



DFG Priority Programme SPP 1736: Algorithms for Big Data

Mahyar Behdju¹ · Ulrich Meyer¹

Published online: 12 December 2017
© The Author(s) 2017. This article is an open access publication

Abstract

Volume, Velocity, and Variety are the three Vs commonly used to define the term *big data*. Simply put, those refer to the increasing amount of new data created, the increasing rate at which it is created, and the increasing number of different formats it has. At the same time, the three Vs describe challenges that require new algorithmic approaches. In order to tackle those challenges, the German Research Foundation established in 2013 the priority programme *SPP 1736: Algorithms for Big Data*. In this article we give a short overview on the research topics represented within this priority programme.

Keywords Algorithms · Applications · Big data · Combinatorial optimization · Graphs and networks

1 Introduction

Computer systems pervade all parts of human activity: transportation systems, energy supply, medicine, the whole financial sector, and modern science have become unthinkable without hardware and software support. As these systems continuously acquire, process, exchange, and store data, we live in a big-data world where information is accumulated at an exponential rate.

The urging problem has shifted from collecting enough data to dealing with its impetuous growth and abundance. In particular, data volumes often grow faster than the transistor budget of computers as predicted by Moore's law (i.e., doubling every 18 months). On top of this, we cannot any longer rely on transistor budgets to automatically translate into application performance, since the speed improvement of single processing cores has basically stalled and the requirements of algorithms that use the full memory hierarchy get more and more complicated. As a result, algorithms have to be massively parallel using memory access patterns with high locality. Furthermore, an x -times machine

performance improvement only translates into x -times larger manageable data volumes if we have algorithms that scale nearly linearly with the input size. All these are challenges that need new algorithmic ideas. Last but not least, to have maximum impact, one should not only strive for theoretical results, but intend to follow the whole *algorithm engineering* development cycle consisting of theoretical work followed by experimental evaluation.

The “curse” of big data in combination with increasingly complicated hardware has reached all kinds of application areas: genomics research, information retrieval (web search engines, ...), traffic planning, geographical information systems, or communication networks. Unfortunately, most of these communities do not interact in a structured way even though they are often dealing with similar aspects of big-data problems. Frequently, they face poor scale-up behavior from algorithms that have been designed based on models of computation that are no longer realistic for big data.

In 2013, the German Research Foundation (DFG) established the priority programme *SPP 1736: Algorithms for Big Data* (<https://www.big-data-spp.de>) where researchers from theoretical computer science work together with application experts in order to tackle some of the problems discussed above. A nationwide call for the individual projects attracted over 40 proposals out of which an international reviewer panel selected 15 funded research projects plus a coordination project (totalling about 20 full PhD student positions) by the end of 2013. Additionally, a few more projects with own funding have been associated in order to benefit from

The research presented in this article has been partially supported by the DFG coordination funds ME 2088/4- {1,2}.

✉ Mahyar Behdju
mahyar@ae.cs.uni-frankfurt.de
Ulrich Meyer
umeyer@ae.cs.uni-frankfurt.de

¹ Goethe University Frankfurt, Frankfurt, Germany

collaboration and joint events (workshops, PhD meetings, summer schools etc.) organised by the SPP.

In the following, we give a short overview on the research topics and groups represented in the programme and highlight a few results obtained within the first funding period (2014–2017). Two project leaders also contributed separate articles within this special issue: H. Bast on *a quality evaluation of combined search on a knowledge base and text*, and M. Mnich on *big data algorithms beyond machine learning*.

