. For example digital identity management may offer inclusion opportunities also to people without legally recognised identification [1]. At the same time, a drive towards digital identity verification is often under scrutiny for their surveillance potential. In terms of facilitating access to data, debates hone in on interoperability standards or so-called mobility data marketplaces.

Privacy remains a significant value. Privacy and how its protection is interpreted requires critical evaluation. Issues that arise are those linked to reproducing prevailing power inequalities. For example, in debates on who ought to access and share what kinds of mobility data, customer data is often excluded as too sensitive. In effect, this means that this data stays with those who can harvest it. It also means that those with access to more detailed customer data have a competitive advantage over others as "innovative" mobility platforms grow. While it remains unclear how to counter this type of development, it is important to re-emphasise that debates about a multiplicity of principles – with privacy being one – have to continue to ensure that mobility research does not consolidate undesirable social patterns.

At the same time, innovations in the field are also shaking up who the relevant players are. Some startups – like, for example, Citymapper – might perform better at facilitating mobility for their subscribers than the transport operators themselves. This eventually also gave them a stake in the mobility debates, even though their practices might not fall under the same regulatory frameworks as public service providers. This is not to say that such changes are necessarily ethically suspect – various business models are emerging, some of which sincerely have social and ecological sustainability at their heart. Examples here might be the establishment of data trusts in some Canadian cities that conceptualise mobility data as a public good. Similarly, the ideals of Gaya X at the EU level are pointing us to a rethinking of priorities. Still, a prevailing pattern is strong incentives to cut corners to stay competitive.

References

1 McKinsey Global Institute. Digital identification: A key to inclusive growth, 2019. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-identification-a-key-to-inclusive-growth.

4.5 Mobility Data Analysis Ethics beyond humans only: Tracking animals and moral agency

WG scribe and other members: Alessandra Raffaetà, Bettina Berendt, Maria Luisa Damiani, Stan Matwin, Chiara Renso, Geoffrey Rockwell

In this group, we considered issues around the dignity and rights of non-humans when it comes to location data. We track, for example, land animals (individuals and herds) and fish, for all sorts of reasons. The location data can be used to study them or to hunt/fish them. This raises a number of ethical questions.

Rights from dignity. Animals with dignity implies the existence of rights for such animals. For example, pets are considered by many to have dignity, so there are laws that punish their mistreatment or abandonment. We can say that rights are proportional to dignity: more dignity, more rights. It is also a matter of relationships. You are tied to your cat or dog so you recognize that they have some rights.

Rights from suffering. The philosopher Peter Singer [1, 2] states that all beings capable of suffering have to be worthy of equal consideration and that giving lesser consideration to beings based on their species is no more justified than discrimination based on skin colour. He argues that animals' rights should be based on their capacity to feel pain more than on their intelligence. Animals can suffer so they should have rights, hence they should be protected.

What about fish? We do not consider fish as individuals but as species. Our attitude is to take care of the species' survival in order to maintain the ecosystem. We have an utilitarian approach with respect to fish. We want to avoid over-exploitation because this can provoke loss of work (for the fishers) or loss of food for some populations. In this context, domain experts like ecologists are useful to establish policies for sustainable ecosystems; such policies are utilitarian rather than tied to the deontological notion of dignity.

What about insects? We usually kill mosquitoes, and in agricultural ethics it is argued that individual insects do not have a "right to life." [7] Animal ethics depend on culture and religion. Let us think about Buddhists or Jains – they try to do no harm to *any* animal.

Issues concerning research and animals

We also discussed issues around the ethics of research and animals.

Tracking animals. We track animals to understand their behaviour; is that ethical or not? We use datasets reporting the movements of animals assuming that they do not have any privacy/ethical concerns. Is it really true? Lennox et al. [3] highlight the risks associated with animal tracking. In fact, bad actors, such as poachers or also photographers, could intercept animals directly using tracking hardware or indirectly from research results, databases and maps that provide the positions of vulnerable animals, and they could chase or simply disturb them. Thus, uncontrolled access to such location data may ultimately compromise the welfare of wild animals and the recovery of species. In [3] some techniques are discussed to protect animal data from misuse, such as data blurring, data aggregation, and data hiding.

[7] https://en.wikipedia.org/wiki/Insects_in_ethics

However, it is worth noticing that anonymization approaches used with human data are less useful in this context because the identity of an individual animal is rarely important. A good trade-off to delivering research data could be providing time-varying density location data extracted from trajectories. In this way, the whole trajectories are not shared and they are protected.

Bridging movement ecology and human mobility. Movement ecology is a relatively new discipline in the field of ecology that studies the spatio-temporal patterns and processes at the basis of animal movement [4] . While data and analytical methods are similar between movement ecology and human mobility, there is surprisingly little interdisciplinary awareness of these similarities. Developing new methods to integrate human mobility data (e.g. road traffic or human recreational activities) in the study of movement ecology would crucially improve the ecologists' understanding of the tight relationship between animal movement and human activities, as for example to unveil the effect of COVID-19 human lockdowns on animal movement and behavior. The vision of an integrated science of movement bridging the two research fields is however hampered by the limited availability of open data on human mobility.

Sustainability of research. As a more general issue, we need to reflect on what research looks like, how it is conducted, and also what its impacts on the environment are. To what extent should we consider not only humans and animals, but also ecosystems and the planet as a whole as moral agents (or patients)? It is important to consider alternative or hybrid formats of conferences that are more inclusive and ultimately more sustainable. In fact, having virtual conferences can allow researchers who are excluded for geographical and financial reasons from participating and at the same time the carbon footprint is reduced [5].

References

1 Peter Singer. *Animal Liberation.* Harper Collins, 1975.

2 Peter Singer. *In Defence of Animals: The Second Wave.* Blackwell, 1985.

3 Robert J Lennox et al. A Novel Framework to Protect Animal Data in a World of Ecosurveillance. *BioScience*, 70(6):468–476, 2020.

4 Federico Ossi, Fatima Hachem, Francesca Cagnacci, Urska Demsar, and Maria Luisa Damiani, editors. *HANIMOB@SIGSPATIAL 2021: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility, Beijing, China, 2 November 2021.* ACM, 2021.

5 Chelsea Miya, Oliver Rossier, and Geoffrey Rockwell. *Right Research: Modelling Sustainable Research Practices in the Anthropocene.* OpenBook, 2021.

5 Online Interactions in a COVID-19-era Dagstuhl Seminar: Design, Experiences, and Reflections

Bettina Berendt (TU Berlin, DE)

This three-day Dagstuhl Seminar took place in January 2022, which for many participants was month 23 of a string of more or less harsh COVID-19 lockdowns and home-office conferences. The seminar had originally been planned as an on-site and then as a hybrid workshop. Given the skyrocketing COVID-19 infection counts throughout the world and the suddenly quickly increasing number of participants who were forced to or chose to participate online, we decided, in the week before the seminar, to hold it fully online. We were quite concerned about this decision, given the stark contrast between many participants' fond memories of the interactivity of pre-COVID Dagstuhl Seminars and the growing Zoom fatigue that we had been witnessing in ourselves and in our colleagues. We therefore contemplated what we could do to make the most of our three days together.

We (the organisers) were happy about the results of our strategy, and we received very positive feedback from several participants. Of course, our choices could not fully reproduce the experience of an on-site Dagstuhl Seminar. However, we believe that online and hybrid meetings have many advantages and will be an important part of how we all will be working in the future. Therefore, in this section, we want to share our experiences and observations. We hope that others can profit from them.

5.1 What do we consider "success"?

There are many possible measures of a Dagstuhl Seminar's success. Based on our personal experience, we especially hoped for (a) academic output (especially of the type(s) aspired to by the participants of the seminar), (b) a pleasant working atmosphere: interactive, inclusive and with the right balance of focussed academic exchange, loose brainstorming, and everything in between, (c) meeting new colleagues or interacting with known colleagues in new ways, so as to enter into meaningful conversations (which often begin as loose brainstormings, often about personal matters), and (d) retention of participants over the course of the seminar.

We believe that we have met goal (a) (with, for example, an op-ed in The Conversation [1] about seminar topics published two weeks after the seminar's end; other activities are ongoing). This is related to the self-reported goal of a majority of participants at the beginning of the seminar ("an extended paper abstract"). All participants who gave us feedback expressed (b), and this is in line with our (the organisers') own impressions. We did not ask participants about (c) but can confirm this to be true for ourselves.

With regard to (d), most participants attended the seminar all day, which we found remarkable given our observations that project meetings and workshops tended to become shorter over the course of the pandemic. Some participants (especially senior ones) had to leave the seminar repeatedly for other duties, and we observed how hard it was for participants in substantially different time zones to participate in all sessions. However, due to the informal organisation this did not present serious problems, and in fact some participants who were not able to participate throughout still managed to be central for their working group's progress and outputs, even if in a partly asynchronous way. We are also especially impressed by and grateful to those participants from +7 hours to -8 hours time zones who participated in all sessions. So in sum, we believe we were reasonably successful along this criterion. Since we do not have statistics about participant retention in other online formats, it is even possible that comparatively speaking, we were very successful.

5.2 What did we do, and what did we learn from it?

We believe that four elements of interaction design and participation choices were key.

5.2.1 Zoom + Gathertown + jointly edited documents

A specific combination of two different types of videoconferencing software, in our case Zoom and Gathertown, proved very helpful for engagement. Our hypothesis had been that Zoom fatigue is not only a physiological effect and the lack of the full breadth of "real" interaction, but also very much a consequence of the types of interaction favoured by Zoom (or similar

VC systems that are essentially generalisations of bilateral video telephony): a meeting at a specific time in which there is one speaker at a time. Zoom breakout rooms allow for a restructuring of a meeting into smaller meetings, but again they create public fora for well-defined groups with set start and end times. Parallel conversation threads are possible (in private chats), but usually only with one other participant at a time, and they always carry a certain feeling of clandestine conversation. So in sum, Zoom appeared to us to be an excellent tool for presentations, for goal-oriented discussions in "work mode" (i.e. under strict time settings), and suitable for the informal interactions needed to get to know people (or get to know them better) at most in small groups in which trust is established through a common goal or between people who already know each other well. Importantly, group membership and partaking in a conversation are binary choices, marked by entries and exits that are rendered very visible by the software.

Some of us had attended work meetings in Gathertown before. Gathertown allows for a categorically different type of being in a group and partaking in a conversation. First, Gathertown employs a spatial metaphor: groups congregate in locations, and one can *see* a group (as a set of avatars, rather than a list of names as in Zoom) and *move* there (and be watched moving by those in the group and others) rather than be "teleported" as in Zoom. Membership is *gradual*: the closer one gets to another person, the better one sees them (and is seen by them), and the better one hears them (and is heard). By positioning one's avatar at the "outskirts" of a group, one can signal a certain looseness of belonging, and if one does this, does not take part actively in the conversation, and then moves on, it feels a lot less abrupt and impolite than dropping out of a Zoom breakout room. In addition, Gathtertown's Minecraft-like environment may particularly appeal to people who enjoy video games, but also for others it offers gratuitous (but still potentially effective) signals of shared environment luxury: the cute watercooler, the plants, the sofa, rather than only the often bleak and distorted camera view of somebody else's home, office, or virtual background. We considered these elements key enablers of the pleasant and informal and fluid conversations that are so generative of creative exchanges with (hitherto) strangers in physical Dagstuhl meetings. We also considered the *spatial* and *motion*-based interaction of Gathertown to be particularly suited to our topic of *mobility* data mining.

We were privileged to have a colleague who is an expert of Gathertown designs and was enthusiastic to help us: we are incredibly grateful to Adem Kikaj of KU Leuven – his work has been a major factor in making this Dagstuhl Seminar a success! We are also grateful to people from whom we learned interaction techniques (in former Dagstuhl Seminars as well as in the planning of the current one). We designed an action model for the seminar and a spatial manifestation of this action model (with combinations of frontal presentations, fishbowl interaction, posters that people would be able to come back to to remember other participants' introductions of themselves and have bilateral or small-group discussions, etc.). In line with the real Dagstuhl experience, we also designed a lobby (with small seating groups, implemented as Gathertown private spaces, to mimick the availability of various meeting rooms in Dagstuhl castle) and a dining hall. We also planned a sports room for joint gymnastics exercises, a feature one of us had seen and appreciated as an activity during breaks in a different conference, and a nod to Dagstuhl's sports equipments/rooms. Adem built this world for us and enabled us to apply the finishing touches (which would only be possible during the seminar and/or live) ourselves.

Another core element in the media mix were jointly editable documents and folders. We used Google docs/folders for the landing page, the participant-to-working-group assignment, shared resources such as participant self-descriptions and tutorial slides, and working group results (each working group had one folder that they were free to fill).

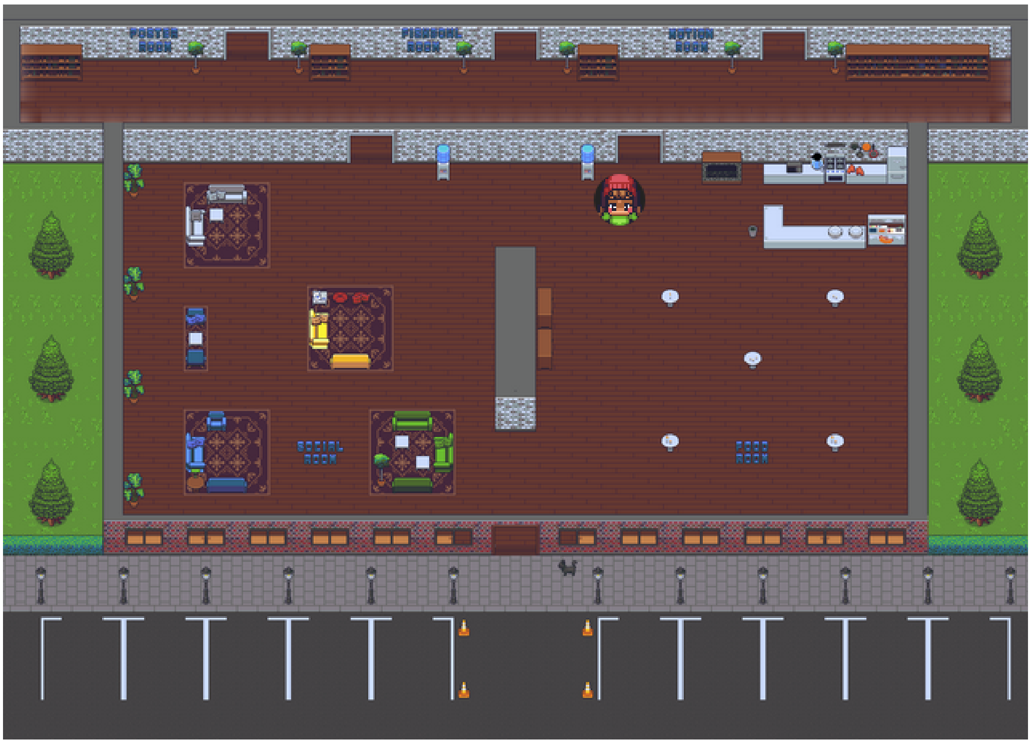


Figure 1 Gathertown: entrance area (map view).

The services themselves of course come with their advantages and disadvantages. We chose two services that are widely known and easy to use (Zoom and Googledocs) and one that, as far as we know, is the only one to provide a certain type of interaction (Gathertown). Familiarity is, in our experience, important, but it does raise questions with regard to providers and their (e.g. data protection) policies. Future designs could prioritise other features more (such as open source or provider characteristics).

5.2.2 Less is more

As explained above, the Gathertown world was quite elaborate and beautifully supportive of different types of planned interactions. In order not to overwhelm participants, we however decided to do the first day fully in Zoom and with Googledocs (for familiarity) and gave participants a first tour of the Gathertown world only during the “reception”, an extra one-hour slot after the close of the official “work” programme. Many (but not all) participants accepted this “invitation”, and we did our best to make this a tour, picking people up at the entrance (and returning to meet individual latecomers), showing them around, explaining the functionalities, etc. Observing their interactions, we (the organisers) quickly decided to use Gathertown for the breaks only and keep the “work” programme in Zoom, and we enforced neither the use of all rooms in the Gathertown world (for example, we supplanted the originally planned poster-room by a simple folder on Googledocs) nor of all functionalities (such as shared whiteboards, private spaces, and the fishbowl). We also dropped the idea of our gymnastics offer, instead only making it known that we’d be in the Gathertown for the “morning coffee” (half an hour before the official start every day), the two coffee breaks and the lunch break, and the “reception”. The “morning coffee” break was taken over by one of

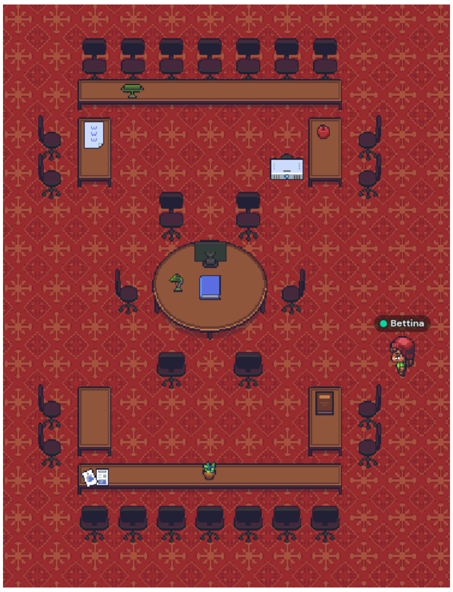


Figure 2 Gathertown: room designed for supporting fishbowl discussions (partial view).

us who likes getting up early, and the "reception" by one who tends to work later during the day. In sum, an even simpler Gathertown world would have sufficed (and may have required less bandwith).[8]

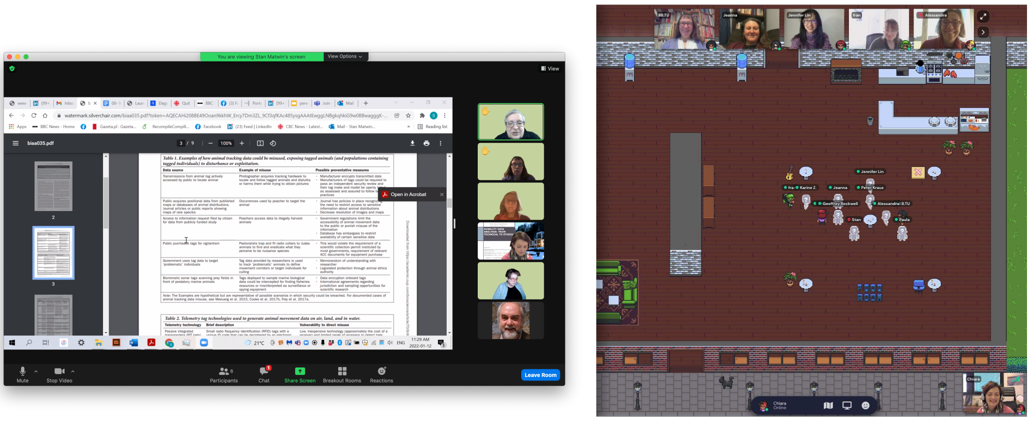


Figure 3 "Work" and "break".

We believe that this adaptivity to the actual interactions was important and prevented a sense of over-structuring that we had observed in earlier online events we had taken part in, a structuring which may contribute to work productivity but can feel stifling at times.

There was one element that we *added* (while all other adaptations were removals): one participant reminded us of the Dagstuhl tradition of randomising (but then enforcing) seating at meals. We implemented this with a simple hack that illustrates that complex Gathertown

[8] We learned after the seminar that a previous Dagstuhl collaborator created a virtual replica of Dagstuhl itself. While we were intrigued by this and probably would have used it had we known about it, based on our experiences, we believe it is not functionally necessary and may even prove somewhat overwhelming. This question should be explored in future work.

spaces can also be pragmatically adapted under time pressure: we put different ready-made objects (a plant, a guitar, ...) next to the different tables, created a random assignment of participants to tables, wrote this on the landing page, and showed it just before he lunch break, asking people to go to their assigned table. These arrangements proved to not work as deterministically as in the physical Dagstuhl, but they did create the same effect: unexpected gatherings and chitchat over lunch with participants one had not known before.

These little tricks encouraged also timid people to interact more with other participants they did not know before. It was not as effective as in physical Dagstuhl, since some people preferred to take a real break from the online event and also had to prepare their food, but a start.

To future organisers who consider using meeting software that is not widely known, we recommend trying out an interaction space such as Gathertown intensively. We did this in a "dress rehearsal" with two organisers and two participants in the week before the seminar, but at the time, we did not anticipate the dining seating arrangements. It was only during the breaks that we realised we had positioned the "tables" too close to each other relative to the visibility/audibility radius in Gathertown, such that conversations at different tables mixed in a confusing manner, and we had to artificially "move into the corner behind the table" to avoid disturbing the neighbouring table. This was amusing rather than really annoying, but we would certainly improve this design element next time.

