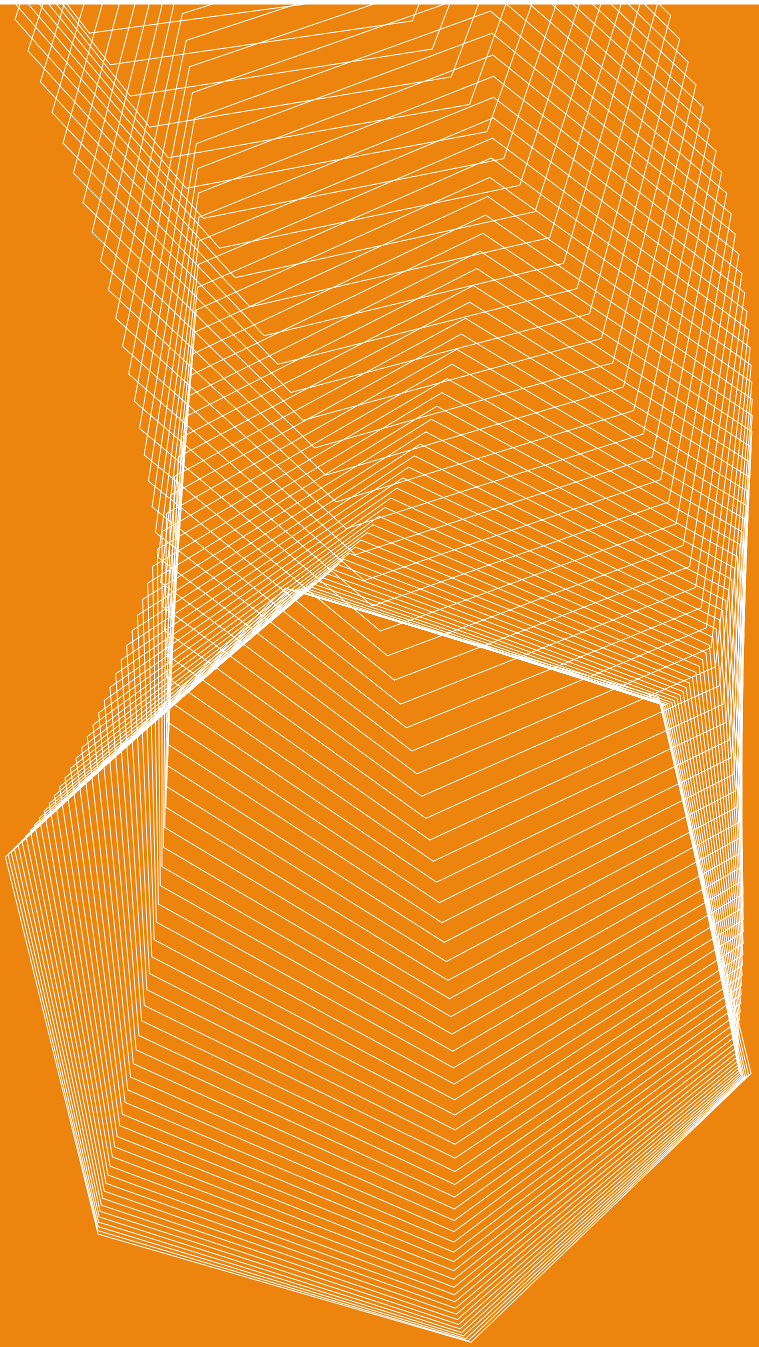


30th International Conference on Scientific and Statistical Database Management

SSDBM 2018

July 9–11

SPONSORS AND LOCAL ORGANIZERS

Gold Sponsor	Bronze Sponsor	Bronze Sponsor	Bronze Sponsor	Partner
WÜRTHPHOENIX	systems Building IT Success	EOS SOLUTIONS A KUMAVISION GROUP COMPANY	alpin streamline your business	IDM SÜDTIROL ALTO ADIGE

SSDBM 2018

unibz

Fakultät für Informatik
Facoltà di Scienze e Tecnologie informatiche
Faculty of Computer Science

WELCOME MESSAGE

The SSDBM 2018 organizing committee is pleased to welcome you to the **30th International Conference on Scientific and Statistical Database Management**, which takes place at the Free University of Bozen-Bolzano, July 9-11, 2018.