2 Funded Research Projects

Most of the funded projects concentrate on big-data algorithms that are not machine learning and frequently tackle more than one of the following areas: (1) technological challenges, (2) fundamental algorithmic techniques, and (3) applications. There are various ways the respective research topics could be clustered—here is one attempt:

2.1 Technological Challenges

Several projects are concerned with algorithmically mastering constraints in the way data can be efficiently accessed, compactly maintained, and processed in parallel. Besides mere execution time and solution quality, further metrics such as energy consumption and limited data lifetime come into play:

P1 *Energy-Efficient Scheduling* S. Albers (TU München)

This project explores methods to reduce the total energy consumption using scheduling, based on speed-scalable processors where typically much less energy is consumed if the processors run slower. While jobs come with deadlines (hence slowing down is not always possible), preemption and migration of jobs open up additional optimization opportunities. The scheduling objective is to minimize the total energy consumption while taking into account all constraints. Albers et al. considered non-homogeneous settings where trade-offs between speed and energy can differ among the set of processors used [2]. The authors improve the state of the art by providing several new approaches that are conceptually easier and hence more practical than previous solutions.

P2 *Dynamic, Approximate, and Online Methods for Big Data* U. Meyer (U Frankfurt/M)

One line of research in the project deals with dynamic, approximate, and online methods in the context of parallelism and memory hierarchies. An important application area is graph algorithms (see [22] and Sect. 3 for

a more detailed treatment of joint results on graph generation in parallel external memory). Another line of research in (P2) aims to use methods from Game Theory (truthful mechanisms) in order to reasonably solve memory assignment problems for concurrently running programmes in shared memory environments. This is particularly challenging if the users do not have to pay money for the RAM their executed programmes occupy: in the absence of money, selfish programmers may claim to need unreasonably large chunks of central memory for a “fast” execution of their programmes. First results in a static setting with fixed RAM chunk sizes appeared in [18], where forced waiting times are used as a currency in order to yield a truthful mechanism that returns solutions minimizing the makespan, i.e. the maximum completion time.

P3 *Distributed Data Streams in Dynamic Environments* F. Meyer auf der Heide (U Paderborn)

The research topic of this project is the design and analysis of distributed algorithms that continuously compute functions of streams of data, arising from many devices of potentially different types. Due to huge volumes and velocity, data can neither be completely stored, nor sent to a central server via a network, nor fully processed in real time. Initial results concern, among others, the communication complexity of so-called distributed aggregation problems; Mäcker et al. [20] considered the expected message complexity for the top- k POSITION MONITORING problem. Here, the task is to compute the IDs of the devices that observe the k largest items at every time step. They also gave an approximation variant [21].

P4 *3D+T Terabyte Image Analysis*¹ R. Mikut and P. Sanders (KIT Karlsruhe)

The data dealt with in this project stems from light-sheet fluorescence microscopy, which is frequently applied in developmental biology in order to perform long-time observations of embryonic development. An exemplary application domain is tracking of objects (e.g., cell nuclei, cytoplasm, nano particles) in microscopic images where different object classes are labelled with particular fluorescent dyes. In the project, time series of high resolution 3D images of developing zebrafish embryos yield more than 10 terabytes per embryo, which is significantly more than state-of-the-art software tools can typically handle. While striving

¹ associated project.

for improved algorithms on modern hardware, it is also important to carefully test the result quality of these new approaches. To this end, the project has successfully investigated methods to create large, realistic, simulated inputs with ground truth that can be used to quantitatively assess result quality [30].

P5 *Kernelization for Big Data (See footnote 1)* M. Mních (U Bonn)

The Project is concerned with kernelization in big-data contexts. Given a concrete optimization question q , a kernelization algorithm compresses a data set A to A' such that q can still be answered from A' . Ideally, the size of A' is much smaller than that of A and depends only polynomially on some structures capturing particular aspects about the optimization question (and not on the size of A). As an example, Etscheid and Mních discussed kernelization techniques for the MAX-CUT problem [14]; more details are provided in Mních's article on *big data algorithms beyond machine learning* in this special issue.

