

# Holistic Graph-Processing Systems: Enabling Real-World Scale and Societal Impact

Alexandru Iosup<sup>\*1</sup>, Ana Lucia Varbanescu<sup>\*2</sup>, Hannes Voigt<sup>\*3</sup>, and Jože Rožanec<sup>†4</sup>

1 VU Amsterdam, NL. alexandru.iosup@gmail.com

2 University of Twente – Enschede, NL. a.l.varbanescu@utwente.nl

3 Neo4j – Leipzig, DE. hannes.voigt@neo4j.com

4 Jozef Stefan Institute – Ljubljana, SI. jmrozanec@gmail.com

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25171, “Holistic Graph-Processing Systems: Enabling Real-World Scale and Societal Impact”.

Motivated by the need to tackle the challenges that massive and complex data production and consumption bring to our interconnected, digital world, this seminar focused on large-scale graph processing as a systematic approach to transform these challenges into opportunities. Graphs provide a universal mathematical abstraction for such data, and they already influence various sectors – such as logistics, drug discovery, or fraud detection. However, we have only begun to realize their potential. Nevertheless, the benefits of graph processing could be canceled out by the rapid increase in data scale and diversity, as well as the increasing complexity in developing, executing, and sharing graph-based algorithms and workflows. The emerging field of graph processing systems promises to tackle these challenges. To make such systems effective and efficient, and facilitate their adoption, we need holistic approaches to cope with data transformation and ingestion, workload and system dynamics, high-tier graph programming and co-design with the platform, the emerging computing continuum, and domain-specific needs, among others.

Our seminar explored the symbiosis of graph systems, machine learning, and network science by bringing together researchers, developers, and practitioners actively working on these topics with a focus on graphs. The seminar featured a mix of invited talks, expert panels, and focused discussion groups. The report documents these different elements, summarizes the main findings, and identifies the open problems and challenges that we will tackle next as a joint community.

**Seminar** April 21–25, 2025 – <https://www.dagstuhl.de/25171>

**2012 ACM Subject Classification** Computing methodologies → Distributed computing methodologies; Computing methodologies → Machine learning; Hardware → Emerging technologies; Information systems → Data management systems; Mathematics of computing → Graph theory

**Keywords and phrases** digital continuum choreography, graph processing optimization, machine learning on graphs, massive graphs, sustainable distributed graph processing

**Digital Object Identifier** 10.4230/DagRep.15.4.79

---

\* Editor / Organizer

† Editorial Assistant / Collector



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

Holistic Graph-Processing Systems: Enabling Real-World Scale and Societal Impact, *Dagstuhl Reports*, Vol. 15, Issue 4, pp. 79–91

Editors: Alexandru Iosup, Ana Lucia Varbanescu, Hannes Voigt, and Jože Rožanec



Dagstuhl Reports

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


## 1 Executive Summary

*Ana Lucia Varbanescu (University of Twente – Enschede, NL)*

*Alexandru Iosup (VU Amsterdam, NL)*

*Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI)*

*Hannes Voigt (Neo4j – Leipzig, DE)*

License  Creative Commons BY 4.0 International license

© Ana Lucia Varbanescu, Alexandru Iosup, Jože Rožanec, and Hannes Voigt

In today’s digital landscape, complexity grows with increasing data volume and degree of interconnection. A suitable data abstraction is crucial for comprehending and navigating this dense network of connections. Starting from Euler’s pioneering work on The Bridges of Königsberg in 1735, graphs have steadily evolved as a robust and adaptable conceptual framework. Graphs are universal representations of concepts, where nodes are markers for distinct entities and edges delineate their interrelations, further enriched with detailed annotations when necessary. Graphs are successful in various domains, like bioinformatics, e-commerce, logistics and transportation networks, urban planning, and even pandemic analysis or vaccine development (e.g., during COVID-19).