As a consequence, there was one great simplification: Conceptually, there was one space for "work" (Zoom and shared documents) and one space for "breaks" (Gathertown). Within the "breaks" space, only the two rooms near the entrance (dining hall and lobby) were actually used, but this may have been a consequence of the number of users and the space being sufficient for them. The two organisers who were responsible for "hosting the breaks" complemented this conceptual separation by a technical one, running the "work" space on one computer and the "breaks" space on a different one. This improved audio and video quality considerably and made it easier to mentally control and survey the spaces. One participant remarked on this clean separation by relating their experiences from another online event, which employed a very elaborate Gathertown world with dedicated spaces for every activity. They had found that too complicated and overwhelming and welcomed our simplicity.

5.2.3 Structure + flexibility

We combined a simple outer structure with guidance on the one hand with freedom in the details of how working groups were to function. The outer structure consisted of (i) a consistent timetable for all three days, (ii) a rough specification of the functions and expected outcomes of each half-day slot, (iii) a self-presentation round in which each participant was asked to fill in the same Powerpoint slide consisting of three questions to present themselves, with slides sent to the organisers ahead of the seminar,[9] (iv) a tutorial round in which three relevant angles on the seminar topic were introduced by experts in the respective fields (see Section 3), (v) working-group formation and reporting rounds in the plenum, and (vi) working-group sessions (see Section 4).

[9] The slide asked for: name, affiliation, "my question" (that I would like to see answered at the seminar), "expected seminar output" (4 choices + free-form), and "an image of an object that symbolizes your interest, motivation, curiosity, doubts, ... regarding this seminar". All templates that we believe could be helpful for future seminars can be downloaded at http://www.master-project-h2020.eu/dagstuhl-materials/.

Media use was kept simple: one landing page (a Googledoc that was adapted during the course of the seminar) which pointed to all other resources, one Zoom main room which was always open and either in use by at least one person or equipped with a slide that indicated the current status (such as "lunch break"), and the Gathertown world which was also always open. A first version of working-group topics was created by the organisers clustering participants' answers to a question about the personal goal for the seminar (included in the self-presentation slide), but then modified in an open group discussion. This as well as the assignment of people to working groups used a jointly edited spreadsheet.[10] Sessions for five working groups were planned adaptively to also take into account when participants in American time zones would be able to attend (better). The working groups were numbered and met in Zoom breakout rooms of the same number, which were only opened during the times previewed in the time plan. Participants were encouraged to also meet with their working groups or with others outside these slots, and the use of Gathertown for such meetings was explained. This last option was however, as far as we know, not used much, since the "break" times were needed for relaxation, and remaining online for socialising after a full day of online work was not really an option for most participants.

One organiser monitored the Zoom main room throughout and another one did so in the Gathertown world. During working-group sessions, this presence was a secondary activity from a different computer. This ensured that participants who arrived at an unusual time (e.g. due to other commitments or time zones) were never "alone" and could always be briefly informed of current activities and locations.

Working groups were free to choose any type of media they wanted and any type of reporting, and other structures (such as the dining seating arrangements) were not enforced. This flexibility was used and appreciated by participants.

Visual consistency was ensured by a graphics expert's providing slide templates for self-presentation and organiser-provided slides: we are very grateful to Beatrice Rapisarda of ISTI/CNR, who also created the "group photograph" included at the beginning of this report from a collection of screenshots we had taken throughout the seminar.

5.2.4 "I am not here"

Many of us had experienced, over the course of the pandemic, a "densification" of work: an increasing number of work meetings from the home office, with meetings often taking longer than previously and positioned back-to-back. Attendance at conferences, which now involved no physical travel, was regularly disturbed by the assumption of colleagues that all the usual work could be done in parallel. Many senior colleagues we spoke to concluded that at some point, they just stopped going to conferences because it was "too much" or "not worth it". We believe that this overload contributed at least as much to Zoom fatigue as the physiological factors and the restricted interaction mentioned above.

Without having coordinated this in advance, we found that (at least) two of the organisers and one other participant had independently decided to break this pattern for this seminar. They had told their colleagues and superiors that they would be "at" Dagstuhl and not

[10] Interestingly, the spreadsheet was chosen by participants over an initially planned process involving spatial movement that one of us had observed as an efficient group formation process in earlier Dagstuhl Seminars. We had envisaged a replication of this technique in the Gathertown spaces, but after one participant proposed the spreadsheet, everybody immediately agreed that this would be the best method. It supported the same functionalities: simultaneous and mutually influencing self-assignemnt and observation of other participants' choices.

available during the three days of the seminar, and severely restricted their email monitoring and processing. One participant from the Americas had even put themselves into the Dagstuhl time zone, ensuring support from their family for this "absence". While the three days were still physiologically very taxing, all three of us agreed that this had been very sensible choices, needed for sustainable participation in future seminars, conferences, etc.

5.3 A sense of place in a hybrid world – and other parallels between the medium and the message

One element stands out from these observations, and it echoes the seminar's themes. Media use and interactions provided a "sense of place" that helped structure both (a) a prima facie non-spatial online world and (b) the hybrid nature of contemporary work. The unique and permanent "entrance points" (the Web-based landing page, the main Zoom room, the Gathertown world) provided stability and simplicity. The clear association of Zoom with "work" and Gathertown with "breaks" (or, in some cases, also corridor talk in parallel to the official working group meetings) helped participants to focus respectively relax. Gathertown's fluid interaction supported informal chats. Once inside a space (in particular in the breakout rooms), individuals and groups had all the freedom of acting there. The additional choice made by some participants to effectively "remove" themselves from their main place of work, in order to be fully at the seminar place, proved beneficial.

The sense of place echoes the *mobility* part of the seminar's themes, *Mobility data analysis: from technical to ethical*. We believe that the other key elements were also reflected: the *technical* obviously in the explorations of online-media design and uses, and the *ethical* through our general approach to research design, which we base on an ethics of care. This involves the recognition of relationships between people as at least as fundamental as autonomy and a desire to create, sustain and honour such relationships in research design. This meta-notion of care and ethics, first explored by the author of this section – together with two participants of this Dagstuhl Seminar – in [2], informed the choices described in the present section. In this sense, the present section is not only an exploration of different videoconferencing systems, but also a complement to the section author's work as the MASTER project's Independent Ethics Advisor.

References

1 Geoffrey Rockwell, Bettina Berendt, Florence M. Chee, Jeanna Matthews, Sébastien Gambs, and Chiara Renso. Ottawa's use of our location data raises big surveillance and privacy concerns. *The Conversation*, 2022, Jan 27. https://theconversation.com/ottawas-use-of-our-location-data-raises-big-surveillance-and-privacy-concerns-175316.

2 Todd Suomela, Florence Chee, Bettina Berendt, and Geoffrey Rockwell. Applying an ethics of care to internet research: Gamergate and digital humanities. *Digital Studies/le Champ Numérique*, 9(1), 2019. http://doi.org/10.16995/dscn.302.

6 Conclusions

Chiara Renso, Bettina Berendt and Stan Matwin

We believe that this seminar represents a successful first step in building a community of scientists around the mobility data ethics theme. The five topics that we identified and discussed in the Working groups are stepping stones from which the community can extrapolate new research topics. Indeed, at many points of our three-day journey, we realized that more research and reflections are needed to properly address these issues. For example, many discussions in the working groups revolved around drilling down into and challenging points that had been topics of the tutorials.

We also saw, in the tutorials as well as the discussions and through the interest in our work that greeted us at The Conversation Canada and in reaction to the article we published there directly after the seminar [1], the extent to which the general public perceives the urgency and importance of the issues we discussed, and the extent to which they expect explanations, answers, and better technology from the research community. It is our responsibility and aim to continue to provide these.

We would like to close with conclusions for future work on interaction design. After the second COVID-19 winter and with political choices worldwide geared towards making us “live with COVID”, all of us find ourselves in a transition to working routines that will likely involve more online interactions than in the past. Some of us believe that online meetings are the future, some of us believe that we should also invest more time and energy into creating successful hybrid meetings, which present additional challenges. Others are more cautious and long for a return to “normality”. Most believe in the need to find a new balance, such as obviating extensive travel for one-day meetings but travelling for longer meetings. (These discussions were also had at our Dagstuhl Seminar.)

We believe in the virtues of online meetings, at the minimum because they promise more inclusiveness (e.g. of mobility-restricted or otherwise vulnerable people, individuals without access to travel funding, persons with care duties, etc.) and are environmentally more sustainable. We also agree that hybrid meetings are particularly challenging, in particular when it comes to forming working groups that involve both on-site and online participants. But through the experience of this seminar, we also understood better that even so-called “fully online” meetings are in fact hybrid – in the sense that every participant remains engulfed in their home-office and/or regular-place-of-work spatial spheres. Creating a “sense of place” that is psychologically, physiologically and socially sustainable, will be a necessary element of future meetings, whether they are “fully online”, “hybrid”, or “on-site” in standard terminology. The sometimes-used term “offline” may actually be a key element for disconnecting from an overload of simultaneous duties. The simple and often ad-hoc uses of existing technology towards these ends that we have described in this text may have worked well in this particular setting, and others will be needed for different settings. But we hope that our lessons learned can serve as an inspiration to others on this trajectory.

Acknowledgements

This work is supported by the MASTER project that has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie-Slodowska Curie grant agreement N. 777695. Bettina Berendt thanks the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113f for funding for this seminar.

References

1 Geoffrey Rockwell, Bettina Berendt, Florence M. Chee, Jeanna Matthews, Sébastien Gambs, and Chiara Renso. Ottawa's use of our location data raises big surveillance and privacy concerns. *The Conversation*, 2022, Jan 27. https://theconversation.com/ottawas-use-of-our-location-data-raises-big-surveillance-and-privacy-concerns-175316.

Participants

- Darren Abramson
Dalhousie University, CA
- Christine Ahrend
TU Berlin, DE
- Bettina Berendt
TU Berlin, DE
- Florence Chee
Loyola University Chicago, US
- Thiery Chevallier
Akka Technologies, FR
- Maria Luisa Damiani
University of Milan, IT
- Josep Domingo-Ferrer
Universitat Rovira i Virgili – Tarragona, ES
- José Antônio Fernandes de Macedo
Universidade Federal do Ceara – Brazil, BR
- Sébastien Gambs
University of Montreal, CA
- Ioannis Kontopoulos
Harokopio University – Athens, GR
- Peter Kraus
European Data Protection Board – Brussels, BE
- Fen Lin
City University – Hong Kong, HK
- Jeanna Matthews
Clarkson University – Potsdam, US
- Stan Matwin
Dalhousie University – Halifax, CA
- Fran Meissner
University of Twente, NL
- Anna Monreale
University of Pisa, IT
- Francesca Pratesi
ISTI-CNR – Pisa, IT
- Alessandra Raffaetà
University of Venice, IT
- Chiara Renso
ISTI-CNR – Pisa, IT
- Paula Reyero-Lobo
The Open University – Milton Keynes, GB
- Geoffrey Rockwell
University of Alberta – Edmonton, CA
- Yannis Theodoridis
University of Piraeus, GR
- Konstantinos Tserpes
Harokopio University – Athens, GR
- Karine Zeitouni
University of Versailles, FR

Report from Dagstuhl Seminar 22031

Bringing Graph Databases and Network Visualization Together

Karsten Klein*[1], Juan F. Sequeda*[2], Hsiang-Yun Wu*[3], and Da Yan*[4]

1 **Universität Konstanz, DE.** karsten.klein@uni-konstanz.de
2 **data.world – Austin, US.** juan@data.world
3 **FH – St. Pölten, AT.** hsiang-yun.wu@fhstp.ac.at
4 **The University of Alabama – Birmingham, US.** yanda@uab.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22031 "Bringing Graph Databases and Network Visualization Together". Due to the ongoing restrictions caused by the COVID-19 pandemic, this purely on-site seminar had a reduced number of participants. Twenty-two researchers and practitioners from the Network Visualization and Graph Database communities met to initiate collaborative work and exchange between the two communities. The seminar served to establish a common understanding of the state of the art and the terminology in both communities, and to connect participants to tackle joint research challenges. Survey talks on the first days laid the foundations for subsequent plenary discussions and working groups. Further lightining talks during the next days gave more detailed insight into specific research questions and practical challenges. The contributions of the seminar include bringing the communities together, the identification of the top areas of research interest, and the characterization of research challenges and research questions. As an outcome, a position paper is planned, and further collaborations and joint publications are on the way.

Seminar January 16–21, 2022 – http://www.dagstuhl.de/22031
2012 ACM Subject Classification Information systems → Data management systems; Human-centered computing; Human-centered computing → Visualization
Keywords and phrases graph databases, network visualization, visual analytics
Digital Object Identifier 10.4230/DagRep.12.1.67

1 Executive Summary

Karsten Klein (Universität Konstanz, DE)
Juan F. Sequeda (data.world – Austin, US)
Hsiang-Yun Wu (FH – St. Pölten, AT)
Da Yan (The University of Alabama – Birmingham, US)

Network analytics through interactive network visualization has been essential in many research and application areas, such as bioinformatics, biomedicine, cyber security, e-commerce, social science, and software engineering. A network is often supported by graph databases with advanced query engines and indexing techniques. Graph databases have substantial contributions by academia and gained strong momentum in the industry, where the focus is on scalable systems using graph query languages that require to be learned by users.

* Editor / Organizer

Even though the Graph Database and Network Visualization communities study the same object, a graph/network, albeit from different perspectives, they do not communicate with each other. By bringing both communities together, we aimed to initiate and foster mutual communication and joint work. The goal of this seminar was to initiate collaborative efforts, to increase the mutual awareness of each others' existing concepts and technologies, and to identify new and complementary research challenges that lead to novel scientific outcomes. We have developed the schedule for the seminar based on our experience from previous successful Dagstuhl Seminars with a balance between prepared talks, plenary discussions, and breakout groups for less structured discussions focused on a selection of highly relevant topics.

The organizers envisioned several core topics for discussion at the Dagstuhl Seminar, as outlined in the proposal:

- Integration of fundamental concepts used in the two communities
- Visual scalability and computational performance
- Visual graph query paradigm
- Responsive visualization of graph query results
- (Qualitative) Evaluation
- Domain-oriented applications

During the plenary discussions on the first day, the participants identified several more specific topics for the work in separate working groups, which lead to the following working group titles:

- Evaluation and Usefulness
- Understanding gaps and opportunities between Graph Databases and Network Visualization
- Visual querying and result visualization

Our aim was to have focused discussions on these topics in which we would be able to make significant progress during the seminar, in order to shape a position paper and to lay the foundations for subsequent collaborations. The discussions showed that there indeed is the need for a closer exchange between the communities in order to improve the mutual understanding and practical solutions, but also to identify research questions that can be tackled jointly. They however also showed the great potential in this exchange and the large interest in both communities for joint work.

We have organized the seminar during the COVID-19 pandemic. Due to various regulations and travel restrictions, only roughly half of the usual number of participants attended in person. The meeting was held purely on-site, with the exception of one participant connecting in via video conferencing. We thank Dagstuhl for equipping the seminar rooms with suitable infrastructure and for putting suitable health and safety regulations in place to create a smooth experience and safe environment for all participants.

Acknowledgments

We would like to thank all participants of the seminar for their contributions and lively discussions, and the scientific directorate of Dagstuhl for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the amazing support by the staff of Dagstuhl.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Lightning Talk on GraphPolaris Visual Graph Analytics System and Research Challenges

Michael Behrisch (Utrecht University, NL & GraphPolaris.com Analytics Platform)

GraphPolaris is a no-code analytics platform for graph analysis. It enables non-data scientists to analyze large and complex datasets without the typically required query scripting and allows exposing the gathered analytical insights directly through effective visualizations. Our inductive exploration workflow engages the user into a visual data analysis dialog, in which an intuitive drag and drop user interface guides the construction of complex analytical queries visually and expressive visualizations give access to on-the-fly result interpretations, thus making graph databases and graph analytics accessible for a wide commercial and research audience.

In this talk, I demonstrated, on the one hand, the current approach towards developing a visual graph analytics platform in GraphPolaris and, on the other hand, contrasted it with the current research challenges.

3.2 Visual Graph Query and Analysis for Tax Evasion Discovery

Walter Didimo (University of Perugia, IT)

We briefly report on a 3-years collaboration with the Italian Revenue Agency, about the design of a decision support system for tax evasion discovery. The system combines network visualization techniques with graph databases and exploits some key ingredients: 1) An intuitive and powerful visual language that allows analysts to define suspicious patterns related to fraudulent schemes; the language also handles temporal information and does not require the knowledge of native GDBMS query languages. 2) A graph pattern matching engine, built on top of the Neo4J graph database, which efficiently retrieves and ranks subgraphs from a large network of taxpayers, based on the previously defined schemes. 3) An interactive environment which makes it possible to visualize the results returned by the graph pattern matching engine and to incrementally explore the network of taxpayers starting from them. Additionally, the system adopts artificial intelligence and information diffusion techniques to automatically assign each taxpayer a risk level based on both classical network centrality indices and on new centrality indices that estimate the level of involvement of a taxpayer in suspicious activities.

3.3 Overview of Graph Query Language Part 1 – Foundations

George Fletcher (TU Eindhoven, NL)

We gave an overview of graph data models with particular attention to RDF and Property Graphs. We then highlighted two basic ingredients used in the design of graph query languages: subgraph pattern matching and path querying. In the first class the languages of conjunctive queries and unions of conjunctive queries were formally defined and illustrated with examples. In the second class reachability, label constrained reachability, and regular path queries were defined and illustrated. We then defined and illustrated languages which combine both ingredients: unions of conjunctions of regular path queries and the regular queries. We concluded with an overview of the complexity of query evaluation and the complexity of query containment for each language.

3.4 Lightning Talk on Visualization and Query Optimization

Pavel Klinov (Stardog – Arlington, US)

The idea of the talk is that there might be an overlap between the graph visualization area and the graph query optimization area. Both graph visualization tools and query optimizers often require graph pre-processing, such as summarization, algorithms to figure out the structural properties of the graph. Visual rendering algorithms require it to highlight the most salient nodes and patterns in the graph while the optimizers use it to enable cardinality estimations for generating efficient query plans. It remains to be seen whether similar summarization or sampling algorithms can turn out useful for both kinds of tools.

3.5 Visualizing large graphs: a brief overview and some research experience

Fabrizio Montecchiani (University of Perugia, IT)

Visualization is a very popular and central task in graph processing pipelines. When the input is a large and complex network, computing an effective visualization is very challenging. The main steps involved in the creation of large-scale graph visualizations are usually simplification, layout, rendering, and interaction. We gave a brief overview of the methods and techniques used to address each of these steps. Particular attention has been paid to node-link layouts and to force-directed methods for computing such layouts. We also presented a vertex-centric multilevel force-directed algorithm to compute node-link layouts on a cloud computing platform.

3.6 Lightning Talk: Case Study: yFiles layout improvements for Graph Database Visualizations

Sebastian Müller (yWorks GmbH – Tübingen, DE)

The presentation was about what improvements were made in the graph drawing library "yFiles" in response to requirements by real world users working with graph database visualizations. With the help of conceptually simple layout techniques, the usefulness of visualizations for graph database exploration purposes was improved dramatically. By incorporating information about different semantic types of nodes and edges in the diagram as stored in the graph database, the algorithms were able to improve the arrangement in the diagram. Also, as an additional improvement, certain patterns that frequently appear in query results were specifically highlighted in the diagram, matching the users' expectation of the query results. For this, paths, stars, chains, and parallel structures were detected and treated especially by the layout algorithm. These improvements are available for various common layout algorithm implementations in yFiles.

3.7 What do Knowledge Graphs, Data Catalogs and Network Visualization have to do with each other?

Juan F. Sequeda (data.world – Austin, US)

Data catalogs are metadata and data management systems that inventory and organize data within an organization. Knowledge Graphs are a means of integrating data and knowledge at scale where concepts and relationships of a domain are manifested in the form of a graph. Thus a data catalog can be powered by a knowledge graph.

A question is how can Network Visualizations help accomplish common tasks in a data catalog such as business glossary definition, ontology engineering, impact analysis, root cause analysis, sensitive data impact. In this presentation, I highlight challenges and opportunities when attempting to integrate Network Visualization with a data catalog powered by a knowledge graph.

3.8 Lightning Talk on Neo4j Bloom

Hannes Voigt (Neo4j – Leipzig, DE)

We gave a brief ad-hoc presentation/demo of Neo4j Bloom, a domain-agnostic, low-code, ad-hoc graph visualization and exploration tool for data experts, data scientists, and data analysts.

3.9 Overview of Graph Query Language Part 2 — Practice

Hannes Voigt (Neo4j – Leipzig, DE)

We gave an overview of three graph query languages used in practice: SPARQL, Cypher, and GQL. The talk illustrated how the basic ingredients for graph query language discussed in Part 1 manifest in these graph query languages. This demonstrated that these languages – even if designed for different data models – are very similar in their capabilities. Still, we pointed out corners in which the languages differ slightly in their capabilities or put different emphases. Specifically for property graph query languages, we highlighted how they make use of the visual benefits of the ascii-art approach used in the graph pattern sublanguage and how it contributes to the adoption of these languages.