This year, SSDBM celebrates its 30th anniversary. By and large the success of the conference series is due to the continuous endorsement and commitment of Arie Shoshani (Lawrence Berkeley National Laboratory). Arie initiated the conference 38 years ago and since then has been serving as the steering committee chair, keeping the conference going and establishing the data-driven approach as a new paradigm of scientific discovery. We would like to express our special thanks to Arie for his tireless support of SSDBM during the past 38 years. At the conference Arie will have the opportunity to look back and share his thoughts about past, present and future of SSDBM.

SSDBM 2018 received 75 research paper submissions and 7 demonstration proposals. We would like to thank all authors for submitting outstanding contributions to the conference, which is vital for the success of the conference. The submissions were carefully evaluated by the program committee, which consisted of 42 members for regular research papers and 6 members for demonstration proposals. A warm thanks to the program committee members for their high-quality and timely handling of all reviews and discussions. This community service requires a lot of work on a tight schedule, and the quality of this work is the foundation of the continued success of SSDBM. Thanks to this effort we can look forward to an exciting program and attractive SSDBM conference in Bolzano. The program includes 23 full-length research papers, 5 short papers that will be presented as posters, and 5 demonstrations.

The SSDBM scientific program will be complemented by two keynotes given by Christian S. Jensen (Data-Intensive Vehicle Routing) and David Maier (Are Green Buildings Healthy?). Christian S. Jensen is Obel Professor of Computer Science at Aalborg University, Denmark. His research concerns data-intensive systems, focusing on temporal and spatio-temporal data management. He is a member of Academia Europaea, the Royal Danish Academy of Sciences and Letters, and the Danish Academy of Technical Sciences. David Maier is Maseeh Professor of Emerging Technologies at Portland State University, with research contributions in database theory, object-oriented technology, scientific databases, and data streams. He received an NSF Young Investigator Award, the 1997 SIGMOD Innovations Award, and a Microsoft Research Outstanding Collaborator Award.

The SSDBM program will also feature a local stakeholder session, where open research issues from the perspective of two local companies will be discussed. Susanne Greiner, a data analyst at Würth Phoenix, will talk about predictive analytics for the management of complex IT systems. Peter Moser, a database developer at IDM Südtirol/Alto Adige, will present challenges for databases in the Open Data Hub Südtirol.

Bongki Moon (Seoul National University, chair), Periklis Andritsos (University of Toronto) and John Wu (Berkeley Lab) generously accepted to serve on the best paper committee. They took on the daunting task to carefully read and evaluate a selection of highly ranked papers and they will present the award for the best paper during the conference.

We would also like to acknowledge the excellent work of all the people involved in the organization of the conference, in particular the demo chair Peer Kröger (LMU Munich), the proceedings chair Dimitris Sacharidis (TU Vienna), the local arrangement chairs Anton Dignös (unibz) and Patrick Ohnewein (IDM Südtirol/Alto Adige), the web chairs Theodoros Chondrogiannis (University of Koblenz) and Luca Miotto (IDM Südtirol/Alto Adige).

The precious support of several colleagues at unibz was indispensable for a smooth local organization and preparation of the conference. We express our gratitude to all of them (in alphabetical order): Anna Anesi, Manuela Degaspari, Nadine Mair and the students Ankita Sadu and Francois Tronche-Macaire.

We are very grateful to the generous contributions of our sponsors and supporters:
* Gold sponsor: Würth Phoenix
* Bronze sponsors: Alpin, EOS Solutions, Systems
* Local organizers: unibz and IDM Südtirol/Alto Adige

On behalf of the organization committee we would like to thank you all for attending the SSDBM 2018 conference.

Welcome in Bozen-Bolzano and enjoy the conference!

Michael Böhlen	**Johann Gamper**
University of Zurich	Free University of Bozen-Bolzano
SSDBM 2018 Program Chair	SSDBM 2018 General Chair

CONFERENCE VENUE

Free University of Bozen-Bolzano / Universitätsplatz 1 - piazza Università, 1 / Italy - 39100, Bozen-Bolzano
All sessions take place in **room D 1.02**

INTERNET

WiFi internet access is provided at the conference venue.
Two networks are available **eduroam** and **openAiR** (no password required).