Although graphs enable complex analysis and decision-making, processing graphs to understand real-world phenomena and to solve real-world problems raises many challenges that threaten to keep graph processing intractable for the current generation of applications. For example, creating graphs from massive data sources or with generative approaches poses multiple challenges, including volume, velocity, and variety. Furthermore, the variability and irregularity of graphs and their processing algorithms challenge the use of established heterogeneous hardware, general-purpose big data solutions, or computing continuum mechanisms. Continuous operation on (streaming) graphs requires new techniques for adaptivity and optimization – e.g., provisioning, allocation, elastic scaling, migration, offloading, partitioning, consolidation, and caching – to be combined across large-scale information and communication technology infrastructure.

Addressing these challenges for graph processing workflows at real-world scale and with societal impact requires a holistic approach that leverages the expertise and synergies of multiple communities related to graph processing. Our seminar did provide a unique opportunity for encounters between these distinct communities, each addressing graph processing at scale. We brought together three essential communities: distributed, parallel, and cluster computing, machine learning for, on, and with graphs, and social and information networks. We therefore facilitated a better understanding of each community’s challenges regarding graph processing, and promoted a synergic relationship to shape holistic, actionable knowledge of graph processing.

In search of the holistic view of massive-scale graph processing, the seminar featured five topics of discussion: (1) massive graph creation with generative and analytical approaches, (2) graph processing algorithms and workflows, (3) graph operations across the digital continuum, (4) adaptivity and optimization to ensure performance, scalability, and sustainability, and (5) applications at real-world scale with near-term societal impact. For each of these topics, the seminar dedicated half a day, featuring expert talks, an expert panel, and break-out session to facilitate discussions. All these activities shaped a comprehensive vision and sketched a roadmap to guide research in graph processing for the upcoming years.

Our main goals were: (i) to establish a uniform vocabulary across communities for issues related to graph processing, (ii) to identify key open graph-processing challenges and opportunities across communities along with ideas for long-term research, (iii) to co-design

a holistic approach (blueprint, reference architecture, experimental methodology) to graph processing in the digital continuum, and (iv) to identify flagship applications for holistic graph processing with real-world scale and societal impact.

Our future plans include high-visibility collaboration and dissemination, with mechanisms such as roadmap white-papers, networks of excellence, EU-level projects, and follow-up Dagstuhl Seminars.

## 2 Table of Contents

### Executive Summary

*Ana Lucia Varbanescu, Alexandru Iosup, Jože Rožanec, and Hannes Voigt . . . . .* 80

### Overview of Talks

|   |    |
|---|----|
| Democratizing Large-Scale Graph Analytics: From Supercomputing to Societal Impact |    |
| <i>David A. Bader . . . . .</i>   | 83 |
| Graphs and Large Language Models: Vision or Reality ?                             |    |
| <i>Angela Bonifati . . . . .</i>  | 83 |
| Structured data has a lot of value and we need to learn how to tap into it        |    |
| <i>Michael Cochez . . . . .</i>   | 84 |
| From Graphs to Design Automation: Custom Systems for Data Analytics               |    |
| <i>Antonino Tumeo . . . . .</i>   | 84 |
| Two Graph-Related Topics  |    |
| <i>M. Tamer Özsu . . . . .</i>  | 85 |

### Working groups

|  |    |
|--|----|
| Holistic graph operations at scale, across the digital continuum   |    |
| <i>Andrea Bartolini, Duncan Bart, Dante Niewenhuis, Jože Rožanec, Daniël ten Wolde, Antonino Tumeo, and Nikolay Yakovets . . . . .</i>                   | 85 |
| Applications at a real-world scale with near-term societal impact  |    |
| <i>Aydin Buluc, Duncan Bart, Stefania Dumbrava, Kamesh Madduri, Dante Niewenhuis, Jože Rožanec, Daniël ten Wolde, and Ana Lucia Varbanescu . . . . .</i> | 86 |
| Adaptivity and optimization to ensure performance, scalability, and sustainability   |    |
| <i>Stefania Dumbrava, Duncan Bart, Peter A. Boncz, Florina M. Ciorba, Dante Niewenhuis, Jože Rožanec, and Daniël ten Wolde . . . . .</i>                 | 87 |
| Graph processing algorithms and workflows  |    |
| <i>Johannes Langguth, Duncan Bart, Sanjukta Bhowmick, Dante Niewenhuis, Jože Rožanec, Ingo Scholtes, and Daniël ten Wolde . . . . .</i>                  | 88 |
| Massive graph creation with generative and analytical approaches   |    |
| <i>Nikolay Yakovets, Bogdan Arsintescu, Duncan Bart, Jože Rožanec, Juan F. Sequeda, and Daniël ten Wolde . . . . .</i>                                   | 88 |