2.2 Graphs

Another cluster of projects is mainly concerned with various kinds of graph problems which become very challenging once the input data is really big:

P6 *Skeleton-based Clustering in Big and Streaming Social Networks* U. Brandes (U Konstanz) and D. Wagner (KIT Karlsruhe)

The scientific goal of the project is to devise novel methods to cluster large-scale static and dynamic online social networks. Their approach is based on skeleton structures, i.e. sparse (sub-)graphs, that represent the essential structural properties of the graphs. Besides supporting efficient clustering approaches, these skeletons are used to find patterns in online social relationships and interactions. An example concerns components of *quasi-threshold graphs*² since they share features frequently found in social network communities. Communities are then detected by finding a quasi-threshold graph that is close to a given graph in terms of edge edit distance. The problem is \mathcal{NP} -hard and existing FPT-approaches also fail to scale on real-world data. Hence, the project introduced *Quasi-Threshold Mover* (QTM), the first scalable quasi-threshold editing heuristic [9]. QTM constructs an initial skeleton forest and

then refines it by moving vertices to reduce the number of edits required. (P6) is also active in graph visualization and graph generation (cf. Sect. 3).

P7 *Engineering Algorithms for Partitioning Large Graphs (See footnote 1)* P. Sanders, Ch. Schulz, and D. Wagner (KIT Karlsruhe)

(Hyper-)Graph partitioning is crucial in many big-data graph applications as it subdivides the problem instance into smaller (and thus more manageable) pieces with little interaction. Unfortunately, HYPERGRAPH PARTITIONING is \mathcal{NP} -hard, and it is even \mathcal{NP} -hard to obtain good approximations. Therefore, in practice, multi-level heuristics are applied. Project (P7) has significantly contributed to the large body of previous work in the area; see [10] for an overview. Their recent k -way partitioning result [1] represents the state of art concerning high-quality hypergraph partitioning: it always computes better solutions and is faster than some of the competitors.

P8 *Competitive Exploration of Large Networks* Y. Disser (TU Darmstadt) and M. Klimm (HU Berlin)

This project looks into algorithms that operate on very large networks and the dynamics that arise from the competition or the cooperation between such algorithms. An initial result concerned the exploration of an unknown undirected graph with n vertices by an agent possessing very small memory [12]. While upper and lower memory bounds of $\Theta(\log n)$ had been shown before for this setting, the project reduced the memory requirement of the agent to $O(\log \log n)$ for bounded-degree graphs in case the agent gets access to another $O(\log \log n)$ indistinguishable markers, called *pebbles*. A pebble can be dropped or collected whenever the agent visits a vertex, leaving or removing a mark. (P8) also showed that for sub-linear agent memory, $\Omega(\log \log n)$ pebbles are required.

P9 *Algorithms for Solving Time-Dependent Routing Problems with Exponential Output Size* M. Skutella (TU Berlin)

Methods for the solution of static routing problems have been successfully optimized over many decades. Unfortunately, real-life applications such as evacuation planning, logistic planning, or navigation systems for road networks crucially depend on dynamic edge costs that change over time (and even depend on the solution). The standard approach to build a huge time-expanded network whose size could be exponential in the input size becomes infeasible for big-data graphs due

² which neither contain paths on four vertices nor cycles on four vertices.

to memory limitation. Hence, project (P9) investigate alternative methods that try to avoid this data explosion. For example, Schlöter and Skutella presented memory-efficient solutions for evacuation problems [27].

- P10 *Local Identification of Central Nodes, Clusters, and Network Motifs in Very Large Complex Networks* K. Zweig (TU Kaiserslautern)

This Project focuses on the development of *local* methods to compute classic network analytic measures like centrality indices, network motifs (subgraphs) and clustering. Commonly, these measures are based on *global* properties of the graph such as the distance between all pairs of vertices or a global ranking of similar pairs of vertices or edges, thus, resulting in at least quadratic time complexity. Hence, it is difficult to scale those fundamental approaches directly to big-data graphs. Recent work in this direction concerns the identification of network motifs in the so-called *fixed degree sequence model* (FDSM) which refers to the set of all graphs with the same degree sequence excluding multi-edges or self-loops. Schlauch and Zweig proposed a set of equations, based on the degree sequence and a simple independence assumption, to estimate the occurrence of a set of subgraphs in the FDSM and empirically supported their findings [26]. Other parts of the research in this project have also been included in a newly published textbook [33].