3.10 Qualitative Evaluation: Opportunities and Pitfalls

Tatiana von Landesberger (Universität Köln, DE)

This talk introduces evaluation studies in graph visualization. It focuses on qualitative evaluation with users. Evaluation studies with users are important when evaluating real value of application-motivated visualizations such as medicine, biology, finance. The visualizations need to be of high quality both from perceptual side and need to fit to the user's task and experience. Based on author's experience, and related literature in conducting evaluations of visualizations for real applications, the talk presents main steps, guidelines and pitfalls in conducting the studies. The talk discusses how the choice of tasks influence the evaluation, the pros/cons of methodological choices such as think-aloud protocols, which measures should be used for the evaluation, what is the value of free feedback for the evaluation. Finally, the talk presents how pitfalls from ill-posed evaluation questions can be turned into interesting research questions.

References

1 Archambault, Daniel, Helen Purchase, and Tobias Hoßfeld. "Evaluation in the Crowd." Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481. Vol. 10264. Springer, 2015.

2 Purchase, Helen C. Experimental human-computer interaction: a practical guide with visual examples. Cambridge University Press, 2012.

3 Isenberg, Tobias, et al. "A systematic review on the practice of evaluating visualization." IEEE Transactions on Visualization and Computer Graphics 19.12 (2013): 2818-2827.

4 Sedlmair, Michael, Miriah Meyer, and Tamara Munzner. "Design study methodology: Reflections from the trenches and the stacks." IEEE transactions on visualization and computer graphics 18.12 (2012): 2431-2440.

5 Ballweg, Kathrin, et al. "Visual Similarity Perception of Directed Acyclic Graphs: A Study on Influencing Factors and Similarity Judgment Strategies." J. Graph Algorithms Appl. 22.3 (2018): 519-553.

6 Kochtchi, Artjom, T. von Landesberger, and Chris Biemann. "Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles." Computer Graphics Forum. Vol. 33. No. 3. 2014.

3.11 An Overview of Graph Analytics

Da Yan (The University of Alabama – Birmingham, US)

Main reference Da Yan, Guimu Guo, Md Mashiur Rahman Chowdhury, M. Tamer Özsu, Wei-Shinn Ku, John C. S. Lui: "G-thinker: A Distributed Framework for Mining Subgraphs in a Big Graph", in Proc. of the 36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020, pp. 1369–1380, IEEE, 2020.
URL https://doi.org/10.1109/ICDE48307.2020.00122

A typical data analytics workflow includes (1) data retrieval, which falls in the domain of Data Engineering for data maintenance and querying; and (2) data analytics, which falls in the domain of Data Analytics. This is of no exception when we consider graph data, where the relevant graph data can be first retrieved from a graph database using query languages such as SPARQL, Neo4j Cypher, and GQL; the obtained graph can then be analyzed using methods in data mining and knowledge discovery, machine learning, and data visualization. Oftentimes, such retrieval and analytics algorithms can be programmed using a graph-parallel programming framework where users only need to specify some important user-defined functions based on application needs, rather than learning and using existing query languages and analytics libraries.

This talk introduces two such programming framework paradigms: (1) think like a vertex (TLAV) which aims to output a value for each vertex, as presented by Google's Pregel; and (2) think like a task (TLAT) for mining subgraphs, as presented by G-thinker, which allows compute-intensive analytics to scale performance with the number of CPU cores in contrast to existing data-intensive analytics tools. This talk also reviews a list of graph analytics tasks: (T1) Graph Traversal for Node Labeling, (T2) Random Walks for Node Scoring/Embedding, (T3) Graph Neural Networks, (T4) Frequent Subgraph Mining, (T5) Dense Subgraph Mining, and (T6) Subgraph Matching. Among them, (T1)-(T3) can be addressed by TLAV systems, while (T4)-(T6) can be efficiently addressed by TLAT systems such as G-thinker and PrefixFPM both published in ICDE 2020.

4 Working groups

4.1 Working Group 1: Evaluation and Usefulness

Juan F. Sequeda (data.world – Austin, US), Walter Didimo (University of Perugia, IT), Nadezhda T. Doncheva (University of Copenhagen, DK), George Fletcher (TU Eindhoven, NL), Stephen G. Kobourov (University of Arizona – Tucson, US), Giuseppe Liotta (University of Perugia, IT), Catia Pesquita (University of Lisbon, PT), and Tatiana von Landesberger (Universität Köln, DE)

This group discussed evaluation and the notion of usefulness of network visualizations. We observe a socio-technical phenomenon when it comes to the wow factor vs usefulness of network visualizations. Our discussions led us to the following realizations:

- We need to understand the diversity of roles around the work of network data management and network visualization.

- Bridging the gap between network visualization and graph database communities should be a win-win for both communities. Network visualization should support graph databases. Graph databases should support network visualization
- It is important to study how much network visualization can help in each role.
- There are specific interactions between the roles, which are not fully understood. How can network visualization help for those interactions?
- The network visualization space needs to be more fully studied, in the light of evaluation and usefulness.

The main topics we discussed were the following:

What is special about graphs/networks? Graphs are "natural" as they often capture the underlying model/problem well: If you need to model objects and relationships between them. A graph captures this with vertices and edges. Furthermore, a node-link diagram is a "natural" representation of a graph because the objects are the nodes, the relationships are the edges/links. There are many variations (edges can be directed/undirected, nodes can have attributes). The takeaway is that the data model, the visualization, and the interaction and meta-modeling paradigms are all the same: a graph. This is what makes graphs/networks special.

On the varieties of "Usefulness". What does it mean for a network visualization to be useful? Let's start with an example for the use case of ontology matching in life science. In the absence of context, we might infer that two references to the term "Gum" might match with high confidence. However, given the context of a dental ontology (network) visualization, we might note that one reference is in the context of the human mouth while the other is in the context of a type of candy, arising during different tasks by different researchers. This greatly lowers the subject matter expert's confidence in the match. This simple example illustrates that the usefulness of visualizations is very multifaceted.

We discussed and itemized several of these facets: the phases of usefulness and the lifecycle of network visualization (from engagement to final outcome); each of access for experts versus non-experts; aesthetics and enjoyability of network visualizations; trustworthiness and interpretability; small versus big data; loose versus highly structured data; data versus metadata; human context and human factors (e.g., profession); cost of interaction.

Roles in graph data and visualization. The final major topic discussed was that of roles around network data work. We identified some major roles, as a starting point to studying their interaction and how these interactions can be better supported by network visualization solutions.

- Domain Practitioner: Has expertise in a specific domain and has a question that needs to be answered. This role is the ultimate motivator (has the $), e.g., Doctor, CEO, Journalist.
- Analyst: Translates the question from the domain to the other roles. Probably the role responsible to provide the answer (or some means to get the answer such as data, visualization, etc) to the Domain Practitioner, e.g., Business Analyst, Data Scientist.
- Knowledge Scientist: Gather knowledge from the domain in order to understand what data should be used in order to answer the question.
- Visualization Scientist: Knows the space of visualization paradigms/metaphors and is able to suggest the best way of visualizing a certain type of data.
- Visualization Engineer: Is in charge of integrating or implementing visualization solutions and algorithms.

- Graph Database (GDB) Admin: Is responsible for the graph database, making sure the infrastructure is up and running, and provides access to the data.
- GDB Engineer: Is responsible to develop the graph database system itself; this role is typically filled in the context of a graph database vendor.
- Data Engineer: Is responsible for bringing in the data, integrating the data following requirements and loading them into a graph database.
- Data Steward: Is responsible for input data sources.
- Data Governance: Is responsible for understanding what data exists, how it is used and that it satisfies organizational requirements (regulations, policies, security, legal)

Some remarks. First, this list is not exhaustive. Second, one person can assume multiple roles, and multiple people can assume one role. Finally, we do not use the word "user" as it is ambiguous/confusing and also has negative connotations in some domains.

The following two additional roles are considered, but they are separated from the others, because they are studying the phenomena that occurs with the previous roles (their are not part of the evaluation process):

- Visualization Researcher: Provides innovation in the visualization field, by studying and experimenting new visualization metaphors/paradigms, layout algorithms, proof-of-concepts, prototype systems, quantitative and qualitative evaluations
- GDB Researcher: Provides innovation in the graph database field, by studying and experimenting new languages, data structures, efficient algorithms for graph queries.

These roles are fulfilled by professors, PhD students, industry researchers, etc.

Challenges and Opportunities. A major topic for further study is to deeply understand the interactions between roles and how we can better support these interactions with network visualization and graph data management solutions. A further general challenge is to study the different roles of visualization in the context of graph databases:

- Use visualization as a communication media between the different actors of the data production chain.
- Use visualization to design user-centric applications for the domain practitioners who need to explore the data and elaborate new information.
- Graph Layout Recommendation based on role, task, data, etc.

Within each of these challenges lies the deeper study of the definition(s) of usefulness, how we might quantify these definitions, and use these metrics for better evaluation of more effective network visualization and graph data management methods.

4.2 Working Group 2: Understanding Gaps and Opportunities Between Graph Databases and Network Visualisation

Karsten Klein (Universität Konstanz, DE), Henry Ehlers (TU Wien, AT), Oliver Kohlbacher (Universität Tübingen, DE), Sebastian Müller (yWorks GmbH – Tübingen, DE), Falk Schreiber (Universität Konstanz, DE), Hannes Voigt (Neo4j – Leipzig, DE), and Markus Wallinger (TU Wien, AT)

Originally constituted as a group to work on use cases and applications, our initial discussions quickly showed that before we could talk about concrete use cases, we needed to lay foundations for our common understanding of the potential interplay between graph databases

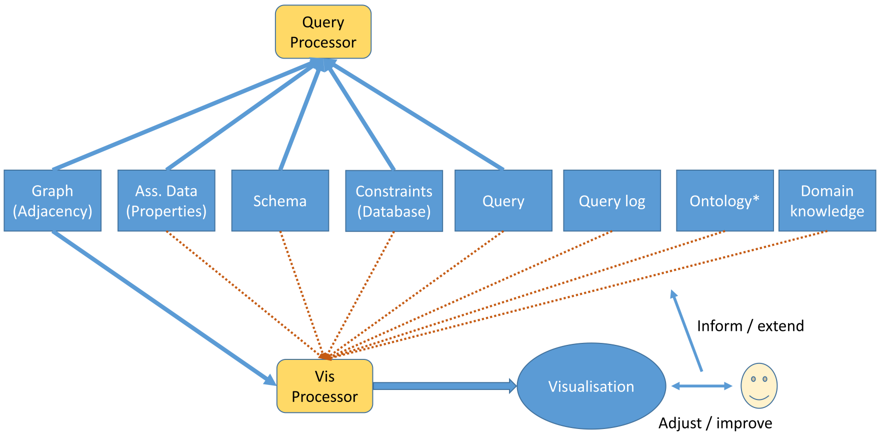


Figure 1 A conceptual diagram of the interplay between data assets, processing entities, and users in interactive graph database visualisation. It shows which sources of information could be used for query and visualisation processor, and how the user would interact with the system. Red arrows indicate currently unused but available information.

and network visualisation. While large potential for synergy and cross-pollination between the two areas is quite evident, we began by structuring the involved aspects and expectations in order to derive a conceptual framework based on which we could better identify gaps and opportunities of this interplay.

Our discussions quickly converged to a network visualisation in which we tried to cover the interplay between data assets and involved stakeholders and entities when graph visualisation is used in interactive interfaces for graph databases. Figure 1 shows the intermediate result that we came up with and which we used for the further discussions. While probably omitting some relevant aspects, it helped us to structure the discussion and to create a common mental map of the interactive graph database visualisation process.

With this conceptual model, we could now analyze the role as well as the aims of the user. The conceptual model then reveals requirements as well as the potential impact of graph visualisations, which in turn led us to a first characterisation of challenges and opportunities. As main gaps and challenges we identified

- A lack of appropriate graph visualisation and navigation methods that are tailored towards users of graph databases
- A lack of methods for projection in graph database systems (i.e., mapping from data assets to graph visualisation), incl. aggregation / abstraction of the graph structure
- Missing integration of concepts from graph database and visualisation communities into one coherent concept
- A definition of "usefulness" for graph visualisation in the graph database context and measures to evaluate it, and in general evaluation of metaphors for specific data/use cases/tasks
- Accessibility of meta-data available in the database system for the visualisation tool, and a systematic understanding of visualisation patterns for meta-data integration
- A lack of annotations for useful domain knowledge about graph topology in database schemas

As opportunities we see both significant potential improvements in practice, in particular regarding data handling, user experience, and more direct interfaces, as well as space for methodological work and new applications for research. These opportunities might concern different stakeholders differently depending on their role – user, developer, product owner, vendor – and area – visualisation, graph databases:

- Development of new graph visualisation and interaction techniques tailored towards graph databases. These can target the layout, encoding, and navigation, but also abstractions, e.g. for overview visualisation and subgraph comparison.
- Availability of currently untapped data sources and use cases for research, as there are huge and diverse graph databases with context available.
- Users can start with an improved out-of-the-box visualisation and apply ready-made templates to common use-cases with subsequent incremental improvement.
- New queries can be automatically derived by interacting with the visualisation as a more intuitive interface than current solutions.
- Query result visualisation has a large potential for improvement (leveraging meta-data and additional data stored with and in the database).
- An efficiency dividend can be achieved through simplified analysis processes.
- Visualisation system developers or vendors can increase the degree of automation and abstraction of their visualisation tool box and simplify its usage.
- Visualisation domain developers can produce domain-specific visualisations in shorter time due to an improved visualisation tool box.
- DB system developer/vendor can improve functional support of visualisation applications (by offering e.g. more powerful graph projection or additional graph schema annotation), and improve performance of query processing for visualisation applications (by leveraging extra knowledge about the application).

In order to structure our discussion on use cases, we made high-level distinctions of these with respect to the following questions: 1) Who tells the story – user or system? 2) What is the user's level of knowledge on the database content? 3) What is the goal – exploration, answering specific questions, or creating visualisations to tell a story?

Building on these discussions, we identified connections to the discussion topics in other groups, for example for the definition and evaluation of usefulness, the roles and types of audience involved, possible exploration patterns, as well as visual query support and corresponding visualisation metaphors. We plan to put our model to the test by creating instantiations of it for specific application use cases that we are familiar with, and to refine it according to the experience we gain in that process.

4.3 Working Group 3: Visual Querying and Result Visualization

Hsiang-Yun Wu (St. Pölten University of Applied Sciences, Austria, hsiang-yun.wu@fhstp.ac.at), Da Yan (The University of Alabama at Birmingham, USA, yanda@uab.edu), Michael Behrisch (Utrecht University, The Netherlands & GraphPolaris.com Analytics Platform, m.behrisch@uu.nl), Carsten Goerg (University of Colorado, USA, carsten.goerg@ucdenver.edu), Katja Hose (Aalborg University, Denmark, khose@cs.aau.dk), Pavel Klinov (Stardog, USA, pavel@stardog.com), and Fabrizio Montecchiani (University of Perugia, Italy, fabrizio.montecchiani@unipg.it)

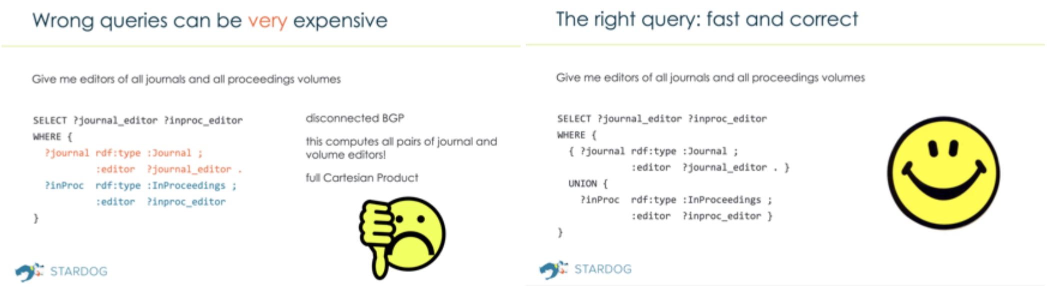
Working group 3 studied (1) the problem of **visual graph querying** which aims to assist user to formulate effective and efficient graph queries, and (2) the problem of **visualizing graph query results** for effective summarization, aggregation and human comprehension.

Visual Graph Querying (VGQ). VGQ is in contrast to the traditional graph database query languages such as SPARQL, Neo4j's Cypher, and GQL. Dagstuhl Seminar 22031 was fortunate to involve industrial attendees from Neo4j, Stardog, GraphPolaris, yWorks and data.world, which are startups and companies with graph database products already integrated with some simple frontend visualization tools. These participants have given in-depth demonstration of their products and use cases on Day 1 presentations. Working group 3 in particular has industrial members from Stardog and GraphPolaris, so in subsequent days we got a lot of chances reviewing concrete real-world use cases to see how graph database queries are applied, such as searching from enterprise knowledge graphs and biological networks.

During the discussion, members with less familiarity in the specific graph querying languages found it not easy to compose and even understand the traditional graph queries. The working group agreed upon the conclusion that learning the grammar of a graph querying language leads to a steep learning curve for end users, such as attendees with graph visualization background rather than graph database background. We expect that tools for formulating a graph query with drag-and-drop visual widgets would provide end users more intuition on what they are searching for, and it is also beneficial to support interactive and explorative query reformulation where users can learn from (at least partial) query results to incrementally revise their queries based on their search intents.

In fact, even our members from the industry admitted that their colleagues can formulate bad queries that accidentally create a large amount of unnecessary information. Figure 2 illustrates such an example to find the editors of all journals and conference proceedings from a backend graph database, where the SPARQL query on the left would unnecessarily lead to an expensive Cartesian product operation that can easily overwhelm computing and memory resources; the correct form of such a query is shown on the right which uses the UNION keyword to allow efficient execution. We expect that some visual widgets can better guide users to avoid formulating bad queries, such as giving a warning sign on excessive intermediate result size in the above SPARQL query example, or even to recommend an equivalent but more efficient query formulation. The implementation, however, would require techniques such as query cardinality estimation and seq2seq deep learning from curated (bad query, correct query) pairs captured in real enterprise operations.

The working group members identified several open challenges to address for effective VGQ. One challenge lies in how to define a set of visual querying paradigms that are effective in real applications. Several possible paradigms were discussed, including (i) pattern matching



Wrong queries can be very expensive

Give me editors of all journals and all proceedings volumes

```
SELECT ?journal_editor ?inproc_editor
WHERE {
  ?journal rdf:type :Journal ;
           :editor  ?journal_editor .
  ?inProc  rdf:type :InProceedings ;
           :editor  ?inproc_editor
}
```

disconnected BGP

this computes all pairs of journal and volume editors!

full Cartesian Product

(a) Expensive query

The right query: fast and correct

Give me editors of all journals and all proceedings volumes

```
SELECT ?journal_editor ?inproc_editor
WHERE {
  { ?journal rdf:type :Journal ;
             :editor  ?journal_editor . }
  UNION {
    ?inProc  rdf:type :InProceedings ;
             :editor  ?inproc_editor }
}
```

(b) Fast and correct query

Figure 2 An examples of a bad and a better queries.

as adopted by existing languages such as SPARQL, (ii) query-by-examples where end users list some desired results for the query engine to learn and recommend the possible queries and query semantics/intentions, (iii) query-by-sketch where end users sketch an incomplete graph query and rely on the learned data schema to auto-complete the actual query or to guide the formulation of the complete query. The group members agreed that effective data schema discovery techniques would be critical for implementing those VGQ paradigms. It is also an open problem to explore whether those query forms are sufficient for real user tasks, and how users can select among these VGQ paradigms. Novel query forms such as Cypher path matching could need to be invented to meet newly discovered querying demands as the field of VGQ progresses forward.

Graph Query Result Visualization. The output of a graph query can be of various forms such as many subgraph instances, or many path instances. Moreover, intermediate results before aggregation/reduction could be huge, requiring effective summative visualizations to make sense of the results that would be otherwise overwhelming to enumerate one by one.

A particularly interesting type of query is the path query as supported by Cypher, where end users specify a path pattern which is then matched against the backend graph database to find all matching path instances. The results are often numerous as indicated by our industrial members, and current systems usually enumerate individual path instances one after another leading to overwhelmingly many results to examine by end users. We envision that more optimized solutions can be easily developed, such as organizing the path instances (including partially matched ones) by tries so that common prefix paths can be shared to avoid redundancy. This method would not only speed up query evaluation, but also reduce the storage space requirement and the number of visual elements to display. Advanced visualization techniques can be integrated, such as making the nodes shared by more paths larger, and making the edges shared by more paths thicker. Of course, path queries are themselves relatively new, so the effectiveness of their result visualization approaches is yet to be verified in real applications.

Some other graph data are geospatial and/or topological in nature (e.g., nodes are associated with coordinates), which enable more effective visualization to bring intuition. Some effective visual representations already exist including radial layout, edge bundling and metro map metaphor, and they have been used in applications such as visualizing metro maps and metabolic pathways, but how to scale them to larger graphs effectively is still an open problem. Possible solutions include multi-scale result representation and hybrid visualization models such as NodeTrix (resp. ChordLink) which collapses dense fragments of a graph into matrices (resp. chord diagrams). See Figure 3 for an illustration.