### Open problems

|  |    |
|--|----|
| Open problems regarding Holistic Graph-Processing Systems to Enable Real-World Scale and Societal Impact |    |
| <i>Jože Rožanec . . . . .</i>  | 89 |

|                               |           |
|-------------------------------|-----------|
| <b>Participants . . . . .</b> | <b>91</b> |
|-------------------------------|-----------|

### 3 Overview of Talks

#### 3.1 Democratizing Large-Scale Graph Analytics: From Supercomputing to Societal Impact

*David A. Bader (NJIT – Newark, US)*

License  Creative Commons BY 4.0 International license  
© David A. Bader


In this talk, Distinguished Professor David A. Bader explores the evolution and impact of large-scale graph analytics, from his pioneering work in Linux supercomputing to today's democratization of massive data science capabilities. The presentation highlights how the open-source Arachne framework, built on Arkouda, enables researchers and organizations to process and analyze graphs containing terabytes of data through an accessible Python interface, while the heavy computational work occurs on powerful backend systems.

Bader discusses three critical application domains of this technology: national security (detecting malicious network activity through relationship patterns), computational neuroscience (analyzing connectomes containing millions of neurons and synapses), and scientometrics (mapping research collaboration networks). These examples demonstrate how graph analytics can solve complex problems that were previously inaccessible due to computational limitations.

The talk emphasizes the parallel between Bader's 1998 development of the Linux-based supercomputer – now the architecture for 100% of the world's top supercomputers with an estimated economic impact of \$100 trillion – and his current mission to democratize access to sophisticated graph analytics capabilities. Through open-source tools and frameworks, Bader's work continues to bridge the gap between cutting-edge computing resources and real-world applications, making powerful data analysis accessible to researchers and practitioners across disciplines.

#### 3.2 Graphs and Large Language Models: Vision or Reality ?


*Angela Bonifati (Lyon 1 University & IUF, FR)*

License  Creative Commons BY 4.0 International license  
© Angela Bonifati

This talk explores the interplay of graph data models and Large Language Models (LLMs), addressing whether their integration is a current reality or an emerging vision. Graphs, as intuitive and powerful abstractions for modeling interconnected data, underpin a wide array of applications from social networks to scientific data analysis. The presentation dives into property graph transformations, emphasizing the need for expressive, declarative tools to enable robust, composable rule-based transformations and graph creation. The second part examines the role of LLMs in generating transformation rules and consistency constraints for property graphs. It discusses techniques for encoding graphs for LLM processing, compares performance across models (LLaMA-3, Mixtral), and highlights challenges in rule quality, scalability, and semantics. The talk concludes by outlining the promise of retrieval-augmented generation (RAG) for enhancing LLM capabilities with external graph-based knowledge, offering a roadmap for integrating symbolic and neural reasoning in future systems.

### 3.3 Structured data has a lot of value and we need to learn how to tap into it

*Michael Cochez (VU Amsterdam, NL)*

License  Creative Commons BY 4.0 International license  
© Michael Cochez


Recent advances in machine learning, especially with language models, have transformed many fields, dramatically raising expectations for artificial intelligence systems. Yet, these advances have also exposed a critical weaknesses: a lack of reliability and a need for a ridiculous amount of training data. In contrast, traditional AI methods offer more dependable but less flexible solutions, often limiting their broad applicability.