2.3 Optimization

Two projects concentrate on generic optimization methods that can be applied in many different scenarios:

- P11 *Scaling Up Generic Optimization* J. Giesen and S. Laue (U Jena)

Dealing with large-scale convex optimization problems, the project developed a generic optimization code generator (*GENO*) which is capable of providing generic, parallel and distributed convex optimization software. Discrete and combinatorial big-data optimization problems can greatly benefit from *GENO* as well as machine learning, data analytics and other fields of research such as network analysis. The *GENO* approach to generic optimization is based on an extension of the alternating direction method of multipliers by Giesen and Laue [16] and is defined by a tight coupling of a modeling language and a generic solver. The modeling language allows to specify a class of (convex) optimization problems, and the generic solver gets instantiated for the specified problem class. Comparing the code produced by *GENO* with state-of-the-art, hand-tuned,

problem-specific implementations show that *GENO* is faster and delivers better results (in terms of accuracy or objective function value for non-convex problems).

- P12 *Fast Inexact Combinatorial and Algebraic Solvers for Massive Networks* H. Meyerhenke (U Köln)

This project focuses on network analysis with three combinatorial optimization tasks with numerous applications: graph clustering, graph drawing, and network flow. Some of those applications are in the biological sciences, where most data sets are massive and contain inaccuracies. Hence, an inexact, yet faster solution process with approximation algorithms and heuristics is often useful. As an example, Bergamini and Meyerhenke [5] proposed the first betweenness centrality approximation algorithms with a provable bound on the approximation error for fully dynamic networks. Another important topic dealt with in (P12) concerns algebraic solvers. In 2016, Bergamini et al. [6] developed two algorithms that accelerate the current-flow computation for one vertex or a reasonably small subset of vertices significantly. The work also provides a reimplementation of the *lean algebraic multigrid solver* by Livne and Brandt [19] and is integrated into the open-source network analysis software *NetworKit* [28], which is freely available to the public.

2.4 Security

Further projects investigate practical cryptographic schemes that do not degrade in big-data contexts, for example when the number of users and ciphertexts grows tremendously:

- P13 *Security-Preserving Operations on Big Data* M. Fischlin (TU Darmstadt) and A. May (U Bochum)

Protecting outsourced data in cloud storage and cloud computing scenarios and when handling big data through third parties is rather complicated, since standard cryptographic means, such as encryption, in general do not work here. This is caused by the very nature of encryption: scrambling all reasonable information, the semantics of the data are hidden and cannot be used by third parties to perform operations, and the option of decrypting the data for the operations would violate the idea of protecting the data from the service provider. Thus, project (P13) works on efficient operations on secured data, targeted as well as through the deployment of functional encryption and indistinguishable obfuscation, certification of cryptographic primitives, and new algorithmic techniques for big cryptographic data. In 2017, Esser et al. [13] proposed new algorithms with small memory consumption for the *LEARNING PAR-*

ITY WITH NOISE (LPN) problem, both classically and quantumly. By using different advanced techniques they obtained a hybrid algorithm that achieves the best currently known run time for any fixed amount of memory.

P14 Scalable Cryptography D. Hofheinz (KIT Karlsruhe) and E. Kiltz (U Bochum)

As mentioned before, in our modern digital society, we rely on encryption and signature schemes for security. However, today's cryptographic schemes do not scale well, and thus are not suited for the increasingly large sets of data they are used on. For instance, the security guarantees currently known for RSA encryption, which is an important type of encryption scheme, degrade linearly in the number of users and cipher texts. Therefore, project (P14) aims to construct cryptographic schemes that scale well to large scenarios. Until now, several practical cryptographic schemes suitable for truly large settings have been developed by the project, such as the first authenticated key exchange protocol whose security does not degrade with an increasing number of users or sessions [3], the first identity-based encryption scheme whose security properties do not degrade in the number of ciphertexts, and the first public-key encryption scheme for large scenarios that does not require a mathematical pairing [15] (awarded the “Best Paper” at the EUROCRYPT 2016 conference). The last-mentioned scheme is solely based upon a very standard computational assumption, namely the *Decisional Diffie-Hellman assumption*, and is thereby efficient.