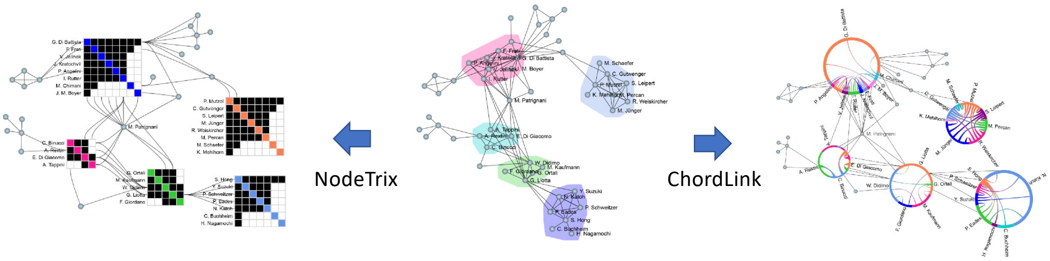
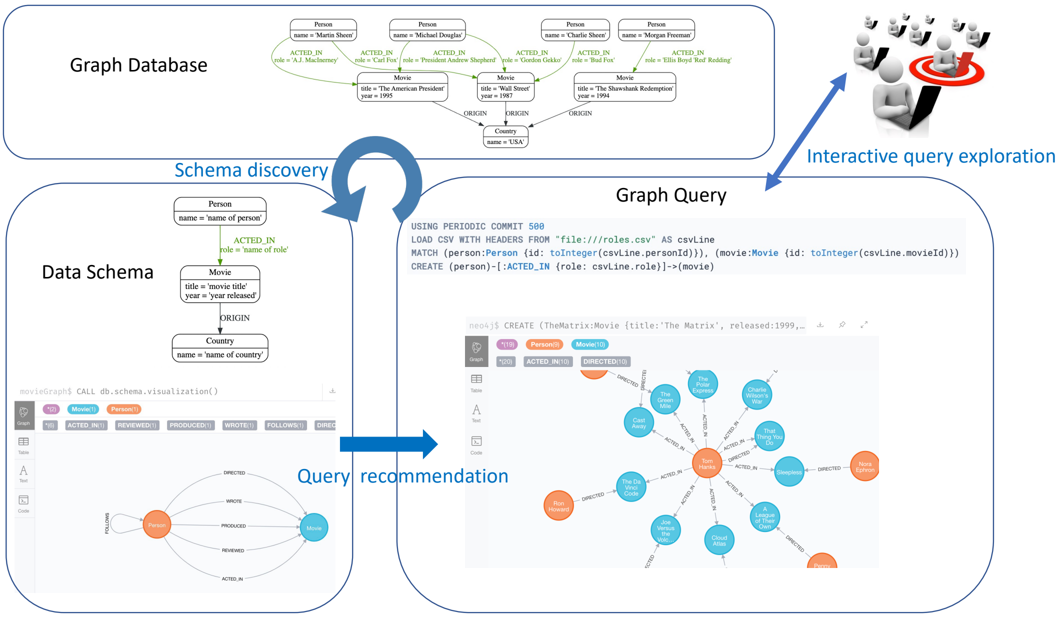


Figure 3 NodeTrix and ChordLink.



```
movieGraph$ CALL db.schema.visualization()
```

```
USING PERIODIC COMMIT 500
LOAD CSV WITH HEADERS FROM "file:///roles.csv" AS csvLine
MATCH (person:Person {id: toInteger(csvLine.personId)}), (movie:Movie {id: toInteger(csvLine.movieId)})
CREATE (person)-[:ACTED_IN {role: csvLine.role}]->(movie)
```

```
neo4j$ CREATE (TheMatrix:Movie {title:'The Matrix', released:1999,…
```

Figure 4 Integrating graph visualization and graph database.

Another interesting topic is to provide provenance explanations for graph query results, and some pioneering work has been conducted in the context of SPARQL, e.g., SPARQLprov published in PVLDB'21.

Summary: Integrating Graph Visualization and Graph Database. Figure 4 summarizes what we have discussed so far, where network visualization can be applied in the various stages of the graph querying pipeline, including data schema discovery, query result visualization and query reformulation recommendation. While a lot of those features have already been integrated into existing graph databases such as Neo4j, more diversified and advanced visualization techniques are yet to be implemented and integrated.

Participants

- Michael Behrisch
Utrecht University, NL &
GraphPolaris.com Analytics
Platform
- Walter Didimo
University of Perugia, IT
- Nadezhda T. Doncheva
University of Copenhagen, DK
- Henry Ehlers
TU Wien, AT
- George Fletcher
TU Eindhoven, NL
- Carsten Görg
University of Colorado –
Aurora, US
- Katja Hose
Aalborg University, DK
- Karsten Klein
Universität Konstanz, DE
- Pavel Klinov
Stardog – Arlington, US
- Stephen G. Kobourov
University of Arizona –
Tucson, US
- Oliver Kohlbacher
Universität Tübingen, DE
- Giuseppe Liotta
University of Perugia, IT
- Fabrizio Montecchiani
University of Perugia, IT
- Sebastian Müller
yWorks GmbH – Tübingen, DE
- Catia Pesquita
University of Lisbon, PT
- Falk Schreiber
Universität Konstanz, DE
- Juan F. Sequeda
data.world – Austin, US
- Hannes Voigt
Neo4j – Leipzig, DE
- Tatiana von Landesberger
Universität Köln, DE
- Markus Wallinger
TU Wien, AT
- Hsiang-Yun Wu
FH – St. Pölten, AT
- Da Yan
The University of Alabama –
Birmingham, US

Report from Dagstuhl Seminar 22042

Privacy Protection of Automated and Self-Driving Vehicles

Frank Kargl*[1], Ioannis Krontiris*[2], André Weimerskirch*[3], Ian Williams*[4], and Nataša Trkulja†[5]

1 **Universität Ulm – Ulm, DE.** frank.kargl@uni-ulm.de
2 **Huawei Technologies – München, DE.** ioannis.krontiris@huawei.com
3 **Lear Corporation and University of Michigan Transportation Research Institute – Ann Arbor, US.** aweimerskirch@lear.com
4 **University of Michigan – Ann Arbor, US.** ianwill@umich.edu
5 **Universität Ulm – Ulm, DE.** natasa.trkulja@uni-ulm.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22042 "Privacy Protection of Automated and Self-Driving Vehicles". The Seminar reviewed existing privacy-enhancing technologies, standards, tools, and frameworks for protecting personal information in the context of automated and self-driving vehicles (AVs). We specifically focused on where such existing techniques clash with requirements of an AV and its data processing and identified the major road blockers on the way to deployment of privacy protection in AVs from a legal, technical, business and ethical perspective. Therefore, the seminar took an interdisciplinary approach involving autonomous and connected driving, privacy protection, and legal data protection experts. This report summarizes the discussions and findings during the seminar, includes the abstracts of talks, and includes a report from the working groups.

Seminar January 23–28, 2022 – http://www.dagstuhl.de/22042

2012 ACM Subject Classification Security and privacy → Human and societal aspects of security and privacy; Security and privacy → Privacy protections; Security and privacy → Privacy-preserving protocols

Keywords and phrases automotive security and privacy, privacy and data protection

Digital Object Identifier 10.4230/DagRep.12.1.83

1 Executive Summary

Frank Kargl (Universität Ulm – Ulm, DE)
Ioannis Krontiris (Huawei Technologies – München, DE)
Nataša Trkulja (Universität Ulm – Ulm, DE)
André Weimerskirch (Lear Corporation – Ann Arbor, US)
Ian Williams (University of Michigan – Ann Arbor, US)

Cooperative, connected and automated mobility (CCAM) has the potential to drastically reduce accidents, travel time, and the environmental impact of road travel. To achieve their goals, connected and automated vehicles (AVs) require extensive data and machine learning algorithms for processing data received from local sensors, other cars, and road-side infrastructure. This immediately raises the question of privacy and data protection. While privacy for connected vehicles has been considered for many years, AV technology is still

* Editor / Organizer
† Editorial Assistant / Collector

in its infancy and the privacy and data protection aspects for AVs are not well addressed. The capabilities of AVs pose new challenges to privacy protection, given that AVs have large sensor arrays that collect data in public spaces. Additionally, AVs capture data not only from other vehicles, but also from many other parties (i.e. pedestrians walking along a street) with very limited possibilities to offer notice and choice about data processing policies. Additionally, the driver will not necessarily be the owner of the vehicle and it may be the case that the majority of AVs are owned by fleets.

Our seminar reviewed existing technologies, standards, tools, and frameworks for protecting personal information in CCAM, investigated where such existing techniques clash with the requirements of an AV and its data processing, and identified gaps and road-blockers that need to be addressed on the way to deployment of privacy protection in AVs from a legal, technical, and ethical perspective. While we ran only a shortened online version of the originally planned seminar due to COVID pandemic limitations, we made very good progress, in particular towards identifying and structuring the challenges. Future meetings will build on the results and will discuss the different challenges in more depth, prioritize the corresponding road blockers, and push for research to overcome them.

Discussions during the seminar were organized in seven sessions with presentations from renowned experts from industry and academia, and a final discussion that collected and structured outcomes. In the concluding session, we identified four main challenges that we present in this report alongside the talk abstracts.

- The first challenge is **ethics** and responsible behavior of companies and other actors that collect and process personal data in such systems. This goes beyond mere regulatory compliance but was seen as a promising path to complement this minimal baseline. Further discussions are required to identify ways to encourage such practices.
- Second, we discussed how **regulation** needs to evolve for future CCAM systems in order to establish a stable baseline. A challenge here will be to identify to what extent sector-specific regulation will be needed to address specifics of CCAM and if regulation of future systems is reasonable and possible.
- A third challenge is the **commercial** environment. Industry has to meet regulations and financial expectations and sometimes even conflicting goals like privacy and safety. Understanding and narrowing these trade-offs while acknowledging that industry has many such constraints that limit its flexibility requires further investigation.
- Last but not least, we see a strong progress in the privacy-enhancing **technology** (PET) as a promising path towards resolving many of the above mentioned problems. At the same time, many PETs have not been designed for the CCAM domain and might not meet its demands in data quality or latency. For this reason, we see the need to further investigate how existing PETs meet CCAM requirements or how they can be developed further to do so.

Generally speaking, there is a lack of incentives for enterprises like original equipment manufacturers (OEMs) to go beyond the legal minimum requirements to manage personal data in a privacy-respecting manner, to design privacy-preserving products, or to make the use of personal data transparent to the data subject. During our discussions one question became prominent: What could be the motivation for OEMs to do more in the field of data protection that goes beyond the bare minimum of legal compliance? Ethical and trustworthy aspects, as well as reputation and brand image could be worth investigating in answering this question. However, the field is massively interdisciplinary making it necessary to convince other involved disciplines of the value of data protection for the automotive sector.

There are several technical solutions available for protecting privacy and facilitating the privacy-by-design approach. However, the up-scaling of these solutions to larger systems and their integration with existing systems often fails because systems aspects and the related interdisciplinary issues are not taken into account. So, further progress is needed in promoting privacy-friendly system engineering, as well as integrating PETs into complete systems, taking into consideration the special requirements of safety and trust in the automotive domain. Overall, there should be a push for joint efforts to define and deploy technologies that are superior to today's solutions and that are commercially feasible since cost and effort are split amongst many participants.

Further progress is also required for the development of best practices, methodologies, and a requirements standard similar to ISO 21434 that supports the engineering of practical privacy solutions in complex systems. This will give OEMs a proper threshold target and allow for efficient solution finding and re-use. That guidance or standard could be a layer on top of regulation, similar to how the UN ECE R155 regulation requires a Cybersecurity Management System (CSMS) for which the ISO 21434 standard defines process requirements.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Introduction to Privacy Protection of Automated and Self-Driving Vehicles

Frank Kargl (Universität Ulm, DE)

This talk opened the seminar with an overview over the field of automotive privacy and how it developed over the years. We started from early works on Car-to-Everything (C2X) and discussed how privacy was considered an important requirement from day one. From this perspective, C2X is an excellent example of privacy-by-design and privacy-by-default. We introduced how changing pseudonyms were designed as a mechanism to protect privacy and prevent location tracking, also highlighting its limitations and the need to balance and trade-off technical privacy against effort and efficiency of applications. As an example, we looked into tracking attacks that can easily reconstruct a vehicle's path from anonymous position samples (if they are available with sufficiently high resolution).

We then continued with the observation that just providing technical solutions is not enough. We also need to consider the field of privacy engineering and the challenges that integrating privacy engineering with the overall engineering process implies. This includes a number of observations. First, system engineering for automotive E/E systems is already a highly complex process that needs to consider many aspects, like safety, real-time requirements, security, and now also privacy. Simplifying and streamlining these processes is important. Second, we need to educate experts for privacy engineering and privacy-enhancing technologies (PETs), while at the same time providing the tools and skills to ordinary developers to enable privacy-by-design in their projects. And third, as automotive industry builds products for world-wide markets, considering all the different privacy regulations is a special challenge.

We then looked into a case study that Prof. Kargl's research group conducted on privacy of electric vehicle charging in 2010 [1]. In the POPCORN protocol [2], they re-engineered the ISO/SAE 15118 protocol towards better technical privacy protection. This research highlighted that with proper privacy engineering, we can build systems that are highly privacy-preserving that at the same time provide rich and personalized functionality. In its conclusion, the talk discussed challenges and questions specific to autonomous vehicles and future cooperative intelligent transportation system (cITS) architectures including:

- Privacy of automotive AI and machine learning mechanisms
- Increased complexity and data flows in cooperative ITS and CCAM
- The conflict between privacy and functional and safety requirements, and whether PETs could help to resolve it
- Challenges to policy-making and law-making

Many of these challenges were in the focus of the talks and discussions throughout the remainder of the seminar week.

References

1 Fazouane, Marouane, Henning Kopp, Rens W. Heijden, Daniel Le Métayer, and Frank Kargl. "Formal verification of privacy properties in electric vehicle charging." In International Symposium on Engineering Secure Software and Systems, pp. 17-33. Springer, Cham, 2015.

2 Christina Höfer, Jonathan Petit, Robert Schmidt, and Frank Kargl. "POPCORN: privacy-preserving charging for eMobility." In Proceedings of the 2013 ACM workshop on Security, privacy & dependability for cyber vehicles, pp. 37-48. 2013.

3.2 The Role of Information in Autonomy and Community

Bryant Walker Smith (University of South Carolina, US)

Issues of privacy and data protection implicate the role of information in the larger goals of autonomy and community. These issues are exacerbated by the combination of increasingly powerful perception, processing, and transmission by advanced motor vehicles. In an analysis of these issues, some common distinctions among advanced vehicles matter, while others probably do not. This is evident in three threshold questions. First, (how) are mobile phones and other connected devices different than vehicles? Second, (how) are V2V-capable vehicles different than conventional vehicles? Third, (how) are automated vehicles different than conventional vehicles?

Connectivity and automation are orthogonal concepts. Connectivity in a narrow sense typically refers to the vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) communication that involves low latency, high reliability, direct(ish) transmission between road users. Connectivity in a broad sense refers to all manner of digital and wireless communication technologies and applications, including telematics, infotainment, over-the-air updates, remote assistance, mobile vehicle apps, and others. Automation typically involves distinctions between assisted driving and automated driving and among trips, vehicles, and features, but for the purposes here the distinction between safety and convenience features may be more significant.

Current and future technologies can monitor inside the vehicle, the vehicle itself, and outside the vehicle. This monitoring is accomplished with a large number and variety of internal-facing and external-facing sensors. Indeed, Congress recently directed NHTSA to soon mandate impaired driver detection systems in new vehicles. Data can be collected by these sensors for multiple overlapping purposes: To operate the system (implicit collection), to develop the system (implicit or intended), to document performance of the system (intended), or merely during operation of the system (incidental). It is also helpful to distinguish among data that are generated, processed, used, stored, and transmitted–on-board or off-board the vehicle.

This leads to three key questions: What can vehicle sensors perceive that humans cannot? What can computers do with that information that humans cannot? And what might companies, governments, and individuals do with these capabilities? This suggests immense possibilities both good and bad, for which canals, railroads, highways, and the Internet offer useful analogies from technologies of the past. Imagine a real-time version of Google's Street View or automated enforcement by private networks in which pedestrians who impede the movement of a company's vehicle are identified through facial recognition and then barred from that company's services as punishment.

This brings us to an antagonistic view of "privacy" as against a variety of actors, including domestic governments acting in their law enforcement or national security or administrative capacity or on behalf of companies, foreign governments acting in equivalent capacities, companies acting in a market, as employers, or on behalf of governments, individuals who are nosy, malicious, or controlling, and so forth.

The domain of safety offers lessons for privacy–since, indeed, both implicate autonomy and community. My argument is that automated vehicles are driven not by hardware or software or humans or nobody but by the companies that develop and deploy them. This is the position taken by the Uniform Law Commission's Automated Operation of Vehicles Act.

In other words, we should focus less on technologies and more on the companies behind those technologies. This means asking not "does the public trust this technology" but, rather, "is the company behind this technology worthy of our trust?" The former question is not useful because the public is fickle, words are not actions, marketing is still coming, and a lot can still change before we reach 100% adoption or acceptance (if ever). Similarly, asking "should the public trust this technology?" is not helpful because future technologies don't yet exist, these technologies will be diverse, and most won't be super dangerous. Indeed, the story of technology, society, law, and progress is about replacing one set of problems with a new set of problems and just hoping that the new set in aggregate is less than the old set–as illustrated by replacing the pollution of the horse with the pollution of the car.

"Is the company behind this technology worthy of our trust?" is an important question because companies act through their human and their machine agents, because technologies are only as safe (or privacy protecting) as the companies behind them, because safety (like privacy) is a marriage rather than a wedding, and because companies can do right even after their technologies fail. Similarly, regulation (whether of safety or privacy) needs the concept of trustworthiness because emerging technologies are complex, stochastic-ky, and dynamic; because developers have the expertise, information, and access to make lifecycle safety (or privacy) cases; because developers need space for technical innovation and regulators need space for regulatory innovation; and because even though regulators won't have all the answers, they can still ask better questions.

The question of how safe is safe enough has two answers: the retrospective (after a failure) and the prospective (before a deployment). Retrospective is more straightforward: a system must be at least as safe as a human in the maneuver and at least as safe as a comparative system, and (more controversially) safer than the last system to fail. Prospective safety is trickier, and my answer is that it means having reasonable confidence that the developer is worthy of our trust.

A trustworthy company, in turn, shares its safety philosophy (by saying what the company is doing, why it believes that to be reasonably safe, and why the public can believe the company), makes a promise to the public (by committing to market only what it believes to be safe, to being candid about its limits and failures, and by mitigating its failures), and then keeps that promise (by appropriately managing public expectations, supervising the entire product lifecycle, and mitigating harms promptly, fully, and publicly). The same approach might be applied to issues of privacy.

Looking for early breaches of trust is key to identifying untrustworthy actors. These breaches include making hyperbolic claims, misrepresenting evidence, failing to update technologies, exploiting the litigation process, and forcing confidential settlements. For example, if we cannot trust a company when it calls it system "full self driving," why should we trust that company when it calls its system safe or private? By analogy: Calling my umbrella a parachute doesn't make it true, but does make my umbrella more dangerous.

This approach also has implications for administrative regulation: Regulate the company rather than the technology; expect a company to vouch for its technologies through a public safety (or privacy) case; focus on processes and systems; identify assumptions and logical progressions; ask questions and challenge answers; and target breaches of public trust. In short (and to quote Voltaire or Spiderman): With great power comes great responsibility.

This detour into safety can also apply to privacy. In some ways, modern privacy law is something of a late 19th century invention. The foundational article, at least in the United States, may have been inspired by the fact that the paparazzi of the day used one technology–the camera–to take photos of a high-society wedding attended by one of the authors and then used another–the modern printing press–to distribute them widely in newspapers. But most people of the day did not enjoy privacy: They shared beds in tenements or other small homes, and they lived in communities of hundreds or thousands in full view of what Jane Jacobs approvingly called "eyes on the street." At the same time, they did have a form of escape that is largely unavailable today: They could physically run, or be chased, away from gossip and rumors and reputations. They could physically move to a new place without the kind of digital trail that would likely follow each of us today.

Privacy has since become a key aspect of the modern human rights doctrine. It can be embraced (or rejected) in ways that are revolutionary or evolutionary. It can be an end in itself or a means to an end. In particular, is may serve as a tool toward the goal of autonomy, where autonomy means the freedom to discover oneself, to be true to oneself, and to live one's own life. At the same time, society necessitates community, and every unit of governance other than the individual is a collective: governments, companies, religious institutions, families. And so both autonomy and community are part of happiness in the classic sense of leading a good life–that is, eudaimonia.

One of the key policy choices, to be determined by society much more than it is dictated by any existing law, is who or what will be empowered: individuals, governments, companies, or other collectives. Consider, very roughly, data protection. The approach in the EU has been to empower individuals through legal rights created by the General Data Protection Regulation (GDPR). The approach in the United States, at least prior to some recent state laws, has been to empower companies by generally enforcing the contracts of adhesion that we have all accepted in order to use many of the products and services essential to our modern lives. And the approach in the People's Republic of China has been to empower government in the storage and access expectations for companies operating in that country.