I am working in the intersection of these approaches which is sometimes called neuro-symbolic AI, particularly with knowledge graphs (KG). I investigate how graph neural networks can bridge the gap between discrete KGs and the statistical world of machine learning, enabling more robust and scalable solutions for tasks like structured and natural language question answering. My work often utilizes data from the medical and biomedical domains. During the seminar, I hope to discuss two sharpen my vision on the future of graph research. Among others, I am interesting in

1. Improving ML Trustworthiness: Many popular ML models are not trustworthy. Can the addition of graph data make a difference? Can we get to a point where we can provide formal guarantees (e.g., error bounds) to enhance trust in graph ML models?
2. Addressing issues in Multi-source KGs: There are challenges when attempting learning on unbalanced KGs, where a larger, potentially biased graph overshadows smaller, more specific ones. How could we overcome this?
3. Get to scalable graph ML models using both smart use of hardware and knowledge on what the content of the graph is.

### 3.4 From Graphs to Design Automation: Custom Systems for Data Analytics

*Antonino Tumeo (Pacific Northwest National Lab. – Richland, US)*

License  Creative Commons BY 4.0 International license  
© Antonino Tumeo

This talk discusses several co-design approaches to enable accelerated graph analytics across the whole continuum of computing. The talk first identifies the key issues in large scale graph processing, then reviews some algorithm-hardware co-design works to enable parallel graph processing on leadership computing systems with thousands of graphic processing units (GPUs). We then provide a full stack case study, the Graph Engine For Multithreaded Systems, which implements a scalable in-memory graph database for distributed high-performance computing clusters. The talk then briefly discusses the role of C++ distributed data structures libraries to design complex and malleable data models (including property graphs) and parallel (graph) algorithms. Finally, we discuss how to generate specialized accelerators for irregular applications and graph algorithms starting from shared memory parallel code descriptions with the SODA Synthesizer.

### 3.5 Two Graph-Related Topics

*M. Tamer Özsu (University of Waterloo, CA)*

License © Creative Commons BY 4.0 International license  
© M. Tamer Özsu

I talk about two topics one of which focuses on graph processing algorithms and techniques, and the other on an interesting use of graphs. The first topic is querying streaming graphs which is difficult because it combines two difficult issues, namely the unboundedness and high velocity of arrivals of streaming data and the difficulty of graph querying. In this space we systematically analyzed query models with clear semantics, developed streaming graph algebra primitives that allowed logical plan generation. Transformation rules were developed to manipulate the logical plans as well as a cost-based optimization framework to find the better plans. All of this was implemented as a prototype over Calcite.

The second topic is graph-based vector indexing. In this space, vectors that are generated by an embedding model form a graph dataset over which an approximate nearest neighbour search (ANNS) is performed. To aid search, indexes are built on this dataset. Graph-based indexes are those that treat each vector as a vertex with edges representing relationships between these vectors. The search takes a query point, starts from an entry point vertex in this index and finds the approximate nearest neighbours of the query point. These indexes can be built using either a refinement approach (e.g., NNDescent) or an iterative approach (e.g., HNSW). Refinement-based approaches start with a random graph (of vectors) and iteratively improve the graph. They have fast index construction time, but lower queries-per-second than iterative approaches and they cannot do incremental inserts of new vectors. On the other hand, iterative approaches insert vectors one-at-a-time, performing an ANNS to determine which other vectors it should be connected to. They have high construction cost, but good query performance and can handle incremental inserts. Our work involves an index, called MIRAGE, that combines the good characteristics of both approaches. It has better construction time and higher queries-per-second.

## 4 Working groups

### 4.1 Holistic graph operations at scale, across the digital continuum

*Andrea Bartolini (University of Bologna, IT), Duncan Bart (University of Twente – Enschede, NL), Dante Niewenhuis (VU Amsterdam, NL), Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI), Daniël ten Wolde (CWI – Amsterdam, NL), Antonino Tumeo (Pacific Northwest National Lab. – Richland, US), and Nikolay Yakovets (TU Eindhoven, NL)*

License © Creative Commons BY 4.0 International license  
© Andrea Bartolini, Duncan Bart, Dante Niewenhuis, Jože Rožanec, Daniël ten Wolde, Antonino Tumeo, and Nikolay Yakovets

**Introduction.** Currently, we lack of a holistic view on graph processing at scale, that would comprehend the whole digital continuum. Therefore, it is key to identify what siloes do currently exist and the reasons behind their persistence as to design mechanisms that would enable graph operations at scale across the digital continuum.