PROCEEDINGS

The conference proceedings will be accessible through the ACM Digital Library at the following URL:
https://dl.acm.org/citation.cfm?id=3221269

WELCOME RECEPTION

The welcome reception on Monday, July 9, 2018 will take place in Room F6 – University Club at the Free University of Bozen-Bolzano.

EXCURSION AND CONFERENCE BANQUET

SSDBM 2018 will include a half-day social program that brings us to the beautiful surroundings of Bozen-Bolzano. We will first go by bus to Eppan on the South Tyrolean wine route, where we visit the castle Englar and the gothic church belonging to the castle. Englar is more than 500 years old and one of many castles in South Tyrol. The visit of the castle will finish with a wine tasting. After that, the bus brings us to Kaltern on the South Tyrolean wine route. From there, we walk for approximately one hour through vineyards to Lake Kaltern (easy walk, no hiking shoes are required). The banquet starts at around 7pm in the restaurant Gretl am See, which is directly located at the Lake Kaltern.

PROGRAM AT A GLANCE

	MONDAY, 9 JULY, 2018
8:00	Conference Registration
9:00 – 10:30	SSDBM Opening and Welcome Keynote by David Maier
10:30 – 11:00	Coffee Break
11:00 – 12:15	**Research Session 1:** Statistical Data Analysis
12:15 – 13:30	Lunch
13:30 – 14:45	**Research Session 2:** Geospatial Data Management
14:45 – 15:15	Coffee Break
15:15 – 16:30	**Research Session 3:** Indexing and Querying Scientific Data
16:30 – 18:00	Short Papers and Demo Session
19:00 – 21:00	Welcome Reception (F6 - University Club)
	TUESDAY, 10 JULY, 2018
9:00 – 10:15	**Research Session 4:** Privacy and Encryption
10:15 – 10:45	Coffee Break
10:45 – 12:00	**Research Session 5:** Complexity and Scale
12:00 – 12:30	Local Stakeholder Session
12:30 – 13:30	Lunch
15:00 – 23:00	Excursion and Conference Banquet
	WEDNESDAY, 11 JULY, 2018
9:00 – 10:30	30 Years of SSDBM and SSDBM 2019 Keynote by Christian S. Jensen
10:30 – 11:00	Coffee Break
11:00 – 12:15	**Research Session 7:** Social and Collaborative Data Management
12:15 – 13:30	Lunch
13:30 – 14:45	**Research Session 8:** GPU–Assisted Scientific Data Analysis
14:45 – 15:15	Coffee Break
15:15 – 16:05	**Research Session 9:** Exploratory Data Analyses

DETAILED PROGRAM

Monday, 9 July, 2018 09:00-10:30
SSDM Opening and Keynote by David Maier

Session chair: *Michael Böhlen*

SSDBM Opening and Welcome
Michael Böhlen (University of Zurich)
Johann Gamper (Free University of Bozen-Bolzano)

Keynote I: Are Green Buildings Healthy?
David Maier (Portland State University)

Monday, 9 July, 2018 10:30-11:00
Coffee Break

Monday, 9 July, 2018 11:00-12:15
Research Session 1: Statistical Data Analysis

Session chair: *Periklis Andritsos*

SBG-Sketch: A Self-Balanced Sketch for Labeled-Graph Stream Summarization
Mohamed S. Hassan (Purdue University); Bruno Ribeiro (Purdue University);
Walid G. Aref (Purdue University)

A Unified Framework of Density-Based Clustering for Semi-Supervised Classification
Jadson Castro Gertrudes (University of Sao Paulo);
Arthur Zimek (University of Southern Denmark);
Jörg Sander (University of Alberta);
Ricardo J. G. B. Campello (James Cook University)

Feature-based Comparison and Generation of Time Series
Lars Kegel (TU Dresden);
Martin Hahmann (TU Dresden);
Wolfgang Lehner (TU Dresden)

Monday, 9 July, 2018 12:15-13:30
Lunch

Monday, 9 July, 2018 13:30-14:45
Research Session 2: Geospatial Data Management

Session chair: *Erich Schubert*

Finding Shortest Keyword Covering Routes in Road Networks
Vassilis Kaffes (University of the Peloponnese);
Alexandros Belesiotis (Athena Research Center);
Dimitrios Skoutas (Athena Research Center);
Spiros Skiadopoulos (University of Peloponnese)