This is a crude model, and like everything else it is full of contradictions and unexpected consequences. For example, in the early 1900s, the US entered the "Lochner era" that empowered big business in the name of individual rights: Because the freedom to contract was paramount (indeed, the courts said, constitutionally protected), states could not enact rules for minimum wages or maximum hours worked. There are still echoes of this in so-called "right to work" states that restrict the power of unions. And in the European Union, there is concern that the GDPR could ultimately empower large companies vis-à-vis their smaller counterparts and ultimately vis-à-vis consumers.

One of my research interests is on using technologies and other tools to appropriately empower collectives. A key: Who inside the community, who is outside, and who decides?

These issues can be considered on the strategic, tactical, and operational levels. And here too there is much tension! It may be that governments should not necessarily adopt the policies that people think they want. For example, a world full of single-occupant vehicle trips on big, wide, fast roads is also a world of obesity and isolation and massive sea-level rise. In fact, there are benefits to "frictions" in life – what some incorrectly call inefficiencies. Sharing space with strangers, for example, can produce empathy and maybe even friendship–although it could also lead to harassment and assault. But it may be that policy should seek to accomplish what people say they want at a strategic level if not necessarily at an operational level. People say they want communities with fresh air and active mobility, for example. One answer is to set, regularly review, and as necessary revisit public policy goals so that changes are happen through deliberation rather than by default.

This presents a challenging strategic and even ethical question: Should innovation be understood primarily as a technical solution (to accomplish that which otherwise is technically impossible) or primarily as a policy solution (to accomplish that which is already technically feasible but not politically feasible)? For example, the public may be far more accepting of privacy and other risks inherent in the status quo than of equivalent risks inherent in new technologies. Similarly, economic interests may be far more entrenched for established technologies than for those that are merely emergent. This is one of the ethical issues explored in the "Ethics of AI in Transport" book chapter in the 2020 Oxford Handbook of Ethics in AI.

3.3 Ethics, Privacy, and Autonomous Vehicles

Adam Henschke (University of Twente, NL)

Main reference Adam Henschke: "Ethics in an Age of Surveillance: Personal Information and Virtual Identities", Cambridge University Press, 2017.
URL http://dx.doi.org/10.1017/9781316417249
Main reference Adam Henschke: "Privacy, the Internet of Things and State Surveillance: Handling Personal Information within an Inhuman System", Moral Philosophy and Politics, Vol. 7(1), pp. 123–149, 2020.
URL http://dx.doi.org/doi:10.1515/mopp-2019-0056

In this talk, I covered a range of conceptual and ethical issues to do with privacy and autonomous vehicles. The talk started with a motivating problem, showing how potentially innocuous vehicle data can have social and political implications. The talk then gave an overview of the "swamp" of different ways that privacy can be conceptualised. When thinking of privacy in an interpersonal sense, it can be descriptive or normative. The talk then shows that we also need to think of privacy in a political sense, as the relationship between citizens and the state. The talk then suggested that technologies like autonomous vehicles may need us to add an international geopolitical sense of privacy, where what matters are the relationships between states. The talk next looked at how autonomous vehicles are different from traditional cars in terms of the capacity to aggregate, share, and use personal information. Finally, the talk offered a way out of the swamp of privacy concepts by suggesting that we ought to be concerned for privacy if and when autonomous vehicles gather, access, use, and/or communicate information that is revealing, powerful, or has a special meaning given a particular context.

3.4 CCAM Research: Experiences in Handling Data Protection Regulations @UULM-MRM

Michael Buchholz (Universität Ulm, DE)

Main reference Michael Buchholz, Johannes Muller, Martin Herrmann, Jan Strohbeck, Benjamin Volz, Matthias Maier, Jonas Paczia, Oliver Stein, Hubert Rehborn, Rudiger-Walter Henn: "Handling Occlusions in Automated Driving Using a Multiaccess Edge Computing Server-Based Environment Model From Infrastructure Sensors", IEEE Intelligent Transportation Systems Magazine, Vol. PP, pp. 2–16, 2021.
URL http://dx.doi.org/10.1109/MITS.2021.3089743

This talk starts with a short introduction of our cooperative, connected and automated mobility (CCAM) research at the Institute of Measurement, Control, and Microtechnology (MRM) at Ulm University (UULM). Besides several automated test vehicles with approval for public traffic, UULM-MRM operates a smart infrastructure installation at an intersection in Ulm-Lehr, a suburb of Ulm. Details can be found in this video: `https://www.youtube.com/watch?v=RFdIpi3buAg`.

For both, test vehicles and smart infrastructure, sensors such as cameras are used, which potentially capture personal data of (other) road users. The second part of this talk focuses on the experiences made with implementing data security concepts as required by GDPR and national laws, as well as the experiences that the research team had with the persons concerned, all from an engineering perspective. For UULM, the Data Protection Act of the State of Baden-Württemberg holds, whereas for companies, the federal law is applicable.

The talk highlights some specific experiences and solutions, such as storing video data in read-only mode to avoid logging of potential changes, or the use of pictograms and QR codes to inform the persons concerned. Additionally, it is reported that sharing the data with others, like companies, is also possible, e.g., by using subcontracting or joint controller contracts. One experience discussed in the talk is the fact that most people seem not to care about the cameras on the vehicles or in the infrastructure, since there have been almost no questions and objections from persons concerned. Finally, personal conclusions are drawn, which address also the requirement of a good communication between the responsible lawyers/data security officers and the engineers.

3.5 Privacy Challenges of Connected and Autonomous Vehicles

Benedikt Brecht (Volkswagen AG – Berlin, DE)

This talk is an opinion contributing to the discussion of the seminar rather than a scientific presentation. It focuses on vehicle data and the potential privacy issues that may arise when accessible via different means. It begins with showing and listing all known and specific privacy-relevant data that is available in modern cars [1]. The listing gives a first understanding as to why getting access to this data could be a privacy issue. Especially as it is relatively simple to access this data, given physical access to such a vehicle, the tools and the know-how that is available on the internet. The section ends with a discussion of potential reasons why this data is not better protected yet, and why this becomes a potential privacy problem when selling a car or parts of it. The next section is a digression to Car-to-Everything (C2X) communication. It lists which data is sent out when enabling

C2X in modern, European cars of a specific make. It also highlights the fact that this data is unencrypted as this is the only solution to solving the requirements of availability (even if not connected to the internet), as well as latency that is required by the safety functions; the overhead created by any key exchange for an effective encryption would not meet these requirements. This section ends with a discussion about whether legitimate interest following GDPR is a sufficient legal basis. The following section discusses why especially autonomous vehicles add massively to the privacy issue due to their extended sensors and the way data is collected in order to make machine learning work. The last section gives a glimpse to a developing effort of ETSI's Technical Committee (TC) Lawful Interception (LI) summarized in their Technical Report (TR) 103 854. They appear to be working on standardizing the access to in-vehicle data (e.g., live camera feeds, planning of routes, customer details etc.) for law enforcement agencies while linking their effort to the ongoing regulation effort of the European Union on Electronic Evidence (E-Evidence).

References

1 c't – *magazin für computertechnik* 2022, Heft 1, pages 20/21. Heise Medien GmbH & Co. KG, Hannover, GER, 2022

3.6 Technologies for Establishing and Managing Trust in CCAM

Thanassis Giannetsos (UBITECH Ltd. – Athens, GR)

Joint work of Thanassis Giannetsos, Ioannis Krontiris
Main reference Benjamin Larsen, Thanassis Giannetsos, Ioannis Krontiris, Kenneth A. Goldman: "Direct anonymous attestation on the road: efficient and privacy-preserving revocation in C-ITS", in Proc. of the WiSec '21: 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks, Abu Dhabi, United Arab Emirates, 28 June – 2 July, 2021, pp. 48–59, ACM, 2021.
URL http://dx.doi.org/10.1145/3448300.3467832

This talk focused on the security, privacy, and trustworthiness, which are key properties that need to be considered as we are moving towards the realization of the 5G C-V2X technology. In this context, converging all of these properties by assessing dynamic trust relationships and defining a trust model and a trust-reasoning framework is of paramount importance; based on which involved entities can establish trust for cooperatively executing safety-critical functions. This will enable both a) cybersecure data sharing between data sources in the cooperative, connected and automated mobility (CCAM) ecosystem that had no or insufficient pre-existing trust relationship, and b) outsource tasks to the MEC and the cloud in a trustworthy way. Beyond the needs of functional safety, trustworthiness management should be included in CCAM's security functionality solution for verifying trustworthiness of transmitting stations and infrastructure. Compounding this issue, new schemes need to be designed that build upon and expand the Zero Trust concept: how to bootstrap vertical trust from the application, the execution environment, and the device hardware from the vehicle up to MEC and cloud environments.

For the latter, a promising solution integrates the use of trusted computing technologies and attestation mechanisms to enable the establishment of such "strong" trust relationships. This includes the integration of TEE technologies that enable highly secure, trusted, and verifiable remote computing capabilities, which can offer guarantees and assurances for the establishment of trust through the required proofs/claims. Such proofs can provide verifiable evidence on their correctness and functional safety, from their trusted launch and configuration to the runtime attestation of both behavioral and low-level concrete execution properties.

On the privacy front, Direct Anonymous Attestation (DAA) offers a promising solution – to overcome the challenges of traditional Public Key Infrastructures (PKIs) – by shifting trust from the backend infrastructure to the edge vehicles. DAA is a cryptographic protocol designed primarily to enhance user privacy within the remote attestation process of computing platforms, which has been adopted by the Trusted Computing Group (TCG). Applying the DAA protocols for securing V2X communication results in the redundancy (and removal) of most of the PKI infrastructure entities, including the pseudonym certificate authority: vehicles can now create their own pseudonym certificates using an in-vehicle trusted computing component (TC), and DAA signatures are used to self-certify each such credential that is verifiable by all recipients. Furthermore, a DAA-based model supports a more efficient revocation of misbehaving vehicles that don't require the use of CRLs, therefore removing all of the computational and communication overhead that comes with it.

Based on the above, some final conclusions were reached as it pertains to the future of CCAM technologies: this new security paradigm is the key element for having certifiable, more agile levels of trustworthiness to automotive services and translates to long-term consumer confidence, which is a requirement for end-user adoption.

3.7 On Exploring the Use of Local Differential Privacy in ITS

Ines Ben Jemaa (IRT SystemX – Palaiseau, FR)

Joint work of Ines Ben Jemaa, Anis Bkakria, Cedric Adjih, Ala'a Almomani, Michael Wolf

In this talk, we focus on the vehicular kinematic data sharing with the edge servers scenarios. One of the main privacy threats of this data sharing is the ability of an attacker to reconstruct the vehicular trajectory and thus learn information about the user profile. Differential Privacy (DP) [1] seems to be a promising solution to protect such data. Compared to other privacy approaches based on anonymization, cryptography or obfuscation, DP offers strong theoretical guarantees of privacy by adding some noise on the data following a specific probabilistic distribution while keeping some desired utility. Thus, it is compatible with the low overhead and computation complexity requirements of embedded systems. While the centralized scheme of DP is based on the data noise addition by a trusted curator, the local scheme, which assumes an untrusted curator, seems more realistic and convenient for our data sharing model. In the local scheme, the noise addition operation is attributed to the end users (i.e. the data originators) before transmitting their data to the server. The Geo-indistinguishability [2] paradigm implements the local scheme of DP and provides interesting properties for location privacy protection. We focused on studying its feasibility in the context of continuous data sharing in the connected and automated driving context and realize that there are some remaining challenges that have to be addressed. One of these challenges is the periodic nature of data sharing which decreases the privacy level and raises the problem of privacy budget allocation. The privacy budget allocation strategy could also directly impact the trade-off between the privacy and the utility of the service. The second challenge is the location correlation risk created by continuous sharing, which is not solved by geo-indistinguishability designed originally for sporadic use.

References

1 C. Dwork. *Differential privacy*. Proceedings Of ICALP, volume 4052 of LNCS, pages 1–12. Springer, 2006

2 Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., Palamidessi. *Geo-indistinguishability: Differentialprivacyfor location-basedsystems*. Proceedings of the ACM Conferenceon Computer and Communications Security, 901–914.

3.8 Automotive Privacy – The Good, The Bad, The Ugly

Mario Hoffmann (Continental Teves – Frankfurt-Sossenheim, DE)

Joint work of Mario Hoffmann, Sarah Syed-Winkler

The future of automotive products and mobility services is digital, autonomous, and personalized. Specifically, modern vehicles with dozens of assist systems, sensors and actuators, and autonomous driving capabilities will become a huge data source for a plethora of new individually tailored mobility services by the end of this decade. In this data industry, modern vehicles will become an important cornerstone for complex cross-domain scenarios with other road users, infrastructures, as well as mobile and back-end systems. One promising example are Smart Cities.

According to the "Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications" by the European Data Protection Board [1] these scenarios rely on the so-called Personally Identifiable Information (PII) and, therefore, are subject to special regulations and laws. One of the most prominent regulations is the European General Data Protection Regulation (EU GDPR, enforced May 2018). Here, a set of binding rules has been defined giving clear obligations for realizing – for instance – user consent, transparency, purpose binding, data minimization, data portability, and the right to be forgotten. Meanwhile, several regions around the globe took the EU GDPR as a blueprint for their own regulations.

The technical interpretation and implementation of these regulations in complex mobility scenarios, however, is a huge challenge. While Privacy Enhancing Technologies – such as Sticky Policies, Zero Knowledge Proofs, and Attribute Based Encryption, are well understood in the IT and Internet domain [2], there is still a technological gap in applying these techniques into the automotive domain. On the one hand, specifically, the PII life cycle in modern cars with its more than 100 Electronic Control Units (ECUs) lacks a consistent and interoperable data protection architecture which includes vehicles of different brands, mobile devices, infrastructures and back-end service environments. On the other hand, drivers and passengers, however, would like to be sure that the privacy settings defined in a car are enforced according to the regulation for each participant and stakeholder in any corresponding mobility scenario. Enjoying a particular personalized mobility service does not necessarily mean that the users' PII is free for any further usage.

In the publicly funded project "AUTOPSY – Automotive Data-Tainting for Privacy Assurance Systems", Continental investigates together with Fraunhofer AISEC and an equally small French consortium technical readiness and applicability of Privacy Enhancing Technologies for future automotive products and mobility services. From June 2021 to May 2024 the project will constantly update guidelines and learning material for the automotive domain regarding the impact of data protection regulations, applicable PETs for both embedded as well as end2end automotive architectures, and will demonstrate selected data tainting techniques. Future innovation projects need to investigate step by step more PETs in order to ensure cross-stakeholder inter-operability and policy enforcement as well as tackle all regulatory goals appropriately in order to fill the automotive privacy engineering toolbox.

References

1 European Data Protection Board, *Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications*. Jan 2020, https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf

2 ENISA, *Data Protection Engineering*. Jan 2022, https://www.enisa.europa.eu/publications/data-protection-engineering/@@download/fullReport

4 Working groups

4.1 Results of Concluding Discussions

Frank Kargl (Universität Ulm – Ulm, DE)
Ioannis Krontiris (Huawei Technologies – München, DE)
Nataša Trkulja (Universität Ulm – Ulm, DE)
André Weimerskirch (Lear Corporation – Ann Arbor, US)
Ian Williams (University of Michigan – Ann Arbor, US)

The seminar was organized in five 3-hour sessions with remote attendance due to the COVID pandemic. There was an introductory session and a concluding session, and the other time was spent on presentations and discussions around ethical, regulatory, legal, commercial, and technology aspects. The seminar sessions were loosely linked to those four topics with many cross-area considerations and discussions. The following summarizes the main results of the discussions.

4.1.1 Ethics

The ethical aspect considers the trustworthy and responsible use of personal data which is a matter of respect to other people including a company's customers. During discussions in the online seminar, we concluded that data protection regulation will not necessarily create ethical behavior but rather compliance. Compliance requires validation, which also creates cost and overhead. Furthermore, compliance does not rule out unethical behavior, e.g., today many terms of use and license agreements lure users to give their consent against their best interest. This observation led to a number of questions being raised and debated.

The working group discussed that stakeholders, in particular companies, might have an incentive to provide privacy protection that goes beyond minimal compliance and to establish a reputation of trustworthiness. There are examples available from the consumer electronics domain, even though only a few. For instance, the public opinion is that Apple is trustworthy in terms of privacy protection, but it is unclear if and how automotive stakeholders can follow Apple's role model.

It is also unclear what a company's incentive to act in a trustworthy way is and which regulatory steps can foster trustworthiness. Ideas include a regulatory requirement for transparency and demonstration of data protection practices. It is uncertain if a push for more trustworthiness as a concept replaces or extends data protection regulation. A concern raised is that perceived trustworthy behavior of a company might be created by clever marketing without the company providing actual trustworthiness. It is unclear how such practices can be stopped though. It appears this leads to a regulatory requirement for more transparency and for demonstrable privacy practices. At the same time, perceived trust and how technical safeguards to protect privacy are communicated to end-consumers is an important aspect that is built on trustworthiness and ethical behavior instead of strict compliance. We also discussed how this would evolve in a larger international setting. There is a strong overlap with the legal domain discussions presented in the next section.

The next discussion block was around the ethical goals of handling and processing personal data. It is unclear how to provide companies guidance on behaving ethically. We discussed that a reasonable expectation of privacy of a reasonable / average person could be a guiding principle where we would, e.g., ask whether such a person might expect to be unobserved by cameras when driving in an AV taxi.

We also identified that there is a lack of understanding of privacy expectations of people in traffic and mobility and this would require further research. For instance, it is necessary to understand under what circumstances end-consumers are willing to share information voluntarily for the greater or personal good, and what data they are willing to share. Experiences from some research projects show that people are usually willing to share data to contribute to research.

Overall, we think that this is an interesting concept to follow-up on and it would definitely help to explore the alternatives to the apparent trade-off between ever stricter regulation and abundant (ab-)use of personal data in cooperative, connected and automated mobility (CCAM) systems.

4.1.2 Regulation

The regulatory and the legal aspect of privacy are often termed data protection. Automated vehicles (AVs) face a number of legal and regulatory challenges that come from the complex intersection of data privacy law (often new and still developing) with a long-standing system of motor vehicle law and regulation. We discussed how we can ensure *compliance* and how we can encourage anything more than legally required compliance minimums. One train of thought is that heavy fines or significant potential legal liability are the only way to ensure even minimal compliance. An alternative argument is that potential reputational damage or the potential to use privacy protection as a marketing tool could be used to encourage compliance and even encourage privacy protection beyond minimum legal compliance. It is a challenge then for companies to prove compliance and for other stakeholders including end-consumers to check for compliance. Much of this discussion overlapped with a wider discussion on corporate trustworthiness, and what it means to be a "trustworthy" company.

A main question to answer is whether privacy issues around AVs (and connected vehicles) rise to a level that requires automotive-specific privacy laws and regulations, or whether privacy of AVs should be addressed through existing (or proposed) laws, regulation, and policy intended to govern privacy issues across a number of technologies.

The team discussed the effects of regulation on future technology and raised the concern that overly stringent privacy rules may lead manufacturers and developers to avoid new technologies or region-limit their deployments. Another challenge is how we communicate legal and regulatory requirements. In particular, how do we ensure that regulatory requirements are understood by engineers and developers in a way that assists the overall policy goal of a law/regulation, and how can we define threshold values around privacy for the average /

reasonable person, as this would also guide the courts in future disputes? The same holds for the communication with the end-consumers, e.g., how can we communicate a vehicle's level of privacy-friendliness or regulatory compliance to end-consumers similar to the idea of the NCAP 5-star safety rating.

A main challenge for industry are different regional data privacy regulations. For instance, the demands of the GDPR differ from those found in state privacy laws in the US, and from other countries' privacy regimes. Safety regulations also differ between countries and regions, so manufacturers may be more willing to deal with regional differences rather than try and apply standards more broadly. It is unclear if and how this will affect the deployment of AVs.

Finally, AVs could be used as a source of data for law enforcement – even as incidental surveillance (recording street activity as part of their regular activities). This might raise concerns both for industry stakeholders and end-consumers, and it is unclear how the stakeholders can balance privacy protection with cooperating with lawful data requests from law enforcement.

4.1.3 Commercial

The working group pointed out a few automotive specific challenges around privacy and self-driving vehicles that are described in the following. There seems to be a *trade-off between privacy and safety and efficiency* in that more detailed widely available data will enable more and better safety systems. In many applications and scenarios, it is unclear what this trade-off is though. It is also unclear who is supposed to decide about the trade-off. In fact, it is unclear if the automotive industry stakeholders should take the lead here or leave it to lawmakers to define regulation, and then the industry stakeholders only comply with said regulation. There are further trade-offs between privacy and traffic efficiency, pollution, and profits. There are also further trade-offs between physical safety such as emotional protection, pollution protection, etc. The same questions arise for these additional trade-off areas.

There are *commercial limitations* that limit the ability to go beyond the minimum legal requirements. OEMs and suppliers move in a highly regulated and competitive space, which limits flexibility of industry stakeholders. There might be other regulations that need to be considered for conflicts or contradiction in the context of privacy protection, such as right-to-repair law. Finally, it might be unclear who owns the data and hence how access to the data should be organized and controlled.