**Key ideas.** Current approaches are fragmented, and there is no single graph abstraction that can express both temporal and spatial metadata. Furthermore, graph workflows are siloed, with few mechanisms that enable interoperability and expressiveness. A gap also

exists between graph representations and how these could be mapped to hardware for efficient processing (e.g., mapping property graphs to GPUs is not trivial). A graph-first framework resembling a water cycle could be designed, considering (i) data streams and lakes deployed over a storage continuum and natively handling data bitemporality and associated with a semantic catalogue, (ii) mechanisms for scalable interfacing such as declarative queries that allow to express workflows and graph views, which are then translated by a (iii) compiler to stored procedures and executed on (iv) heterogeneous hardware across a compute continuum.

**Conclusions.** There is consensus that we need (a) a standard graph format that would reduce fragmentation and enable end-to-end graph processing, (b) a standardized representation as to allow to describe graph problems, (c) composable abstractions (frameworks that decouple graph modelling from performance concerns) providing a shared vocabulary between user needs and algorithm and system design, and (d) reimagine graph-processing hardware and systems (with a focus on sparse matrix computation and ensuring the storage and compute are co-located as to reduce the movement of data).

## 4.2 Applications at a real-world scale with near-term societal impact

*Aydin Buluc (Lawrence Berkeley National Laboratory, US), Duncan Bart (University of Twente – Enschede, NL), Stefania Dumbrava (ENSIIE & TélécomSudParis, FR), Kamesh Madduri (Pennsylvania State University – University Park, US), Dante Niewenhuis (VU Amsterdam, NL), Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI), Daniël ten Wolde (CWI – Amsterdam, NL), and Ana Lucia Varbanescu (University of Twente – Enschede, NL)*

**License** © Creative Commons BY 4.0 International license

© Aydin Buluc, Duncan Bart, Stefania Dumbrava, Kamesh Madduri, Dante Niewenhuis, Jože Rožanec, Daniël ten Wolde, and Ana Lucia Varbanescu

**Introduction.** To identify the future challenges and opportunities in graph processing, it is essential to identify what real-world applications with societal impact will look like. Understanding the requirements and characteristics of such applications helps identify the modeling and design choices required.

**Key ideas.** When developing applications at a real-world scale with near-term societal impact, it is crucial to consider how data is modeled into a graph, as this determines the approaches that become feasible for use and processing those graphs downstream. Key concerns related to (a) what methodologies are appropriate to extract data and create a graph, (b) how data should be represented in the graph – also addressing privacy and concerns where appropriate, (c) how graph systems should be designed to support multiple modeling choices, optimizations, and infrastructure capabilities without resulting in vendor lock-in, and (d) what approaches could be used for cleansing, completing, fixing, and processing those graphs. Promising application domains that could benefit from graph-based approaches include artificial data generation, healthcare, transportation, and logistics.

**Conclusions.** There is no perfect graph system. Data modeling choices, system optimizations, and infrastructure decisions depend heavily on application needs. Privacy-preserving modeling, federation, and distributed graph processing remain open research challenges. Future directions include better system/application co-design, reducing vendor lock-in, and making knowledge graph construction, processing, and graph-based applications development more accessible.



### 4.3 Adaptivity and optimization to ensure performance, scalability, and sustainability

*Stefania Dumbrava (ENSIIE & TélécomSudParis, FR), Duncan Bart (University of Twente – Enschede, NL), Peter A. Boncz (CWI – Amsterdam, NL), Florina M. Ciorba (Universität Basel, CH), Dante Niewenhuis (VU Amsterdam, NL), Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI), and Daniël ten Wolde (CWI – Amsterdam, NL)*

**License** © Creative Commons BY 4.0 International license

© Stefania Dumbrava, Duncan Bart, Peter A. Boncz, Florina M. Ciorba, Dante Niewenhuis, Jože Rožanec, and Daniël ten Wolde


**Introduction.** While performance, scalability, and sustainability are regarded as relevant, efforts in this direction remain fragmented, making it difficult to consolidate them in frameworks and systems that would benefit the end users.