2.5 Text Applications

In spite of huge improvements over the last decade, efficiently mining big text data remains an important topic:

P15 Efficient Semantic Search on Big Data H. Bast (U Freiburg)

Within the predecessor priority programme on *algorithm engineering* (2007–2013), H. Bast and her group have developed *semantic full-text search*, a deep integration of full-text and ontology search. Their search engine *Broccoli* [4] is able to handle queries like “Astronauts who walked on the moon and who were born in 1925–1930” or “German researchers who work on algorithms” where parts of the required information is contained in ontologies, whereas other parts only occur in text documents. In (P15), they aim to scale semantic search to text sets and ontologies being about 100 times larger than in their original *Broccoli* engine while increasing query quality at the same time. More details can be found in

their article on *a quality evaluation of combined search on a knowledge base and text* included in this special issue.

P16 Massive Text Indices J. Fischer (TU Dortmund) and P. Sanders (KIT Karlsruhe)

The world wide web, digital libraries, biological sequences like DNA, or proteins all constitute large textual data that need to be stored, structured, searched, and compressed efficiently. The amount of such data has grown much faster than the storage and computation capacities of common desktop computers, by several orders of magnitude. Data structures for texts satisfying those needs, for instance suffix arrays or inverted indexes, are called text indexes and are the basic building block of all text-based applications, including well-known services like internet search engines. Since algorithms and data structures for texts are fundamentally different from those for other kind of data, project (P16) aims to develop an own algorithmic toolbox for large texts using both shared memory and distributed memory parallelism, focusing on general-purpose text indexes related to suffix arrays due to their applicability to any text type and their extended functionality. One step towards that goal was basic research on building blocks, yielding results, such as an extensive journal paper [8] that studies practical parallel string sorting algorithms based on the most important classical sorting algorithms. Another important step was to build a prototype of a tool for implementing algorithms that process large data sets on distributed memory machines. The result, *Thrill* [7], is based on C++, offers a rich set of operations on distributed arrays such as map, reduce, sort, merge, and prefix-sum. It can fuse pipelines of local operations into tight loops optimized at compile time, considerably outperforming established tools such as *Spark* or *Flink*.

2.6 Bio Applications

Similarly, new methods in bioinformatics are required as reduced costs for obtaining raw data is resulting in a data flood that becomes increasingly hard to process:

P17 Graph-Based Methods for Rational Drug Design O. Koch and P. Mutzel (TU Dortmund)

The development of a new drug is a complex and costly process. The identification and optimization of bioactive molecules is to a large extent supported by computer-based methods, e.g. the semi-automated classification of molecules and their functional relationships. This is a big-data problem, since the theoretical chemi-

cal space is estimated to contain around 10^{62} molecules. Many approaches within rational drug design are based on the basic hypothesis that structural similar molecules also show a similar biological effect. Common similarity measures use fast but inexact chemical fingerprints, yielding a high number of false positives. By proposing the MAXIMUM SIMILAR SUBGRAPH (MSS) paradigm, an extension of the \mathcal{NP} -complete MAXIMUM COMMON SUBGRAPH problem with allowed deviations with respect of similar bioactivity, the project (P17) introduced an exact comparison method based on searching and clustering graph representations of molecules, where atoms are the vertices and bonds between the atoms are represented by edges. In 2017, Schäfer and Mutzel presented *StruClus* [25], a structural clustering algorithm for large-scale datasets of small labeled graphs based on the MSS paradigm. This algorithm achieves high quality and (human) interpretable clusterings, has a runtime linear in the number of graphs and outperforms competing clustering algorithms. The project also continuously develops *Scaffold Hunter* [24], a flexible visual analytics framework for the analysis of chemical compound data. One application of this tool is to identify whole scaffolds that are exchangeable by similar shape. This leads to a reduced graph which allows for a more efficient MSS computation. *Scaffold Hunter* was initiated as a collaboration with the group of H. Waldmann (Max Planck Institute of Molecular Physiology, Dortmund).