There are also challenges around the *technical solutions*. It is cumbersome to analyze each use-case/application and then design and implement a custom privacy solution. Maintenance and extensions are also rather cumbersome since each use-case that touches on the existing applications might alter the picture and require additional privacy solutions, and so do advances in privacy research and new attacks. Therefore, it appears there is a need for a framework, comprehensive technical guidance, and/or a standard to design, implement and audit privacy.

Companies might be able to utilize privacy for a *competitive advantage*. There are companies that implement privacy protection beyond legal compliance and that utilize their effort for a competitive advantage. It is unclear whether such an approach would be successful in the automotive sector, especially since the margins in the automotive industry are rather small compared to consumer electronics. Therefore, it is rather unclear whether any automotive stakeholder is willing to take the lead on privacy protection, especially since many tech companies don't appear to approach the topic beyond legal compliance.

4.1.4 Technology

The privacy-enhancing technologies (PETs) can be used to provide fundamental data protection principles to an AV system, e.g., minimizing personal data use, maximizing data security, and empowering individuals. Examples of such PETs include differential privacy, homomorphic encryption, secure multiparty computation, etc. The group particularly discussed the case of differential privacy that guarantees privacy protection in the presence of arbitrary auxiliary information. Differential privacy has been adapted to the context of location-based services to personalize the information provided to a user. In the context of the AV system, we can apply differential privacy to vehicle location data (often termed geo-indistinguishability). Notably, the system can add noise to vehicle location data to obfuscate the actual position of the driver or passengers. We identified two trade-offs:

- AVs count on an unprecedented amount of data to make decisions and usually it is a strict requirement that this data is accurate in order to allow the implementation of safety-critical services. This creates a tension with the fact that data needs to be handled in a secure and privacy preserving manner.
- The cryptography behind PETs can be computationally expensive like, for example, in privacy preserving machine learning. On the other hand, safety critical applications have such strict time constraints that makes it impossible to execute several cryptographic operations within these constraints.

A challenge is how to converge privacy protection with safety, based on the strict requirements of computational efficiency and time constraints. Emerging technologies that we can consider as solutions to overcoming this problem include Multi-access Edge Computing (MEC), which brings processing power near the vehicle to meet ultra-low latency requirements. With the help of MEC, massive computation and storage tasks need not be handled in the vehicle with its limited power and resources. Instead, these functionalities can be offloaded to the MEC which can handle it in a more cost-effective way in real-time. At the same time, 5G as the underlying communication paradigm can guarantee the strict service level agreements (bandwidth, zero latency, etc.).

Facilitating the secure and private collaboration between entities is a complex task, especially in the domain of CCAM that relies on cooperation and communication between vehicles and nodes. In this context, several entities that belong to different trust domains must interact with each other to exchange privacy sensitive data in order to enable safety-critical collaborative services. However, if these interactions are not properly managed, it can be the cause of privacy leaks. Therefore, there is a need to establish a high level of trust into received data and the functions that rely on this data. This in turn requires new trust assessment methods, in order to enable vehicles and nodes to assess the trust level of its neighbouring stations and received data and to take critical driving decisions.

A second challenge is how we can assess dynamic trust relationships and define appropriate trust models for involved entities. In this context the group investigated potential mechanisms. A promising approach is the employment of Trusted Computing and the enactment of remote attestation for producing verifiable claims on system properties and integrity.

Participants

- Ala'a Al-Momani
Universität Ulm – Ulm, DE
- Ines Ben Jemaa
IRT SystemX – Palaiseau, FR
- Benedikt Brecht
Volkswagen AG – Berlin, DE
- Michael Buchholz
Universität Ulm – Ulm, DE
- Thanassis Giannetsos
UBITECH Ltd. – Athens, GR
- Adam Henschke
University of Twente – Enschede, NL
- Mario Hoffmann
Continental Teves – Frankfurt-Sossenheim, DE
- Frank Kargl
Universität Ulm – Ulm, DE
- Alexander Kiening
Denso Automotive – Eching, DE
- Ioannis Krontiris
Huawei Technologies – München, DE
- Jason Millar
University of Ottawa – . Ottawa, CA
- Kyriaki Noussia
University of Reading – Reading, GB
- Christos Papadopoulos
University of Memphis – Memphis, US
- Jonathan Petit
Qualcomm – Boxborough, US
- Chrysi Sakellari
Toyota Motor Europe – Brussels, BE
- Yu Shang
Huawei Technologies – Shanghai, CN
- Lauren Smith
Cruise – Washington, US
- Nataša Trkulja
Universität Ulm – Ulm, DE
- Jessica Uguccioni
Law Commission of England and Wales – London, GB
- Bryant Walker Smith
University of South Carolina – Columbia, US
- André Weimerskirch
Lear Corporation – Ann Arbor, US
- Ian Williams
University of Michigan – Ann Arbor, US
- Harald Zwingelberg
ULD SH – Kiel, DE

Report from Dagstuhl Seminar 22051

Finite and Algorithmic Model Theory

Albert Atserias*[1], Christoph Berkholz*[2], Kousha Etessami*[3], and Joanna Ochremiak*[4]

1 **UPC Barcelona Tech, ES.** atserias@cs.upc.edu
2 **HU Berlin, DE.** berkholz@informatik.hu-berlin.de
3 **University of Edinburgh, GB.** kousha@ed.ac.uk
4 **University of Bordeaux, FR.** joanna.ochremiak@gmail.com

Abstract

Finite and algorithmic model theory (FAMT) studies the expressive power of logical languages on finite structures or, more generally, structures that can be finitely presented. These are the structures that serve as input to computation, and for this reason the study of FAMT is intimately connected with computer science. Over the last four decades, the subject has developed through a close interaction between theoretical computer science and related areas of mathematics, including logic and combinatorics. This report documents the program and the outcomes of Dagstuhl Seminar 22051 "Finite and Algorithmic Model Theory".

Seminar January 30–February 4, 2022 – http://www.dagstuhl.de/22051
2012 ACM Subject Classification Theory of computation → Finite Model Theory
Keywords and phrases automata and game theory, database theory, descriptive complexity, finite model theory, homomorphism counts, Query enumeration
Digital Object Identifier 10.4230/DagRep.12.1.101

1 Executive Summary

Albert Atserias (UPC Barcelona Tech, ES)
Christoph Berkholz (HU Berlin, DE)
Kousha Etessami (University of Edinburgh, GB)
Joanna Ochremiak (University of Bordeaux, FR)

Finite and Algorithmic Model Theory research revolves around the study of the expressive power of various logics on finite and finitely presented structures, and the connections between this and different computational models and mathematical formalisms.

The methods and tools of finite and algorithmic model theory (FAMT) have played an active role in the development of several areas of computer science. Finite or finitely representable structures are those that serve as inputs to computation, and the study of the expressive power of logical languages on such structures has led to fundamental insights in diverse areas, including database theory, computational complexity, random structures and combinatorics, verification, automata theory, proof complexity, and algorithmic game theory.

Over the past four decades FAMT has established itself as a rich research field with a strong and evolving community of researchers with a shared agenda. Much of the progress can be traced to the regular meetings of the community: the last such meeting was at Dagstuhl in 2017, and before that at Les Houches in 2012.

[*] Editor / Organizer

Finite and Algorithmic Model Theory, *Dagstuhl Reports*, Vol. 12, Issue 1, pp. 101–118
Editors: Albert Atserias, Christoph Berkholz, Kousha Etessami, and Joanna Ochremiak
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The principal goals of this seminar included:

1. To identify fresh challenges in FAMT arising from some of the main application areas as well as newly emerging ones.
2. To make connections between core research in FAMT and other subfields of theoretical computer science, such as the theory of combinatorial and continuous optimization algorithms, and the theory of homomorphism counts and limit structures.
3. To transfer knowledge from emerging techniques in core FAMT back to the connected subfields and application areas.
4. To strengthen the research community in FAMT, especially by integrating younger members into it.
5. To provide continuity for what has been a successful model of regular seminars for building and consolidating the productive research community in FAMT.

One of the main goals of this Dagstuhl Seminar was to capitalize on the progress and the potential impact of some of the latest developments in FAMT and related areas. Such developments include:

a) The recently established connections between symmetric models of classical computation and bounded-variable counting logics. The symmetric counterparts of classical models of computations include threshold circuits, linear and semidefinite programs, and algebraic circuits. These new results have already been used to establish new upper and lower bounds for large families of algorithms by FAMT tools.

b) The theories of homomorphism counts and limit structures in combinatorics. A recent trend of work establishes that distinguishing the structures by the number homomorphisms they admit from certain classes of patterns, or to certain classes of patterns, is a fruitful alternative to distinguishing them by the logic formulas they satisfy.

c) Enumeration and counting methods including their use in database query processing, among others. One of the goals of this line of research is to understand and classify the logical queries for which it is possible to compute a compact representation of the output from which the query results can be obtained, efficiently, and on demand.

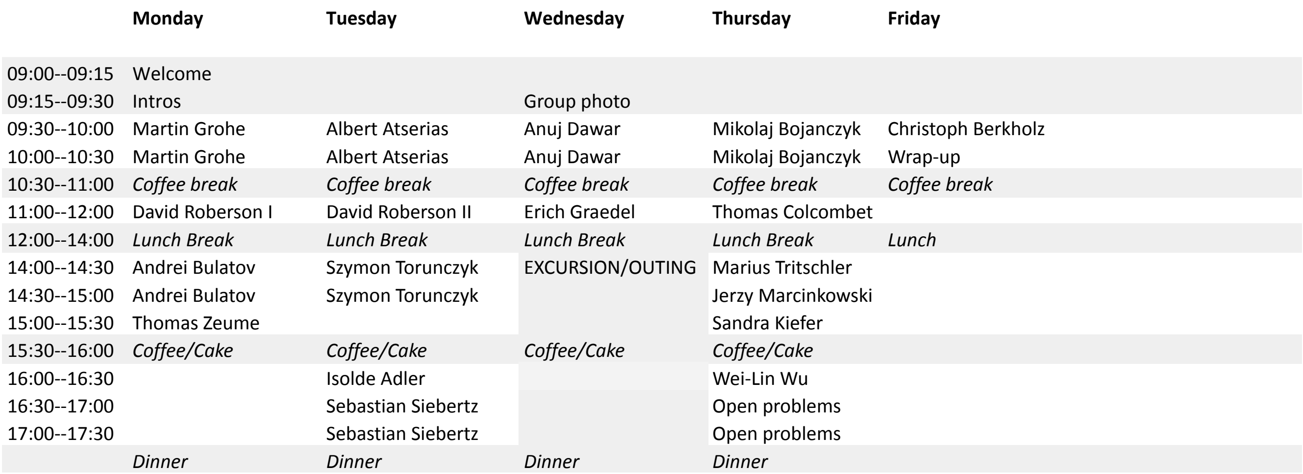
Organization and Activities

The organizers developed a schedule consisting of a number of invited survey talks, a number of talks focused on regular contributions proposed by participants, and an open problem session.

The seminar took place in person at Dagstuhl, with essentially all talks (except one) delivered by speakers attending Dagstuhl in person. The timing of the seminar coincided with the height of the Covid Omicron wave in Germany and Europe. This resulted in a number of late covid-related cancellations. Some talks, including invited talks, had to be cancelled, and the total number of participants (24) was fewer than originally planned.

Outcomes

Despite the many organizational challenges presented by the covid surge around the time of the workshop, the seminar was highly stimulating and surprisingly successful for those who were able to attend, and achieved many of our goals. (And thankfully there were no cases of covid during the seminar among those who did attend in person.)



	Monday	Tuesday	Wednesday	Thursday	Friday
09:00--09:15	Welcome				
09:15--09:30	Intros		Group photo		
09:30--10:00	Martin Grohe	Albert Atserias	Anuj Dawar	Mikolaj Bojanczyk	Christoph Berkholz
10:00--10:30	Martin Grohe	Albert Atserias	Anuj Dawar	Mikolaj Bojanczyk	Wrap-up
10:30--11:00	*Coffee break*	*Coffee break*	*Coffee break*	*Coffee break*	*Coffee break*
11:00--12:00	David Roberson I	David Roberson II	Erich Graedel	Thomas Colcombet	
12:00--14:00	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch*
14:00--14:30	Andrei Bulatov	Szymon Torunczyk	EXCURSION/OUTING	Marius Tritschler	
14:30--15:00	Andrei Bulatov	Szymon Torunczyk		Jerzy Marcinkowski	
15:00--15:30	Thomas Zeume			Sandra Kiefer	
15:30--16:00	*Coffee/Cake*	*Coffee/Cake*	*Coffee/Cake*	*Coffee/Cake*	
16:00--16:30		Isolde Adler		Wei-Lin Wu	
16:30--17:00		Sebastian Siebertz		Open problems	
17:00--17:30		Sebastian Siebertz		Open problems	
	Dinner	*Dinner*	*Dinner*	*Dinner*	

The final program included invited tutorial talks by Martin Grohe on homomorphism counts (delivered online, to an in person audience at the workshop, due to a last minute cancellation for Grohe caused by COVID), David Roberson I and II on quantum isomorphism and its connection to homomorphism counting. These and other related talks at the seminar highlighted the exciting ongoing work aimed at delineating the power of homorphism counting on various classes of graphs and structures, with surprising connections to other areas of mathematics.

Another invited talk was by Albert Atserias on symmetric computation and descriptive complexity (replacing a last minute cancellation), highlighted the exciting recent developments at the intersection of FAMT and combinatorial optimization. Another invited talk on query enumeration also had to be cancelled due to COVID. The full schedule including all the other talks of the Seminar can be found in the adjoined table.

One of the traditions of the series of workshops on Finite and Algorithmic Model Theory is to have a session in which some of the attendants present open problems and directions for further research. In this occasion, an hour of the afternoon of Thursday was devoted to such a session. A couple of days earlier we made a public call for presentations of open problems. Volunteers would write down their name on an easel pad. By Thursday, three volunteers came forward who gave 10 minute presentations (aprox.) of their proposals.

First, Erich Graedel presented an open problem on the topic of his earlier talk on Wednesday. Shortly put, the open problem asks to develop a proof theory for the emerging field of semiring semantics for logic formulas. The motivation comes from its potential applications in database theory and game analysis. Second, Isolde Adler presented an open problem on a topic covered in an earlier talk by Torunczyk. In brief, the open problem asks to study the relationships between the various notions of model-theoretic stability for classes of hypergraphs and more general relational structures. Third, Kousha Etessami gave a short talk on his recent work on applications of Tarski's Fixed-Point Theorem to economic game theory. In a nutshell, the question asks how much faster can one compute an arbitrary fixed-point of a monotone operator on a grid lattice than it takes to compute the least or greatest fixed-point by the standard iteration method. A detailed exposition of this open problem and its motivations can be found in a later section of this report. Finally, during his talk on Tuesday, Albert Atserias announced an open problem related to the optimal hardness of approximating the minimum vertex cover on graphs. Since it was presented earlier during a talk, this open problem was not presented during the session. A detailed description of this open problem appears also in a later section of this report.

Overall, the organizers regard the seminar as a resounding success despite the difficult circumstances, and judging by the very positive feedback from participants, they agreed. We look forward to the next meeting of the FAMT community, hopefully within a few years, whether at Dagstuhl or elsewhere.

The organizers are grateful to the Scientific Directorate of the Center for its support of this workshop and the staff of Schloss Dagstuhl for their organisation of our stay (including regular covid testing) and their hospitality, despite the many challenges posed by covid.

2 Table of Contents

3 Overview of Talks

3.1 First-order property testing

Isolde Adler (University of Leeds, GB)

Joint work of Isolde Adler, Polly Fahey, Frederik Harwath, Noleen Köhler, Pan Peng
Main reference Isolde Adler, Noleen Köhler, Pan Peng: "On Testability of First-Order Properties in Bounded-Degree Graphs", in Proc. of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 – 13, 2021, pp. 1578–1597, SIAM, 2021.
URL http://dx.doi.org/10.1137/1.9781611976465.96

Property testing (for a property P) asks for a given graph, whether it has property P, or is "structurally far" from having that property. A "testing algorithm" is a probabilistic algorithm that answers this question with high probability correctly, by only looking at small parts of the input. Testing algorithms are thought of as "extremely efficient", making them relevant in the context of large data sets.

In this talk I will present recent positive and negative results about testability of properties definable in first-order logic and monadic second-order logic on classes of bounded-degree graphs.

3.2 Symmetric computation, symmetric LP-lifts, and more...

Albert Atserias (UPC Barcelona Tech, ES)

Joint work of Albert Atserias, Anuj Dawar, Joanna Ochremiak
Main reference Albert Atserias, Anuj Dawar, Joanna Ochremiak: "On the Power of Symmetric Linear Programs", J. ACM, Vol. 68(4), pp. 26:1–26:35, 2021.
URL http://dx.doi.org/10.1145/3456297
Main reference Albert Atserias, Joanna Ochremiak: "Definable Ellipsoid Method, Sums-of-Squares Proofs, and the Isomorphism Problem", in Proc. of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018, Oxford, UK, July 09-12, 2018, pp. 66–75, ACM, 2018.
URL http://dx.doi.org/10.1145/3209108.3209186

Descriptive complexity can be thought of as addressing the same fundamental questions as computational complexity theory but in a different model of computation. Unlike Turing machines or Boolean circuits, that get their inputs as strings, the machines and circuits of this model get their inputs as unordered sets of tuples of atomic objects; e.g., graphs. The computation is then supposed to develop in a symmetry-preserving way. In the talk, we discuss the power of fixed-point logic with counting, FPC, and of linear programming lifts, LP-lifts, as models of symmetric polynomial-time computation. We also report on the status of the program that asks to settle some of the most relevant questions of complexity theory in this restricted model, and do so unconditionally.

3.3 Number of variables vs. formula size in existential-positive FO

Christoph Berkholz (HU Berlin, DE)

Joint work of Christoph Berkholz, Hubie Chen
Main reference Christoph Berkholz, Hubie Chen: "Compiling Existential Positive Queries to Bounded-Variable Fragments", in Proc. of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2019, Amsterdam, The Netherlands, June 30 – July 5, 2019, pp. 353–364, ACM, 2019.
URL http://dx.doi.org/10.1145/3294052.3319693

A crucial property of bounded-variable fragments of first-order logic is that they can be evaluated in polynomial time. It is therefore a useful preprocessing step to rewrite, if possible, a first-order query to a logically equivalent one with a minimum number of variables. However, it may occur that reducing the number of variables causes an increase in formula size. We investigate this trade-off for the existential-positive fragment of first-order queries, where variable minimisation is decidable in general. In particular, we study the blow-up in the formula size when compiling existential-positive queries to the bounded variable fragment of positive first-order logic. While the increase of the formula size is always at most exponential, we identify situations (based on the signature and the number of variables) where only a polynomial blow-up is needed. In all other cases, we show that an exponential lower bound on the formula size of the compiled formula that matches the general upper bound. This exponential lower bound is unconditional, and is the first unconditional lower bound for formula size with respect to the studied compilation; it is proved via establishing a novel interface with circuit complexity which may be of future interest.

3.4 Polyregular Functions

Mikolaj Bojanczyk (University of Warsaw, PL)

Transducers are like automata, but instead of accepting/rejecting they produce an output, such as a string or a tree. In my talk, I will discuss some recent results transducers, mainly about the class of polyregular transducers, which can be seen as a candidate for the notion of "regular" string-to-string transducers of polynomial growth. I will discuss how this class can be characterised in many different ways, including logic, automata, and λ-calculus. From a finite model theory point of view, the most interesting fact is that it makes sense to consider mso interpretations where positions in the output string are represented by k-tuples of positions in the input string. It turns out that, somewhat surprisingly, such functions are closed under composition if the inputs and outputs are strings.

References

1 M. Bojanczyk. "Polyregular functions". CoRR abs/1810.08760 , 2018.

3.5 Complexity classification of counting graph homomorphisms modulo a prime number

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

Joint work of Amirhossein Kazeminia, Andrei A. Bulatov
Main reference Amirhossein Kazeminia, Andrei A. Bulatov: "Counting Homomorphisms Modulo a Prime Number", in Proc. of the 44th International Symposium on Mathematical Foundations of Computer Science, MFCS 2019, August 26-30, 2019, Aachen, Germany, LIPIcs, Vol. 138, pp. 59:1–59:13, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.
URL http://dx.doi.org/10.4230/LIPIcs.MFCS.2019.59

Counting graph homomorphisms and its generalizations such as the Counting Constraint Satisfaction Problem (CSP), its variations, and counting problems in general have been intensively studied since the pioneering work of Valiant. While the complexity of exact counting of graph homomorphisms (Dyer and Greenhill, 2000) and the counting CSP (Bulatov, 2013, and Dyer and Richerby, 2013) is well understood, counting modulo some natural number has attracted considerable interest as well. In their 2015 paper Faben and Jerrum suggested a conjecture stating that counting homomorphisms to a fixed graph H modulo a prime number is hard whenever it is hard to count exactly, unless H has automorphisms of certain kind. In this paper we confirm this conjecture. As a part of this investigation we develop techniques that widen the spectrum of reductions available for modular counting and apply to the general CSP rather than being limited to graph homomorphisms.