**Key ideas.** Adaptivity and optimization are required to ensure performance, scalability, and sustainability. When speaking about scalability, we can consider weak scalability (scape up) or strong scalability (speed up), scale out (add more machines) or scale up (more resources inside a single machine). When contextualizing scalability w.r.t. graphs we consider scaling data (graph grows) and scaling workload (operations on graphs grow). When speaking about sustainability, we consider energy efficiency (brown, green energy), power consumption, carbon (embodied and operational), cost, required rare earth materials, and water consumption (for building and cooling). Finally, when considering performance, we can consider algorithmic or system performance, or performance related to productivity and usability of tools and systems. When considering algorithmic and system performance, we can characterize it with metrics that refer to execution time, response time, throughput, resource utilization, and performance portability. When considering performance from the productivity point of view, we can characterize it considering lines of code or external dependencies, but also through usability criteria, such as ease of deployment or ease of use of a given tool or system. When analyzing adaptivity and optimization to ensure performance, scalability, and sustainability, we observe that the graph ecosystem is fragmented. Many frameworks support only subsets of functionality and there is little convergence. That fragmentation is evident when considering supported functionality across graph tools and systems, where there is no single definition of correctness. Furthermore, such fragmentation makes it more difficult to select the right algorithm implementation (hardware and software) when considering complex algorithms.

**Conclusions.** To reduce fragmentation across the graph community, we propose focusing on standards rather than implementation details. Such standards could, in certain cases, replace the notion of correctness by specifying expectations over an algorithmic execution of system output. Furthermore, high-level abstractions, such as DSLs, could be useful to increase code portability and optimization across architectures. To ensure the best algorithms are available to everyone, it is necessary to promote best practices that provide algorithm and model implementations are usable beyond an academic endeavor. Furthermore, we require a community effort to catalog and benchmark algorithms, in order to understand their performance and available implementations. Artificial Intelligence can be leveraged to determine the optimal hardware and software implementation for a particular algorithm. We would ideally strive for models trained on small graphs whose inference can be extrapolated to large graphs. Explainable Artificial Intelligence can provide additional insights into understanding why a particular setup is suitable for a specific task.

#### 4.4 Graph processing algorithms and workflows

*Johannes Langguth (Simula Research Laboratory – Oslo, NO), Duncan Bart (University of Twente – Enschede, NL), Sanjukta Bhowmick (University of North Texas, US), Dante Niewenhuis (VU Amsterdam, NL), Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI), Ingo Scholtes (Universität Würzburg, DE), and Daniël ten Wolde (CWI – Amsterdam, NL)*

**License**  Creative Commons BY 4.0 International license

© Johannes Langguth, Duncan Bart, Sanjukta Bhowmick, Dante Niewenhuis, Jože Rožanec, Ingo Scholtes, and Daniël ten Wolde

**Introduction.** Graph processing and graph applications have similar but different concerns. Nevertheless, ensuring synergies between those communities can help advance the field of graph processing at a higher pace.

**Key ideas.** The application and graph processing communities are similar, but have different concerns. The main objective of the graph database community is data management, while applications focus on performance. This discrepancy makes it challenging for the communities to align. For example, people are familiar with GraphDBs, but cannot extract the information they need due to low performance. Additionally, many communities work on the same challenges without considering progress made on them in a different community. Instead, we should combine forces. However, interoperability between the different communities is non-trivial. Creating a complete overarching system that satisfies all communities is challenging due to the variety in needs.

**Conclusions.** We need to focus on improving usability for all users. Current systems lack intuitive abstractions, especially for non-experts. We need to combine forces and avoid addressing problems that might have been solved in a different community. Additionally, we should focus more on data quality. Data cleaning is still a widespread bottleneck, but is understudied. There is also a responsibility mismatch: data producers have little incentive to clean data and what does clean data mean could be different depending on the downstream requirements.