P18 *Algorithmic Foundations for Genome Assembly*
A. Srivastav (U Kiel), Th. Reusch (GEOMAR Kiel),
and Ph. Rosenstiel (Uniklinikum Schleswig-Holstein)

This is a joint project between A. Srivastav from the Department of Computer Science of Kiel University, T. Reusch from GEOMAR Helmholtz Centre for Ocean Research Kiel and Ph. Rosenstiel from the Institute of Clinical and Molecular Biology University Medical Center Schleswig-Holstein, dealing with the genome assembly problem: given a high number of sequences of an unknown genome, called *reads*, which may contain errors, and perhaps some extra information, the task is to reconstruct the genome. The objectives of (P18) are the development of a comprehensive mathematical model for genome assembly as an optimization problem, the engineering and theoretical analysis of distributed and streaming assemblers and distributed probabilistic data structures to hold intermediate information, the engineering of an assembler based on the *maximum-likelihood* method, and applications to marine species investigated in the group of Th. Reusch and to the variational calling problems in the group of Ph. Rosenstiel.

In 2017, Wedemeyer et al. [32] presented their read filtering algorithm *Bignorm*. They show how probabilistic data structures and biological parameters can be used to drastically reduce the amount of data prior to the assembly process and demonstrate its significance by the assembly of genomes of single-celled species.

2.7 Coordination Project

In addition to the research projects mentioned above, there is also a coordination project headed by U. Meyer. This project provides financial and organizational support for yearly colloquia of the whole priority programme, summer schools and smaller dedicated workshops and trainings, a guest programme, and gender equality measures. It also maintains the webpage of the priority programme <https://www.big-data-spp.de/>.

3 Scientific Output and SPP collaborations

During the first funding period, the SPP did not only publish more than 150 peer-reviewed papers, but also developed, extended and maintained a number of software libraries, e.g.: *Broccoli* [4] for semantic search, *GENO*³ for generic optimization code generation, *NetworKit* [28] for network analysis, *STXXL* [11] for external-memory computing, and *Thrill* [7] for distributed batch data processing. The priority programme also creates visibility by its national and international events (e.g., Summer/Winter schools in Chennai 2016 and Tel Aviv 2017).

A particular feature of a priority programme is the intended collaboration between its participating researchers. The efficient generation of huge artificial input graphs for benchmarking turned out to become a highly active field of joint research: already more than ten papers in this area have been published by SPP members (see [22] for a recent overview), out of which three are co-authored between different SPP projects: (P11) and (P12) consider faster generation of random hyperbolic graphs [31], (P6) and (P12) propose how to generate scaled replicas of real-world complex networks [29], and (P6) and (P2) give improved generation algorithms for random graphs according to FDSM.

Examples for other joint publications include sparsification methods for social networks [17] by (P6) and (P12), and improved parallel graph partitioning for complex networks [23] by (P7) and (P12).

The second funding period for the Big Data priority programme has just started and most of the projects reviewed above also belong to the consortium of the second phase.