TIPP: Parallel Delaunay Triangulation for Large-Scale Datasets
Cuong Nguyen (University of Mississippi);
Philip J. Rhodes (University of Mississippi)

GeoSparkViz: A Scalable Geospatial Data Visualization Framework in the Apache Spark Ecosystem
Jia Yu (Arizona State University);
ongsi Zhang (Arizona State University);
Mohamed Sarwat (Arizona State University)

Monday, 9 July, 2018 14:45-15:15
Coffee Break

Monday, 9 July, 2018 15:15-16:30
Research Session 3: Indexing and Querying Scientific Data
Session chair: *Walid Aref*

Multidimensional Range Queries on Modern Hardware
Stefan Sprenger (Humboldt-Universität zu Berlin);
Patrick Schäfer (Humboldt-Universität zu Berlin);
Ulf Leser (Humboldt-Universität zu Berlin)

COMPASS: Compact array storage with value index
Haoyuan Xing (Ohio State University);
Gagan Agrawal (Ohio State University)

Distributed Caching for Processing Raw Arrays
Weijie Zhao (University of California, Merced);
Florin Rusu (UC Merced); Bin Dong (LBNL);
Kesheng Wu (LBNL); Anna Y. Q. Ho (CalTech); Peter Nugent (LBNL)

Monday, 9 July, 2018 16:30-18:00
Short Papers and Demo Session

Optimizer Time Estimation for SQL Queries
Bhashyam Ramesh (Teradata India Pvt Ltd);
C Jaiprakash (Teradata India Pvt Ltd);
Naveen Sankaran (Teradata India Pvt Ltd);
Jitendra Yasaswi (Teradata India Pvt Ltd)

Scheduling Data-intensive Scientific Workflows with Reduced Communication
Ilia Pietri (University of Athens);
Rizos Sakellariou (University of Manchester)

PARADISO: An Interactive Approach of Parameter Selection for the Mean Shift Algorithm
Daniyal Kazempour (LMU Munich);
Anna Beer (LMU Munich); Johannes-Y. Lohrer (LMU Munich);
Daniel Kaltenthaler (LMU Munich); Thomas Seidl (LMU Munich)

Towards an Efficient and Effective Framework for Evolution of Scientific Databases
Robert E. Schuler (USC Information Sciences Institute);
Carl Kesselman (USC Information Sciences Institute)

Maximizing Area-Range Sum for Spatial Shapes (MAxRS^3)
Muhammed Mas-ud Hussain (Northwestern University);
Goce Trajcevski (Iowa State University)

PathGraph: Querying and Exploring Big Data Graphs
Dario Colazzo (U. Paris Dauphine); Vincenzo Mecca (Università della Basilicata);
Maurizio Nolé (Università della Basilicata); Carlo Sartiani (Università della Basilicata)

Crossing an OCEAN of Queries: Analyzing SQL Query Logs with OCEANLog
Andreas M. Wahl (FAU Erlangen-Nürnberg);
Gregor Endler (FAU Erlangen-Nürnberg);
Peter K. Schwab (FAU Erlangen-Nürnberg);
Sebastian Herbst (FAU Erlangen-Nürnberg);
Julian Rith (FAU Erlangen-Nürnberg);
Richard Lenz (FAU Erlangen-Nürnberg)

In-Database Analytics with ibmdbpy
Edouard Fouché (Karlsruhe Institute of Technology);
Alexander Eckert (IBM Deutschland R&D);
Klemens Böhm (Karlsruhe Institute of Technology)

Visual Querying of Large Multilayer Graphs
Erick Cuenca (Lirmm, University of Montpellier);
Arnaud Sallaberry (Lirmm, University of Montpellier);
Dino Ienco (IRSTEA); Pascal Poncelet (Lirmm, University of Montpellier)

Federated Database System for Scientific Data
Sangchul Kim (Seoul National University);
Bongki Moon (Seoul National University)

Monday, 9 July, 2018 19:00-21:00
Welcome Reception (F6 – University Club)

Tuesday, 10 July, 2018 09:00 - 10:15
Research Session 4: Privacy and Encryption

Session chair: *John Wu*

Towards Meaningful Distance-Preserving Encryption
Christine Tex (Karlsruhe Institute of Technology);
Martin Schäler (Karlsruhe Institute of Technology);
Klemens Böhm (Karlsruhe Institute of Technology)