#### 4.5 Massive graph creation with generative and analytical approaches

*Nikolay Yakovets (TU Eindhoven, NL), Bogdan Arsintescu (Microsoft Corp. – Mountain View, US), Duncan Bart (University of Twente – Enschede, NL), Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI), Juan F. Sequeda (data.world – Austin, US), and Daniël ten Wolde (CWI – Amsterdam, NL)*

**License**  Creative Commons BY 4.0 International license

© Nikolay Yakovets, Bogdan Arsintescu, Duncan Bart, Jože Rožanec, Juan F. Sequeda, and Daniël ten Wolde

**Introduction.** Real-world data is oftentimes not directly obtainable in a graph shape/format. This data can be transformed into a graph using a schema or ontology. In some cases, graphs are needed that cannot be obtained from real-world data. In these cases, the graphs must be synthetically generated.

**Key ideas.** Graph generation can be divided into two categories. Creating graphs from existing data sources (with an inherent graph structure, like social networks, or other forms such as time series data) or synthetic graph generation. Creating graphs from existing data


is essential for real-world applications. A significant challenge in generating graphs from existing data sources is the data modeling problem, i.e., defining a suitable schema/ontology. Creating a good-quality graph is essential for the quality and reliability of graph applications. However, there is currently no clear definition of what constitutes a “good” graph, which also heavily depends on the specific application at hand. Synthetic graph generation is crucial for benchmarking or stress testing applications under immense computational loads, and it can also be critical in ensuring anonymity when handling sensitive data that could identify individuals, such as diseases that affect only a small number of people. Generating synthetic graphs that resemble real-world graphs for a specific use case is a significant challenge to solve.

**Conclusions.** A catalogue should be created with metrics and qualities that distinguish between “good” and “bad” graphs extracted from real data and ontologies. This could include credibility/traceability, entropy, or statistical properties (degree distribution, clustering, and path lengths, among others). We consider automatic ontology detection is a crucial research direction for streamlining graph generation from real-world data without human intervention. Generative AI could be a promising research direction for the generation of synthetic graphs.

## 5 Open problems

### 5.1 Open problems regarding Holistic Graph-Processing Systems to Enable Real-World Scale and Societal Impact

*Jože Rožanec (Jozef Stefan Institute – Ljubljana, SI)*

**License**  Creative Commons BY 4.0 International license  
© Jože Rožanec

The recent seminar successfully identified and delineated several open problems within the domain of holistic graph-processing systems. Addressing these challenges is critical for enabling systems that can operate at real-world scale and maximize societal impact. From the issues discussed, we have prioritized and selected five problems of scientific and engineering importance, which are briefly described below.

- **Application-centric graph usage matrix:** the goal is to establish a structured analytical framework to systematically characterize graph utilization across diverse application domains, drawing inspiration from the foundational work of Asanovic et al. [1]. This analysis must extend beyond a mere quantification of how intensely are graph technologies utilized and must rigorously assess usage across four critical dimensions: (i) graph generative approaches and graph analytics, (ii) graph processing algorithms and workflows, (iii) graph operations at scale across the digital continuum, and (iv) adaptivity and optimization to ensure performance, scalability, and sustainability. Furthermore, a quantitative characterization should be developed to map applications across the trade-off space defined by accuracy, performance, and portability. This will facilitate the creation of a workflow model to guide the selection of appropriate algorithms. The matrix will be complemented by a curated, diverse collection of exemplars: documented instances of successful algorithm-implementation pairs focusing on algorithmic building blocks that drive high performance, productivity, and portability within real-world application contexts.

- **Workload abstraction and scalable reference architecture:** building upon the insights derived from the application-centric graph usage matrix, the objective is to define formal workload abstractions (Basic Graph Operations). These Basic Graph Operations must be composable elements for the construction of complex algorithms, workflows, and scalable, interacting data processing pipelines. Furthermore, we propose the development of a reference architecture for scalable graph processing. This architecture must leverage the defined workload abstractions and directly address the needs identified by the application-centric graph usage matrix. The design must weight distributed computing architectures and single-node architectures, considering their respective strengths and weaknesses and educate how processing demands can be framed to achieve the desired outcomes.
- **The Memex Architecture: integrating graphs, metalearning, and causality:** we aim to lay the foundation for an architecture that would allow to build a holistic knowledge graph and leverage it for self-improvement. The knowledge graph will semantically link information regarding system execution processes, resource performance metrics, identified system anomalies, underlying infrastructure properties, and crucially, causal relationships among these elements. We envision that this architecture can be leveraged for metalearning, enabling the system to continuously self-optimize and improve its operational efficiency over time.
- **Hardware acceleration and graph compression techniques:** a systematic analysis is required to determine the short-term and long-term implications of emerging hardware accelerators on the efficiency of holistic graph operations. Concurrently, there is a necessity to catalogue and evaluate graph compression techniques that can be efficiently processed by these accelerators.
- **Scientific benchmarking standardization:** perceive a significant opportunity to establish a new generation of rigorous, scientifically grounded graph processing benchmarks. These benchmarks could be built upon the foundations laid by the application-centric graph usage matrix, the formal characterization of algorithms, and the capabilities of emerging accelerators. The primary goal is to introduce a transparent methodology for comparing and contrasting algorithms, datasets, processing workloads, and the heterogeneous underlying computing infrastructure used for their execution.