³ <http://www.geno-project.org/>

Hence, we expect a number of further scientific results due to these established cooperations.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Akhremtsev Y, Heuer T, Sanders P, Schlag S (2017) Engineering a direct k -way hypergraph partitioning algorithm. In: Fekete SP, Ramachandran V (eds) Proceedings of the Nineteenth Workshop on Algorithm Engineering and Experiments, ALENEX 2017, Barcelona, Spain, Hotel Porta Fira, January 17–18, 2017, pp 28–42. SIAM
- Albers S, Bampis E, Letsios D, Lucarelli G, Stotz R (2016) Scheduling on power-heterogeneous processors. In: Kranakis E, Navarro G, Chávez E (eds) LATIN 2016: Theoretical Informatics—12th Latin American Symposium, Ensenada, Mexico, April 11–15, 2016, Proceedings, vol 9644 of Lecture Notes in Computer Science, pp 41–54. Springer
- Bader C, Hofheinz D, Jager T, Kiltz E, Li Y (2015) Tightly-secure authenticated key exchange. In: Dodis Y, Nielsen JB (eds) Theory of Cryptography—12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23–25, 2015, Proceedings, Part I, vol 9014 of Lecture Notes in Computer Science, pp 629–658. Springer
- Bast H, Bäurle F, Buchhold B, Haußmann E (2014) Semantic full-text search with broccoli. In: Geva S, Trotman A, Bruza P, Clarke CLA, Järvelin K (eds) The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia—July 06–11, 2014, pp 1265–1266. ACM
- Bergamini E, Meyerhenke H (2016) Approximating betweenness centrality in fully dynamic networks. *Internet Math* 12(5):281–314
- Bergamini E, Wegner M, Lukarski D, Meyerhenke H (2016) Estimating current-flow closeness centrality with a multigrid laplacian solver. In: Gebremedhin AH, Boman EG, Uçar B (eds) 2016 Proceedings of the Seventh SIAM Workshop on Combinatorial Scientific Computing, CSC 2016, Albuquerque, New Mexico, USA, October 10–12, 2016, pp 1–12. SIAM
- Bingmann T, Axtmann M, Jöbstl E, Lamm S, Nguyen HC, Noe A, Schlag S, Stumpp M, Sturm T, Sanders P (2016) Thrill: high-performance algorithmic distributed batch data processing with C++. In: Joshi J, Karypis G, Liu L, Hu X, Ak R, Xia Y, Xu W, Sato A, Rachuri S, Ungar LH, Yu PS, Govindaraju R, Suzumura T (eds) 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5–8, 2016, pp 172–183. IEEE
- Bingmann T, Eberle A, Sanders P (2017) Engineering parallel string sorting. *Algorithmica* 77(1):235–286
- Brandes U, Hamann M, Strasser B, Wagner D (2015) Fast quasi-threshold editing. In: Bansal N, Finocchi I (eds) Algorithms-ESA 2015—23rd Annual European Symposium, Patras, Greece, September 14–16, 2015, Proceedings, vol 9294 of Lecture Notes in Computer Science, pp 251–262. Springer
- Buluç A, Meyerhenke H, Safro I, Sanders P, Schulz C (2016) Recent advances in graph partitioning. In: Kliemann L, Sanders P (eds) Algorithm Engineering—Selected Results and Surveys, vol 9220 of Lecture Notes in Computer Science, pp 117–158
- Dementiev R, Kettner L, Sanders P (2008) STXXL: standard template library for XXL data sets. *Softw Pract Exper* 38(6):589–637
- Disser Y, Hackfeld J, Klimm M (2016) Undirected graph exploration with $\theta(\log \log n)$ pebbles. In: Krauthgamer R (ed) Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10–12, 2016, pp 25–39. SIAM
- Esser A, Kübler R, May A (2017) LPN decoded. In: Katz J, Shacham H (eds) Advances in Cryptology—CRYPTO 2017—37th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 20–24, 2017, Proceedings, Part II, vol 10402 of Lecture Notes in Computer Science, pp 486–514. Springer
- Etscheid M, Mnich M (2016) Linear kernels and linear-time algorithms for finding large cuts. In: Hong S (ed) 27th International Symposium on Algorithms and Computation, ISAAC 2016, December 12–14, 2016, Sydney, Australia, vol 64 of LIPIcs, pp 31:1–31:13. Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik
- Gay R, Hofheinz D, Kiltz E, Wee H (2016) Tightly cca-secure encryption without pairings. In: Fischlin M, Coron J (eds) Advances in Cryptology—EUROCRYPT 2016—35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8–12, 2016, Proceedings, Part I, vol 9665 of Lecture Notes in Computer Science pp 1–27. Springer
- Giesen J, Laue S (2016) Distributed convex optimization with many convex constraints. *CoRR*, abs/1610.02967
- Hamann M, Lindner G, Meyerhenke H, Staudt CL, Wagner D (2016) Structure-preserving sparsification methods for social networks. *Social Netw Anal Mining* 6(1):22:1–22:22
- Kovács A, Meyer U, Ventre C (2015) Mechanisms with monitoring for truthful RAM allocation. In: Markakis E, Schäfer G (eds), Web and Internet Economics—11th International Conference, WINE 2015, Amsterdam, The Netherlands, December 9–12, 2015, Proceedings, vol 9470 of Lecture Notes in Computer Science, pp 398–412. Springer
- Livne OE, Brandt A (2012) Lean algebraic multigrid (LAMG): fast graph laplacian linear solver. *SIAM J Sci Comput* 34(4):B499–B522
- Mäcker A, Malatyali M, Meyer auf der Heide F (2015) Online top-k-position monitoring of distributed data streams. In: 2015 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2015, Hyderabad, India, May 25–29, 2015, pp 357–364. IEEE Computer Society
- Mäcker A, Malatyali M, Meyer auf der Heide F (2016) On competitive algorithms for approximations of top-k-position monitoring of distributed streams. In: 2016 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2016, Chicago, IL, USA, May 23–27, 2016, pp 700–709. IEEE Computer Society
- Meyer U, Penschuck M (2017) Large-scale graph generation and big data: an overview on recent results. *Bull EATCS* 122
- Meyerhenke H, Sanders P, Schulz C (2017) Parallel graph partitioning for complex networks. *IEEE Trans Parallel Distrib Syst* 28(9):2625–2638
- Schäfer T, Kriege N, Humbeck L, Klein K, Koch O, Mutzel P (2017) Scaffold hunter: a comprehensive visual analytics framework for drug discovery. *J. Cheminf* 9(1):28:1–28:18
- Schäfer T, Mutzel P (2017) Struclus: scalable structural graph set clustering with representative sampling. In: Cong G, Peng W, Zhang WE, Li C, Sun A (eds) Advanced Data Mining and Applications—13th International Conference, ADMA 2017, Singapore, November 5–6, 2017, Proceedings, vol 10604 of Lecture Notes in Computer Science, pp 343–359. Springer