Publishing Spatial Histograms Under Differential Privacy
Soheila Ghane (Melbourne University);
Lars Kulik (University of Melbourne);
Kotagiri Ramamohanarao (University of Melbourne)

Declarative Cartography under Fine-Grained Access Control
Thomas Jensen (June, Danske Bank);
Marcos Antonio Vaz Salles (University of Copenhagen);
Michael Vindahl Bang (June, Danske Bank)

Tuesday, 10 July, 2018 10:15-10:45
Coffee Break

Tuesday, 10 July, 2018 10:45-12:00
Research Session 5: Complexity and Scale

Session chair: *Marcos António Vaz Salles*

Numerically Stable Parallel Computation of (Co-)Variance
Erich Schubert (Heidelberg University);
Michael Gertz (Heidelberg University)

NoSingles: a Space-Efficient Algorithm for Influence Maximization
Diana Popova (University of Victoria);
Naoto Ohsaka (University of Tokyo);
Ken-ichi Kawarabayashi (National Institute of Informatics);
Alex Thomo (University of Victoria)

Point Pattern Search in Big Data
Fabio Porto (LNCC); João Guilherme Rittmeyer (LNCC);
Eduardo Ogasawara (CEFET-RJ);
Alberto Krone-Martins (University of Lisbon);
Patrick Valduriez (INRIA); Dennis Shasha (NYU)

Tuesday, 10 July, 2018 12:00-12:30
Local Stakeholder Session

Session chair: *Johann Gamper*

ITOA: Predictive Analytics to enhance IT operation
Susanne Greiner, Data Scientist at Würth Phoenix

Challenges for Databases in the Open Data Hub Südtirol
Peter Moser, IDM Südtirol / Alto Adige, Ecosystem ICT & Automation

Tuesday, 10 July, 2018 12:30-13:30
Lunch

Tuesday, 10 July, 2018 15:00-23:00
Excursion and Conference Banquet
Best paper award

Wednesday, 11 July, 2018 09:00-10:30
30 Years of SSDBM, SSDBM 2019 and Keynote by Christian S. Jensen

Session chair: *Johann Gamper*

30 Years of SSDBM and SSDBM 2019
Arie Shoshani (Lawrence Berkeley National Laboratory)

Keynote II: Data-Intensive Vehicle Routing
Christian S. Jensen (Aalborg University)

Wednesday, 11 July, 2018 10:30-11:00
Coffee Break

Wednesday, 11 July, 2018 11:00-12:15
Research Session 6: Social and Collaborative Data Management

Session chair: *Ulf Leser*

ERMrest: A web service for collaborative data management
Karl Czajkowski (USC Information Sciences Institute);
Carl Kesselman (USC Information Sciences Institute);
Robert E. Schuler (USC Information Sciences Institute);
Hongsuda Tangmunarunkit (USC Information Sciences Institute)

Metadata-Driven Error Detection
Larysa Visengeriyeva (TU Berlin);
Ziawasch Abedjan (TU Berlin)

Selecting Representative and Diverse Spatio-Textual Posts over Sliding Windows
Dimitris Sacharidis (TU Vienna);
Paras Mehta (FU Berlin);
Dimitrios Skoutas (Athena Research Center);
Kostas Patroumpas (University of Piraeus);
Agnès Voisard (FU Berlin)

Wednesday, 11 July, 2018 12:15-13:30
Lunch

Wednesday, 11 July, 2018 13:30-14:45
Research Session 7: GPU-Assisted Scientific Data Analysis

Session chair: *Florin Rusu*

GPU-Based Parallel Indexing for Concurrent Spatial Query Processing
Zhila Nouri (University of South Florida);
Yi-Cheng Tu (University of South Florida)

Massively-Parallel Break Detection for Satellite Data
Malte von Mehren (University of Copenhagen);
Fabian Gieseke (University of Copenhagen);
Jan Verbesselt (Wageningen University);
Sabina Rosca (Wageningen University);
Stéphanie Horion (University of Copenhagen);
Achim Zeileis (Universität Innsbruck)

Order-Independent Constraint-Based Causal Structure Learning for Gaussian Distribution Models using GPUs
Christopher Schmidt (Hasso Plattner Institute);
Johannes Huegle (Hasso Plattner Institute);
Matthias Uflacker (Hasso Plattner Institute)