3.6 On Uniformisation

Thomas Colcombet (CNRS – Paris, FR)

Joint work of Thomas Colcombet, Alex Rabinovich

In this talk, I will present several questions and results related to uniformisation questions for first-order and monadic second-order theory (MSO). Uniformisation consists in, given a formula $\phi(X,\bar{p})$, where $\bar{p}$ is a tuple of parameters, to find a formula $\psi(X,\bar{p})$ such that: (1) If $\psi(X,\bar{p})$ holds, then $\phi(X,\bar{p})$ holds, and (2) is $\phi(X,\bar{p})$ holds for some X, then there exists a unique X such that $psi(X,\bar{p})$ holds. in other words, if a formula has a solution, it is possible to define a unique solution. Of course these questions are parametrised by the logic and the family of models under consideration.

Question pertaining to uniformisation are various. Some logics may admit uniformisation on some models. Some logics may require a more expressive logic for uniformisation. And uniformisability can be seen as a decision procedure.

In this talk, I will survey some results of these forms on finite and infinite structure, for first-order or monadic second-order logic. In particular, I will explain how the algebraic approach can be used to answer some uniformisability questions for MSO logic over countable ordinals. More precisely, it was is known since 1998 by Shelah and Lifshes that MSO is uniformisable in itself over all ordinals shorter than ω^ω (non-uniformly), and that it is not true beyond. I will present recent unpublished new results in collaboration with Alex Rabinovich: (a) there is a single construct that can be added to MSO in order to uniformise MSO over all countable ordinals (the ability to choose a unique cofinal set of order-type omega); and (2) that it is possible to decide, given an MSO formula, whether it can be uniformised in MSO itself over all countable ordinals.

3.7 Cohomology and Finite Model Theory

Anuj Dawar (University of Cambridge, GB)

Joint work of Samson Abramsky, Adam Ó Conghaile, Anuj Dawar

In this talk I gave a brief introduction to algebraic topology intended for the finite model theory audience, leading up to a definition of cohomology. With this in hand, I discussed some potential applications of the ideas in finite model theory. In particular, we look at a collection of partial solutions to a constraint satisfaction problem as a presheaf and identify cohomological obstructions to satisfiability. This allows us to draw conclusions about the expressive power of fixed-point logic extended with operators for solving systems of linear Diophantine equations.

3.8 Homomorphism Tensors and Linear Equations

Martin Grohe (RWTH Aachen University, DE)

Joint work of Martin Grohe, Gaurav Rattan, Tim Seppelt
Main reference Martin Grohe, Gaurav Rattan, Tim Seppelt: "Homomorphism Tensors and Linear Equations", CoRR, Vol. abs/2111.11313, 2021.
URL https://doi.org/10.48550/arXiv.2111.11313

Lovász (1967) showed that two graphs G and H are isomorphic if and only if they are homomorphism indistinguishable over the class of all graphs, i.e. for every graph F, the number of homomorphisms from F to G equals the number of homomorphisms from F to H. Recently, homomorphism indistinguishability over restricted classes of graphs such as bounded treewidth, bounded treedepth and planar graphs, has emerged as a surprisingly powerful framework for capturing diverse equivalence relations on graphs arising from logical equivalence and algebraic equation systems. In this paper, we provide a unified algebraic framework for such results by examining the linear-algebraic and representation-theoretic structure of tensors counting homomorphisms from labelled graphs. The existence of certain linear transformations between such homomorphism tensor subspaces can be interpreted both as homomorphism indistinguishability over a graph class and as feasibility of an equational system. Following this framework, we obtain characterisations of homomorphism indistinguishability over two natural graph classes, namely trees of bounded degree and graphs of bounded pathwidth, answering a question of Dell et al. (2018).

3.9 Semiring Semantics and Strategy Analysis

Erich Grädel (RWTH Aachen, DE)

Joint work of Erich Grädel, Niels Lücking, Matthias Naaf
Main reference Erich Grädel, Niels Lücking, Matthias Naaf: "Semiring Provenance for Büchi Games: Strategy Analysis with Absorptive Polynomials", CoRR, Vol. abs/2106.12892, 2021.
URL https://arxiv.org/abs/2106.12892

This paper presents a case study for the application of semiring semantics for fixed-point formulae to the analysis of strategies in Büchi games. Semiring semantics generalizes the classical Boolean semantics by permitting multiple truth values from certain semirings. Evaluating the fixed-point formula that defines the winning region in a given game in an appropriate semiring of polynomials provides not only the Boolean information on who wins, but also tells us how they win and which strategies they might use. This is well-understood for reachability games, where the winning region is definable as a least fixed point. The case of Büchi games is of special interest, not only due to their practical importance, but also because it is the simplest case where the fixed-point definition involves a genuine alternation of a greatest and a least fixed point. We show that, in a precise sense, semiring semantics provide information about all absorption-dominant strategies – strategies that win with minimal effort, and we discuss how these relate to positional and the more general persistent strategies. This information enables further applications such as game synthesis or determining minimal modifications to the game needed to change its outcome. Lastly, we discuss limitations of our approach and present questions that cannot be immediately answered by semiring semantics.

3.10 Logarithmic Weisfeiler-Leman Identifies All Planar Graphs

Sandra Kiefer (RWTH Aachen, DE)

Joint work of Martin Grohe, Sandra Kiefer
Main reference Martin Grohe, Sandra Kiefer: "Logarithmic Weisfeiler-Leman Identifies All Planar Graphs", in Proc. of the 48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference), LIPIcs, Vol. 198, pp. 134:1–134:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
URL http://dx.doi.org/10.4230/LIPIcs.ICALP.2021.134

The Weisfeiler-Leman (WL) algorithm is a well-known combinatorial procedure for detecting symmetries in graphs and it is widely used in graph-isomorphism tests. It proceeds by iteratively refining a colouring of vertex tuples. The number of iterations needed to obtain the final output is crucial for the parallelisability of the algorithm.

In my talk, I presented recent work concerning the number of iterations of the WL algorithm on planar graphs. To be more precise, we found that there is a constant k such that every planar graph can be identified (that is, distinguished from every non-isomorphic graph) by the k-dimensional WL algorithm within a logarithmic number of iterations. This generalises a result due to Verbitsky (STACS 2007), who proved the same for 3-connected planar graphs.

The number of iterations needed by the k-dimensional WL algorithm to identify a graph corresponds to the quantifier depth of a sentence that defines the graph in the $(k+1)$-variable fragment C^{k+1} of first-order logic with counting quantifiers. Thus, our result implies that every planar graph is definable with a C^{k+1}-sentence of logarithmic quantifier depth.

3.11 Determinacy of Real Conjunctive Queries. The Boolean Case

Jerzy Marcinkowski (University of Wroclaw, PL)

Joint work of Jarosław Kwiecień, Jerzy Marcinkowski, Piotr Ostropolski-Nalewaja
Main reference Jarosław Kwiecień, Jerzy Marcinkowski, Piotr Ostropolski-Nalewaja: "Determinacy of Real Conjunctive Queries. The Boolean Case", CoRR, Vol. abs/2112.12742, 2021.
URL https://arxiv.org/abs/2112.12742

In their classical 1993 paper [CV93] Chaudhuri and Vardi notice that some fundamental database theory results and techniques fail to survive when we try to see query answers as bags (multisets) of tuples rather than as sets of tuples. But disappointingly, almost 30 years after [CV93], the bag-semantics based database theory is still in its infancy. We do not even know whether conjunctive query containment is decidable. And this is not due to lack of interest, but because, in the multiset world, everything suddenly gets discouragingly complicated. In this paper, we try to re-examine, in the bag semantics scenario, the query determinacy problem, which has recently been intensively studied in the set semantics scenario. We show that query determinacy (under bag semantics) is decidable for boolean conjunctive queries and undecidable for unions of such queries (in contrast to the set semantics scenario, where the UCQ case remains decidable even for unary queries). We also show that – surprisingly – for path queries determinacy under bag semantics coincides with determinacy under set semantics (and thus it is decidable).

3.12 Quantum isomorphism is equivalent to equality of homomorphism counts from planar graphs I & II

David Earl Roberson (Technical University of Denmark – Lyngby, DK)

Joint work of Laura Mancinska, David E. Roberson
Main reference Laura Mancinska, David E. Roberson: "Quantum isomorphism is equivalent to equality of homomorphism counts from planar graphs", in Proc. of the 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020, pp. 661–672, IEEE, 2020.
URL http://dx.doi.org/10.1109/FOCS46700.2020.00067

Over 50 years ago, Lovász proved that two graphs are isomorphic if and only if they admit the same number of homomorphisms from any graph. Other equivalence relations on graphs, such as cospectrality or fractional isomorphism, can be characterized by equality of homomorphism counts from an appropriately chosen class of graphs. Dvořák [J. Graph Theory 2010] showed that taking this class to be the graphs of treewidth at most k yields a tractable relaxation of graph isomorphism known as k-dimensional Weisfeiler-Leman equivalence. Together with a famous result of Cai, Fürer, and Immerman [FOCS 1989], this shows that homomorphism counts from graphs of bounded treewidth do not determine a graph up to isomorphism. Dell, Grohe, and Rattan [ICALP 2018] raised the questions of whether homomorphism counts from planar graphs determine a graph up to isomorphism, and what is the complexity of the resulting relation. We answer the former in the negative by showing that the resulting relation is equivalent to the so-called quantum isomorphism [Mančinska et al, ICALP 2017]. Using this equivalence, we further resolve the latter question, showing that testing whether two graphs have the same number of homomorphisms from any planar graph is, surprisingly, an undecidable problem, and moreover is complete for the class coRE (the complement of recursively enumerable problems). Quantum isomorphism is defined in terms of a one-round, two-prover interactive proof system in which quantum provers, who are allowed to share entanglement, attempt to convince the verifier that the graphs are isomorphic. Our combinatorial proof leverages the quantum automorphism group of a graph, a notion from noncommutative mathematics.

3.13 First-Order Logic with Connectivity Operators: expressiveness and model-checking

Sebastian Siebertz

Joint work of Michał Pilipczuk, Nicole Schirrmacher, Sebastian Siebertz, Szymon Toruńczyk, Alexandre Vigny
Main reference Nicole Schirrmacher, Sebastian Siebertz, Alexandre Vigny: "First-Order Logic with Connectivity Operators", 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. 34:1–34:17, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.
URL http://dx.doi.org/10.4230/LIPIcs.CSL.2022.34
Main reference Michal Pilipczuk, Nicole Schirrmacher, Sebastian Siebertz, Szymon Torunczyk, Alexandre Vigny: "Algorithms and data structures for first-order logic with connectivity under vertex failures", CoRR, Vol. abs/2111.03725, 2021.
URL https://arxiv.org/abs/2111.03725

First-order logic (FO) can express many algorithmic problems on graphs, such as the independent set and dominating set problem parameterized by solution size. On the other hand, FO cannot express the very simple algorithmic question whether two vertices are connected. We enrich FO with connectivity predicates that are tailored to express algorithmic graph properties that are commonly studied in parameterized algorithmics. By adding the atomic predicates $conn_k(x, y, z_1, \ldots, z_k)$ that hold true in a graph if there exists a path between (the valuations of) x and y after (the valuations of) $z_1, \ldots, z_k$ have been deleted, we obtain *separator logic FO+conn*. We thereby obtain a logic that can express many interesting problems such as the feedback vertex set problem and elimination distance problems to first-order definable classes.

We first study the expressive power of the new logic. We then study the model-checking problem and prove that from the point of view of parameterized complexity, under standard complexity theoretical assumptions, the frontier of tractability of separator logic is almost exactly delimited by classes excluding a fixed topological minor. From the atomic case of connectivity predicates we obtain the first deterministic data structure for connectivity under batched vertex failures where for every fixed number of failures, all operations can be performed in constant time.

This talk is based on the results presented in [2] and [1].

References

1 Michał Pilipczuk, Nicole Schirrmacher, Sebastian Siebertz, Szymon Toruńczyk, and Alexandre Vigny. Algorithms and data structures for first-order logic with connectivity under vertex failures. *arXiv preprint arXiv:2111.03725*, 2021.

2 Nicole Schirrmacher, Sebastian Siebertz, and Alexandre Vigny. First-order logic with connectivity operators. In *30th EACSL Annual Conference on Computer Science Logic (CSL 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

3.14 Model Checking on Interpretations of Classes of Bounded Local Cliquewidth

Szymon Torunczyk (University of Warsaw, PL)

Joint work of Édouard Bonnet, Jan Dreier, Jakub Gajarský, Stephan Kreutzer, Nikolas Mählmann, Pierre Simon, Szymon Toruńczyk
Main reference Michal Pilipczuk, Nicole Schirrmacher, Sebastian Siebertz, Szymon Torunczyk, Alexandre Vigny: "Algorithms and data structures for first-order logic with connectivity under vertex failures", CoRR, Vol. abs/2111.03725, 2021.
URL https://arxiv.org/abs/2111.03725

We present a fixed-parameter tractable algorithm for first-order model checking on interpretations of graph classes with bounded local cliquewidth. Notably, this includes interpretations of planar graphs, and more generally, of classes of bounded genus. To obtain this result we develop a new tool which works in a very general setting of dependent classes and which we believe can be an important ingredient in achieving similar results in the future.

3.15 On guarded team logics

Marius Tritschler (TU Darmstadt, DE)

Guarded logics and team logics are logics of imperfect information, in different ways. When combined, some fragments have nice model theoretic properties while still being reasonably expressive. In particular, a technique called "strategy tree transfer" can be applied to show a finite model property for these fragments.

3.16 On capturing some equivalence relations by homomorphism counts

Wei-Lin Wu (University of California – Santa Cruz, US)

Joint work of Albert Atserias, Phokion G. Kolaitis, Wei-Lin Wu
Main reference Albert Atserias, Phokion G. Kolaitis, Wei-Lin Wu: "On the Expressive Power of Homomorphism Counts", 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 http://dx.doi.org/10.1109/LICS52264.2021.9470543

A classical result by Lovasz states that two graphs are isomorphic if and only if they have the same left profile, i.e., they have the same homomorphism counts "from" all graphs. A similar result by Chaudhuri and Vardi states that two graphs are isomorphic if and only if they have the same right profile, in other words, they have the same homomorphism counts "to" all graphs. By restricting the left or right profile to a class of graphs, we have a relaxation of isomorphism. In this talk, I will share some results about what relaxations of isomorphism do/don't coincide with a restriced left or right profile.

3.17 Iltis: Learning logic in the web

Thomas Zeume (Ruhr-Universität Bochum, DE)

Joint work of Thomas Zeume, Gaetano Geck, Christine Quenkert, Marko Schmellenkamp, Jonas Schmidt, Felix Tschirbs, Fabian Vehlken
Main reference Gaetano Geck, Christine Quenkert, Marko Schmellenkamp, Jonas Schmidt, Felix Tschirbs, Fabian Vehlken, Thomas Zeume: "Iltis: Teaching Logic in the Web", CoRR, Vol. abs/2105.05763, 2021.
URL https://arxiv.org/abs/2105.05763

The Iltis project provides an interactive, web-based system for teaching the foundations of formal methods. It is designed with the objective to allow for simple inclusion of new educational tasks; to pipeline such tasks into more complex exercises; and to allow simple inclusion and cascading of feedback mechanisms. Currently, exercises for many typical automated reasoning workflows for propositional logic, modal logic, and some parts of first-order logic are covered.

In this talk I will address (algorithmic) challenges and solution approaches for building such systems, but also show how the system can be used in logic instruction.

4 Open problems

4.1 Find C3-equivalent graphs with a factor-2 gap in the size of their minimum vertex covers

Albert Atserias (UPC Barcelona Tech, ES)

The minimum vertex cover of a graph G, denoted by $k(G)$, is the least cardinality of a set of vertices that touches every edge. By rounding the straightforward linear programming relaxation of the problem, $k(G)$ can be approximated within a factor of two in polynomial time. It is conjectured that, for every positive real $\epsilon \in (0,1]$, the problem of approximating $k(G)$ by a factor better than $(2-\epsilon)$ is NP-hard. Is it also hard for k-variable counting logic C^k for every fixed k? A precise statement of the problem is the following:

Question: Is it true that for every natural number k and every real number $\epsilon \in (0,1]$ there exist C^k-equivalent graphs G and H with $k(G) > (2-\epsilon)k(H)$? One way to ensure $k(G) > (2-\epsilon)k(H)$ is by having G have independence number at most $o(n)$ and H have independence number at least $(1/2 - -o(1))n$, where $n = |V(G)| = |V(H)|$ and $n \to \infty$.

For $k = 2$ and arbitrary $\epsilon \in (0,1]$, graphs as required in the question can be easily found (see [1]). For $k = 3$ and $\epsilon = 0$ it follows from the results in [2] that such graphs do not exist. For arbitrary positive integer k and $\epsilon \in [0.73, 1]$, graphs as required are shown to exist in [1]. For $k \geq 3$ and $\epsilon \in (0, 0.74]$, the problem is open.

References

1 Albert Atserias, Anuj Dawar: Definable Inapproximability: New Challenges for Duplicator. J. Log. Comput. 29(8): 1185-1210 (2019)

2 Albert Atserias, Elitza N. Maneva: Sherali-Adams Relaxations and Indistinguishability in Counting Logics. SIAM J. Comput. 42(1): 112-137 (2013)

4.2 Can the current bounds for computing a Tarski fixed point on a finite (grid) lattice be improved?

Kousha Etessami (University of Edinburgh, GB)

Let $[N] = \{1, \dots, N\}$.

Consider a function

$$f : [N]^d \to [N]^d$$

mapping the finite d-dimensional euclidean grid lattice with sides of length N to itself.

Suppose the function f is *monotone* with respect to the standard coordinate-wise partial order on vectors in $[N]^d$, meaning that for all $x, y \in [N]^d$, if $x \leq y$ then $f(x) \leq f(y)$. By Tarski's (1953) fixed point theorem, such a monotone function f must have a fixed point, i.e., there must exist a point

$$x' \in [N]^d$$

such that $f(x') = x'$.

How hard is it to compute such a fixed point when (a) the function f is given to us as a black box and we wish to find a fixed point with a minimum number of queries? or, (b) the function f is given to us explicitly but succinctly using a boolean circuit, with $d \cdot \log(N)$ input gates and $d \cdot \log(N)$ output gates.

For (a), it is easy to see that standard (Kleene) value iteration, starting from the bottom $\bar{1}$ (or, respectively, top $\bar{N}$) of the lattice $[N]^d$ requires at most $d \cdot N$ queries to find the least (respectively, greatest) fixed point of f. On the other hand, suppose we don't care which fixed point we compute. Suppose any fixed point will do. Can we do better?

This problem has a number of important applications. In particular, efficient algorithms for this problem (polynomial in both d and $\log(N)$) would imply polynomial time algorithms for supermodular games (a very well studied class of games in economic theory), as well as for both Condon's simple stochastic games and Shapley's stochastic games, which are long standing open questions (see [1].)

There is a black-box algorithm due to Dang, Qi, and Ye [2], which requires at most $\log^d(N)$ queries to find a fixed point, and uses a combination of recursion and binary search. This algorithm has recently been improved by Fearnley, Palvolgyi, and Savani [3], who give an upper bound of $\log^{\lceil \frac{2d}{3} \rceil}(N)$ queries for finding a fixed point. There has also been some even more recent (unpublished) results which improve the exponent further, but it remains exponential in d.

What is the best upper bound we can obtain for computing a Tarski fixed point? This is very much an open question. Etessami, Papadimitriou, Rubinstein, and Yannakakis [1] provide a lower bound of $\Omega(\log^2(N))$ queries in the black-box model for (a), already for 2-dimensional monotone functions $f : [N]^2 \to [N]^2$. They also show that a total search version of the white-box Tarski problem (b) is in both the complexity classes PPAD and PLS. (It thus follows from recent results that the Tarski problem is also in the total search complexity classes CLS and EOPL.)

References

1 K. Etessami, C. Papadimitriou, A. Rubinstein, M. Yannakakis, "Tarski's Theorem, Supermodular Games, and the Complexity of Equilibria", Proceedings of 11th Innovations in Theoretical Computer Science conference (ITCS'20), 2020.

2 C. Dang, Q. Qi, and Y. Ye. “Computational models and complexities of Tarski’s fixed points”. Technical Report, Stanford University, 2012.

3 J. Fearnley, D. Palvolgyi, and R Savani: A Faster Algorithm for Finding Tarski Fixed Points. arXiv:2010.02618 (2021). (Earlier version in STACS 2021).