As an academic and professional community focused on advancing holistic graph-processing systems, we consider addressing these open problems is fundamental to realizing the next generation of scalable and sustainable systems capable of generating significant real-world and societal-scale impact.

## References

- 1 Asanovic, Krste and Bodik, Ras and Catanzaro, Bryan and Gebis, Joseph and Husbands, Parry and Keutzer, Kurt and Patterson, David and Plishker, William and Shalf, John and Williams, Samuel Webb. *The landscape of parallel computing research: A view from berkeley*. 2006

## Participants

- Bogdan Arsintescu  
Microsoft Corp. –  
Mountain View, US
- David A. Bader  
NJIT – Newark, US
- Duncan Bart  
University of Twente –  
Enschede, NL
- Andrea Bartolini  
University of Bologna, IT
- Sanjukta Bhowmick  
University of North Texas, US
- Peter A. Boncz  
CWI – Amsterdam, NL
- Angela Bonifati  
Lyon 1 University & IUF, FR
- Aydin Buluc  
Lawrence Berkeley National  
Laboratory, US
- Kuan-Hsun Chen  
University of Twente –  
Enschede, NL
- Florina M. Ciorba  
Universität Basel, CH
- Michael Cochez  
VU Amsterdam, NL
- Khuzaima Daudjee  
University of Waterloo, CA
- Stefania Dumbrava  
ENSIIE & TélécomSudParis, FR
- Aaron Eberhart  
metaphacts – Walldorf, DE
- Brian Elvesaeter  
SINTEF – Oslo, NO
- Reza Farahani  
Alpen-Adria-Universität  
Klagenfurt, AT
- Peter Haase  
Metaphacts – Walldorf, DE
- Kathrin Hanauer  
Universität Wien, AT
- Alexandru Iosup  
VU Amsterdam, NL
- Johannes Langguth  
Simula Research Laboratory –  
Oslo, NO
- Michelle Li  
Harvard University – Boston, US
- Andrew Lumsdaine  
Pacific Northwest National  
Laboratory – Seattle, US &  
University of Washington –  
Seattle, US
- Kamesh Madduri  
Pennsylvania State University –  
University Park, US
- Timothy G. Mattson  
Human Learning Group –  
Ocean Park, US
- Martin Molan  
Comtrade AI – Ljubljana, SI
- Dante Niewenhuis  
VU Amsterdam, NL
- M. Tamer Özsu  
University of Waterloo, CA
- Radu Prodan  
Alpen-Adria-Universität  
Klagenfurt, AT
- Matei Ripeanu  
University of British Columbia –  
Vancouver, CA
- Jože Rožanec  
Jozef Stefan Institute –  
Ljubljana, SI
- Ingo Scholtes  
Universität Würzburg, DE
- Daniel Thilo Schroeder  
SINTEF – Oslo, NO
- Juan F. Sequeda  
data.world – Austin, US
- Daniël ten Wolde  
CWI – Amsterdam, NL
- Antonino Tumeo  
Pacific Northwest National Lab. –  
Richland, US
- Ana Lucia Varbanescu  
University of Twente –  
Enschede, NL
- Laurentiu Vasiliu  
Peracton – Galway, IE
- Hannes Voigt  
Neo4j – Leipzig, DE
- Nikolay Yakovets  
TU Eindhoven, NL