26. Schlauch WE, Zweig KA (2016) Motif detection speed up by using equations based on the degree sequence. *Social Netw Anal Min* 6(1):47:1–47:20
27. Schlöter M, Skutella M (2017) Fast and memory-efficient algorithms for evacuation problems. In: Klein PN (ed) *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16–19*, pp 821–840. SIAM
28. Staudt C, Sazonovs A, Meyerhenke H (2016) Networkit: a tool suite for large-scale complex network analysis. *Netw Sci* 4(4):508–530
29. Staudt CL, Hamann M, Safro I, Gutfraind A, Meyerhenke H (2016) Generating scaled replicas of real-world complex networks. In: Cherifi H, Gaito S, Quattrociocchi W, Sala A (eds) *Complex Networks and Their Applications V - Proceedings of the 5th International Workshop on Complex Networks and their Applications (COMPLEX NETWORKS 2016), Milan, Italy, November 30–December 2, 2016*, vol 693 of *Studies in Computational Intelligence*, pp 17–28. Springer
30. Stegmaier J, Arz J, Schott B, Otte JC, Kobitski A, Nienhaus GU, Strähle U, Sanders P, Mikut R (2016) Generating semi-synthetic validation benchmarks for embryomics. In: *13th IEEE International Symposium on Biomedical Imaging, ISBI 2016, Prague, Czech Republic, April 13–16, 2016*, pp 684–688. IEEE
31. von Looz M, Özdayi MS, Laue S, Meyerhenke H (2016) Generating massive complex networks with hyperbolic geometry faster in practice. In: *2016 IEEE High Performance Extreme Computing Conference, HPEC 2016, Waltham, MA, USA, September 13–15, 2016*, pp 1–6. IEEE
32. Wedemeyer A, Kliemann L, Srivastav A, Schielke C, Reusch TB, Rosenstiel P (2017) An improved filtering algorithm for big read datasets and its application to single-cell assembly. *BMC Bioinf* 18(1):324
33. Zweig KA (2016) *Network analysis literacy—a practical approach to the analysis of networks*. *Lecture notes in social networks*. Springer, Wien