Participants

- Isolde Adler
University of Leeds, GB
- Albert Atserias
UPC Barcelona Tech, ES
- Christoph Berkholz
HU Berlin, DE
- Mikolaj Bojanczyk
University of Warsaw, PL
- Andrei A. Bulatov
Simon Fraser University – Burnaby, CA
- Thomas Colcombet
CNRS – Paris, FR
- Anuj Dawar
University of Cambridge, GB
- Kousha Etessami
University of Edinburgh, GB
- Diego Figueira
CNRS &
Université de Bordeaux , FR
- Erich Grädel
RWTH Aachen, DE
- Martin Grohe
RWTH Aachen University, DE
- Sandra Kiefer
RWTH Aachen, DE
- Aliaume Lopez
ENS – Gif-sur-Yvette, FR
- Jerzy Marcinkowski
University of Wroclaw, PL
- Rémi Morvan
University of Bordeaux, FR
- Martin Otto
TU Darmstadt, DE
- David Earl Roberson
Technical University of Denmark – Lyngby, DK
- Tim Seppelt
RWTH Aachen, DE
- Sebastian Siebertz
Universität Bremen, DE
- Szymon Torunczyk
University of Warsaw, PL
- Marius Tritschler
TU Darmstadt, DE
- Alexandre Vigny
Universität Bremen, DE
- Igor Walukiewicz
University of Bordeaux, FR
- Wei-Lin Wu
University of California – Santa Cruz, US
- Thomas Zeume
Ruhr-Universität Bochum, DE

Report from Dagstuhl Seminar 22052

The Human Factors Impact of Programming Error Messages

Brett A. Becker*[1], Paul Denny*[2], Janet Siegmund*[3], Andreas Stefik*[4], and Eddie Antonio Santos†[5]

1 **University College Dublin, IE.** brett.becker@ucd.ie
2 **University of Auckland, NZ.** p.denny@auckland.ac.nz
3 **TU Chemnitz, DE.** siegj@hrz.tu-chemnitz.de
4 **University of Nevada - Las Vegas, US.** stefika@gmail.com
5 **University College Dublin, IE.** eddie.santos@ucdconnect.ie

Abstract

The impacts of many human factors on how people program are poorly understood and present significant challenges for work on improving programmer productivity and effective techniques for teaching and learning programming. Programming error messages are one factor that is particularly problematic, with a documented history of evidence dating back over 50 years. Such messages, commonly called compiler error messages, present difficulties for programmers with diverse demographic backgrounds. It is generally agreed that these messages could be more effective for all users, making this an obvious and high-impact area to target for improving programming outcomes. This report documents the program and the outputs of Dagstuhl Seminar 22052, "The Human Factors Impact of Programming Error Messages", which explores this problem. In total, 11 on-site participants and 17 remote participants engaged in intensive collaboration during the seminar, including discussing past and current research, identifying gaps, and developing ways to move forward collaboratively to address these challenges.

Seminar January 30–February 02, 2022 – http://www.dagstuhl.de/22052

2012 ACM Subject Classification General and reference → Surveys and overviews; Human-centered computing → Human computer interaction (HCI); Social and professional topics → Computing education; Applied computing → Education; Software and its engineering → General programming languages

Keywords and phrases Accessibility; Compilers; Compiler Error Messages; Computer Science Education; Computer-Human Interaction; Computing Education; CS1; Error Messages; HCI; Human Factors; Human Computer Interaction; Novice Programmers; Programming Errors; Programming Error Messages; Programming Languages; Software Engineering; Syntax Errors

Digital Object Identifier 10.4230/DagRep.12.1.119

* Editor / Organizer
† Editorial Assistant / Collector

The Human Factors Impact of Programming Error Messages, *Dagstuhl Reports*, Vol. 12, Issue 1, pp. 119–130
Editors: Brett A. Becker, Paul Denny, Janet Siegmund, Andreas Stefik, and Eddie Antonio Santos
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Brett A. Becker (University College Dublin, IE, brett.becker@ucd.ie)
Paul Denny (University of Auckland, NZ, p.denny@auckland.ac.nz)
Janet Siegmund (TU Chemnitz, DE, siegj@hrz.tu-chemnitz.de)
Andreas Stefik (University of Nevada - Las Vegas, US, stefika@gmail.com)

Programming error messages (commonly called compiler error messages) pose challenges to programmers – from novices to professionals – with evidence dating from the 1960s to present day. In this seminar, we explored the nature of these challenges and particularly why they remain largely unaddressed. We further investigated the specific challenges that different users including children, non-native English speakers, and those of varying ability experience when faced with programming error messages. Finally, we sought to identify the most promising avenues to assess the effectiveness of error messages, how to improve them with large, demonstrated effect, and how to produce appropriate messages for different users with different needs. To this end, we assembled experts from many sub-disciplines of Computer Science, including Programming Languages, HCI, Computer Science Education, and Software Engineering as well as Learning Sciences. Due to travel restrictions imposed by the COVID-19 pandemic, we ran the seminar in a hybrid format, with 11 on-site participants collaborating with 17 remote participants. We formulated a schedule for the seminar that sought to maximise outcomes for all participants.

By combining the expertise of these different disciplines, we could identify gaps in knowledge and high-priority areas to build a basis for future work. This is the starting point for a long-lasting contribution to the field. By uniting communities that have to date been working largely in isolation, we sought to gather appropriate and useful data, broaden perspectives, build consensus among diverse stakeholders, and begin cross-community efforts working in unison going forward.

During the seminar, we had sessions focusing on existing literature and practice, including experience reports from language maintainers, expert users, researchers, and educators, to develop a shared understanding of the evidence that exists with regard to effective programming error messages. This included the group working together to synthesise existing data sets, which we view as an important exercise given that large corpora of errors and error message data are generally not openly available. This can make it difficult for researchers to answer seemingly simple questions such as: “What are the most frequent error messages encountered by [students, professionals, blind, non-native English speakers] in language x?” or, “What are evidence-based examples of effective and/or ineffective error messages?”. As a result, we identified several fruitful avenues in the form of cross-discipline collaboration, data sharing opportunities, and improved focus on what steps are needed to improve efforts to answer these questions.

One of the key objectives of the seminar was to identify areas for immediate research. Several problems have been identified, including but not limited to:

1. linking error messages with the context of the problems and programs being worked on when error messages are generated;
2. understanding the metacognitive aspects involved in the process of interpreting and reacting to error messages;
3. identifying what error messages arise from inconsistent conceptions of how a program runs or how it is structured;

4. the effects of error message presentation on error message interpretation (e.g. visual queues, structured error messages, etc.);
5. error message classification schemes;
6. metrics to measure and assess error messages;
7. identifying when error messages are "wrong", what occurs when they steer the programmer in the wrong direction, and how this can be avoided;
8. determining the design factors of programming error messages that differ across demographic groups (e.g., expertise level, disabilities, native/natural language, etc.).

Another key objective was to establish new cross-community efforts to improve programming error messages in practice, leveraging the strengths of each community. Participants discussed open research problems and identified those of high priority and interest including those listed above. Several interdisciplinary research objectives have been established, and seeds were sewn to form teams to collaboratively address these research questions.

2 Table of Contents

Executive Summary

Overview of Talks

Panel discussions

3 Overview of Talks

3.1 After ++5 Decades of Error Messages, Where Do We Go (right) Now?

Brett A. Becker (University College Dublin, IE)

Joint work of Brett A. Becker, Paul Denny, Raymond Pettit, Durell Bouchard, Dennis J. Bouvier, Brian Harrington, Amir Kamil, Amey Karkare, Chris McDonald, Peter-Michael Osera, Janice L. Pearce, James Prather

Main reference Brett A. Becker, Paul Denny, Raymond Pettit, Durell Bouchard, Dennis J. Bouvier, Brian Harrington, Amir Kamil, Amey Karkare, Chris McDonald, Peter-Michael Osera, Janice L. Pearce, James Prather: "Compiler Error Messages Considered Unhelpful: The Landscape of Text-Based Programming Error Message Research", in Proc. of the Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE 2019, Aberdeen, Scotland Uk, July 15-17, 2019, pp. 177–210, ACM, 2019.

URL http://dx.doi.org/10.1145/3344429.3372508

Text-based error messages are the primary mechanism provided to programmers by a programming language to fix errors in code. Error messages have been studied for over 50 years in the context of education, software engineering, human-computer interaction, and programming languages. They are often described as confusing and misleading, and are known to cause frustration, particularly for students and novices, but also for professionals. They are situated in the complex boundary between the human user and the system and involve constructed computer languages, human language, cognition, and complex concepts such as the notional machine. They have impacts on humans and society, including the economy, the workplace, and human and societal productivity. They also likely impact diversity, equality, and inclusion as they affect various people and groups differently. The historical literature has called for several avenues to be explored in order to improve their effectiveness. Although progress has been made on this front in recent years many areas remain largely under-explored. New advancements in cognitive science and neuroscience, as well as approaches based on artificial intelligence show promise in improving error messages. However much fundamental work remains. This presentation provides an overview of this landscape, presents a corpus of 330 scientific papers on programming error messages spanning 1965–2019 [1], and provides several avenues for future work – from fundamental human factors research to those based on emerging technologies.

References

1 Becker, Brett A., & Denny, Paul, & Pettit, & Bouchard, Durell & Bouvier, Dennis J. & Harrington, Brian & Kamil, Amir & Karkare, Amey & McDonald, Chris & Osera, Peter-Michael & Pearce, Janice L. & Prather, James, *Compiler Error Messages Considered Unhelpful: The Landscape of Text-Based Programming Error Message Research*. In Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR '19), Association for Computing Machinery, New York, New York USA, 2019, `https://doi.org/10.1145/3344429.3372508`

3.2 Teaching Novices to Read (gcc) Programming Error Messages?

Dennis Bouvier (Southern Illinois Univ. Edwardsville, US)

Until better tools are commonly used, students could benefit from being better prepared to make use of error messages in current systems. Presented are: (1) Ideas for essential topics on for novice programmers learning to use error messages effectively. The topics include (1a) terminology, (1b) compiler technology for producing error messages, and (1c) strategies for using error messages effectively. (2) Ideas for evaluating the efficacy of the lesson, including (2a) debugging activity, and (2b) a questionnaire of personal debugging practice. In the debugging activity, students are presented with a sequence of programs that each have one error message. The student then (i) compiles the program. (ii) reports the error message found, (iii) explains the error message in their own words, and (iv) proposes a fix for the code without actually editing the code nor recompiling. The questionnaire of personal debugging practice asks the student to report on activities they employ in response to error messages. These activities were experienced by students as the third lesson of an introduction to programming course using the C programming language. The debugging activity and questionnaire were repeated two-thirds through the course. Data and analysis to be reported at a later date.

3.3 Blackbox

Neil Brown (King's College London, GB)

Blackbox is a large dataset of programming activity, collected worldwide from users of the BlueJ IDE that helps novices learn to program Java. In this session I will describe the dataset and then guide attendees through an activity to explore some of the data and see how novices respond to error messages "in the wild". Finally, I will describe Blackbox Mini, a new small subset of the data with a simple schema.

3.4 Improving Programming Error Messages: Why Should We Bother?

Paul Denny (University of Auckland, NZ)

It is well known that programming error messages can be notoriously difficult for novices to understand. This is a long standing problem and one that, if solved, would have clear positive impacts on student learning. At this seminar, all participants bring with them a wonderful diversity of motivations, interests and expertise for tackling this problem. I will articulate my own motivations, which include that students have suffered long enough, and will support these with empirical data illustrating the large positive effects that better error message have on the programming performance of novices.

3.5 Error Messages: Six Seconds of Progress from 25 Years of Research

Kathi Fisler (Brown University, Providence, US) and Shriram Krishnamurthi (Brown University, Providence, US)

We have made significant progress in our theoretical and practical understanding of error messages. Some of this work is in languages and environments outside the mainstream; in other cases, these are general results that can be applied widely.

The talk presents eight vignettes describing results, challenges, and cautions for interpretation.

3.6 Observing Programmers with EEG and Eye Tracking

Janet Siegmund

To gain deeper insights into how programming error messages are perceived by programmers, we can use established methods from neuroscience and cognitive psychology, especially electroencephalography and eye tracking. We demonstrate a possible experiment with EEG and eye tracking and show what we could do with the data.

3.7 On the Location of Errors in Source Code

Tobias Kohn (University of Cambridge, GB)

Compilers tend to present errors at specific locations in the source code, indicating where exactly the parsing process or name resolution ground to a halt. We might assume that this should also be the region to make corrections and amend the source code. Closer inspection, however, reveals that a substantial part of the errors are not that easily pinned to a specific location. In fact, it might be more fruitful to think of compiler errors primarily as *inconsistencies* in the source code. Consider a variable name, for instance, that is spelled differently – who is to say that the variable's declaration must be the correct spelling with all other deviating occurrences being wrong?

On the basis of several examples of student code we make a case that indeed most programming errors may be better viewed as matters of inconsistencies in source code rather than clear-cut mistakes. Presenting programming errors as such inconsistencies might also work towards the learners understanding as to why the computer has trouble "understanding" the program code, with less emphasis on arbitrary syntactic rules that must be followed.

3.8 Novices vs Expert Errors: One Size Fits None?

Linda McIver (ADSEI, Glen Waverley, AU)

A lively discussion of what we need from error messages, and whether novice and expert programmers' needs are irreconcilably different. The consensus, if there was one, seemed to be that novices and experts require different things, but there were a few assumptions that require further testing.

3.9 Error Message Topic Potpourri

Prajish Prasad (IIT Bombay, IN)

In this session, I facilitated a discussion where we discussed three topics related to error messages. For each topic, I provided some background and focus questions. Discussions were centred around these focus questions.

The first topic was related to "The Effect of Error Messages in Learners' Metacognition and Self-Regulation". Metacognition, in simple words is "thinking about one's own thinking". When learners encounter a programming task, applying metacognition and self-regulation strategies involve identifying strategies that have been successful or unsuccessful in the past and evaluating these strategies based on feedback. We discussed how error messages can hinder learners' metacognitive processes and whether we could provide general metacognitive scaffolds to help learners recover from these errors.

The second topic was related to "Error Messages and Live Coding". Live coding is the process of writing code live on a computer. In recent times, several developers as well as instructors have been doing live coding. We discussed whether general patterns or guidelines can be extracted from live coding videos that show expert developers encountering and fixing error messages.

Finally, we discussed about "Error Messages in Web Programming Languages", and the challenges in designing and improving error messages for such languages. The web has so many languages and implementations, that deciphering errors often requires a strong understanding of the architecture of the web, of browsers, and of languages.

3.10 Terminology Used When Discussing Programming Error Messages

Eddie Antonio Santos (University College Dublin, IE)

I was interested in the seminar participants' views and attitudes about the words we use when talking about programming error messages (PEMs). I divided this talk into two sections:

1. I had participants post their favourite error message. "Favourite" was defined however the submitter chose.

2. I had participants collectively edit a Google Document adding to a list of salient words found within programming error messages: and a list of words that us – educators, language implementers, and researchers – use when talking *about* PEMs.

Our discussion revealed that our attitudes towards the wording found within contemporary PEMs is largely negative, and can be improved in terms of tone. We note that for experts, words like "illegal", "offending", "fatal", are okay, however novices may be negatively affected by these "scary" words. We suggest that compiler developers may wish to add another stage to the compiler explicitly dedicated to providing feedback to the programmer. This module should be maintained by a professional in human-computer interaction. We also float the idea that "error" is the wrong word to use when discussion feedback whatsoever.

4 Panel discussions

4.1 inHUMANe Programming Error Messages – Position Panel

Dennis Bouvier (Southern Illinois Univ. Edwardsville, US), Kerstin Haring (University of Denver, US), Felienne Hermans (Leiden University, NL), Amy Ko (University of Washington, Seattle, US), Michael Kölling (King's College London, GB), and James Prather (Abilene Christian University, US)

The organisers of the workshop posed position statements to the panel. Panellists discussed both sides of several positions, including (1) Error messages for novice programmers should be avoided at all costs, (2) Kids should learn programming before learning to read and write natural languages, and (3) Internationalised programming error messages are harmful.

With respect to position 1, the position was supported and opposed by most panellists. For example, the argument can be made that error message understanding is crucial for forming professional skills and/or for building knowledge of how programming systems work; likewise the argument can be made when the goal is building confidence with computing, or using computing as a means to study another discipline, learning should not be hampered by dealing with avoidable error messages.

With respect to position 2, the position is supported by some panellists considering that programming is (or will be) a fundamental skill, and that learning programming early in life could be beneficial. Most panellists pushed back on the use of the word "before" in the position, stating that learning to code could, and perhaps should, coincide with learning natural language(s).

With respect to position 3, multiple panellists posed the possibility of using programming error messages expressed in a natural language being learned as a way to expand the programmer's experience in the language while programming; though, it was not entirely clear that such suggestions were entirely genuine. Also suggested was that American programmers would not be able to adapt to messages in a foreign language, and so all error messages should be in English. As this round progressed, it became clearer that, at least, half of the panellists' positions were somewhat facetious. Taking a tangent in the discussion, panellists suggested that no written language be used – errors could be conveyed by "screaming" and/or "electric shock" or memes.

The final point of discussion was “how is, or how can, change of programming technology be influenced for the better?” Michael Kölling suggested that “creating new better things” is the way to influence the way forward, and gave Snap*!* as an example. Amy Ko suggested that organised community effort can be a way forward, and that influencing “error messages” is a unique opportunity for influencing. Ultimately, the panel suggested this was an Hegelian dialectic; and that making error messages more understandable might fit some goals, but not others.

4.2 Directions for Error Message Research

Jan Pearce (Berea College, US), Neil Brown (King’s College London, GB), Linda McIver (ADSEI, Glen Waverley, AU), Jens Mönig (SAP SE, Walldorf, DE), and Raymond Pettit (University of Virginia, Charlottesville, US)

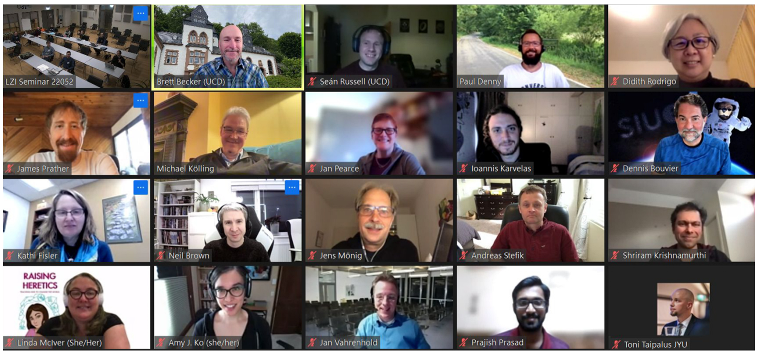
Neil Brown, Linda McIver, Jens Mönig, Jan Pearce, and Raymond Pettit served as discussants, presenting their views on the impact of error messages on identity and self-efficacy, the needs of novice programmers versus expert programmers, what research is needed in these key areas, as well as what concerns panellists have as reviewers of conference papers on error messages. Panellists also discussed how one might make the delivery of error messages more nuanced as well as more accessible to those with disabilities. Neil Brown talked about activity traces from the Blackbox data, and how error messages can often send students down the wrong route to fix their code, thus increasing their frustration and damaging their confidence. Linda McIver talked about how many students come to programming convinced that it’s too hard, and that unintelligible error messages are seen as proof. These students are looking for evidence that they can’t program, and bad error messages provide that proof. Jens Mönig discussed his concern that expectations regarding the positive and negative impacts of error messages may be too large given that these impacts are so heavily dependent upon many other contextual variables that are difficult to disambiguate. Jan Pearce shared and led a discussion on student-produced artefacts that self-describe the negative impacts error messages have personally had on them. Participants discussed whether these self-described negative impacts were due to the error message content or the error itself. Raymond Pettit discussed the differences in the ways novices and experts relate to error messages. Experts are accustomed to receiving these messages and use them as a tool in helping to fix a program, whereas novices often see the error message as the problem. Attendees commented on the ideas shared.

Participants

- Annabelle Bergum
Universität des Saarlandes – Saarbrücken, DE
- Joe Dillane
University College Dublin, IE
- Kerstin Haring
University of Denver, US
- Elisa Madeleine Hartmann
TU Chemnitz, DE
- Felienne Hermans
Leiden University, NL
- Tobias Kohn
University of Cambridge, GB
- Jens Mönig
SAP SE – Walldorf, DE
- Raymond Pettit
University of Virginia – Charlottesville, US
- Ma. Mercedes T. Rodrigo
Ateneo de Manila University – Quezon City, PH
- Eddie Antonio Santos
University College Dublin, IE
- Janet Siegmund
TU Chemnitz, DE

Remote Participants

- Brett A. Becker
University College Dublin, IE
- Dennis Bouvier
Southern Illinois Univ. Edwardsville, US
- Neil Brown
King's College London, GB
- Paul Denny
University of Auckland, NZ
- Kathi Fisler
Brown University – Providence, US
- Ioannis Karvelas
University College Dublin, IE
- Amy Ko
University of Washington – Seattle, US
- Michael Kölling
King's College London, GB
- Shriram Krishnamurthi
Brown University – Providence, US
- Linda McIver
ADSEI – Glen Waverley, AU
- Jan Pearce
Berea College, US
- Prajish Prasad
IIT Bombay, IN
- James Prather
Abilene Christian University, US
- Seán Russell
University College Dublin, IE
- Andreas Stefik
University of Nevada – Las Vegas, US
- Toni Taipalus
University of Jyväskylä, FI
- Jan Vahrenhold
Universität Münster, DE


LZI Seminar 22052
Brett Becker (UCD)
Seán Russell (UCD)
Paul Denny
Didith Rodrigo
James Prather
Michael Kölling
Jan Pearce
Ioannis Karvelas
Dennis Bouvier
Kathi Fisler
Neil Brown
Jens Mönig
Andreas Stefik
Shriram Krishnamurthi
RAISING HERETICS
Linda McIver (She/Her)
Amy J. Ko (she/her)
Jan Vahrenhold
Prajish Prasad
Toni Taipalus JYU