Wednesday, 11 July, 2018 14:45-15:15
Coffee Break

Wednesday, 11 July, 2018 15:15-16:05
Research Session 8: Exploratory Data Analyses

Session chair: *Michael Shekelyan*

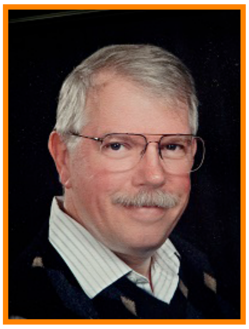
Learning Interesting Attributes for Automated Data Categorization
Koninika Pal (TU Kaiserslautern);
Sebastian Michel (TU Kaiserslautern)

Efficient Anti-community Detection in Complex Networks
Sebastian Lackner (Heidelberg University);
Andreas Spitz (Heidelberg University);
Matthias Weidemüller (Heidelberg University);
Michael Gertz (Heidelberg University)

ABSTRACTS

Monday, 9 July, 2018 09:00-10:30
SSDM Opening and Keynote by David Maier

Session chair: *Michael Böhlen*



SSDBM Opening and Welcome

Michael Böhlen (University of Zurich)
Johann Gamper (Free University of Bozen-Bolzano)

Keynote I: Are Green Buildings Healthy?

David Maier (Portland State University)

Buildings account for 40% of carbon dioxide emission in the US (and more during construction). Thus, the goal of Green Buildings to minimize resource usage and carbon footprint during construction and use is not surprising. While Green Buildings may be healthy for the environment, given that the average person spends 90% of his or her time indoors, it is reasonable to ask if the occupants of such buildings are safe and productive. These goals can work against each other. For example, limiting hot water flow and temperature to reduce water and electricity usage can encourage the growth of harmful bacterial mats. Reducing outdoor air flow to lower heating and cooling costs can raise the rebreathed fraction of air in a room, contributing to disease transmission. Obtaining quantitative results on human well-being and performance in Green Buildings is challenging, but certainly must start with a characterization of conditions in and around buildings. This talk will elucidate the requirements for data and analysis infrastructure needed to investigate the performance of Green Buildings, by looking at specific needs for sample research questions. From there, I go to a proposed architecture for such a system, and then discuss initial experiences in prototyping such a system based on data from a Building Management Systems and additional sensor platforms. I then discuss several interesting problems that have emerged from this work, including: (a) Agile data integration, particularly temporal and spatial alignment of datasets. (b) Exploiting data annotations to support visualization and analysis. (c) Capturing the human element in building performance.

Bio: David Maier is Maseeh Professor of Emerging Technologies at Portland State University. Prior to his current position, he was on the faculty at SUNY-Stony Brook and Oregon Graduate Institute. He has spent extended visits with INRIA, University of Wisconsin–Madison, Microsoft Research, and the National University of Singapore. He is the author of books on relational databases, logic programming, and object-oriented databases, as well as papers in database theory, object-oriented technology, scientific databases, and data streams. He is a recognized expert on the challenges of large-scale data in the sciences. He received an NSF Young Investigator Award in 1984, the 1997 SIGMOD Innovations Award for his contributions in objects and databases, and a Microsoft Research Outstanding Collaborator Award in 2016. He is also an ACM Fellow and IEEE Senior Member. He holds a dual B.A. in Mathematics and in Computer Science from the University of Oregon (Honors College, 1974) and a PhD in Electrical Engineering and Computer Science from Princeton University (1978).

Monday, 9 July, 2018 10:30-11:00
Coffee Break

Monday, 9 July, 2018 11:00-12:15
Research Session 1: Statistical Data Analysis

Session chair: *Periklis Andritsos*

SBG-Sketch: A Self-Balanced Sketch for Labeled-Graph Stream Summarization

Mohamed S. Hassan (Purdue University);
Bruno Ribeiro (Purdue University);
Walid G. Aref (Purdue University)

Applications in various domains rely on processing graph streams, e.g., communication logs of a cloud-troubleshooting system, road-network traffic updates, and interactions on a social network. A labeled-graph stream refers to a sequence of streamed edges of distinct types that form a labeled graph. Due to the large volume and high velocity of these streams, it is often more practical to incrementally build a lossy-compressed version of the graph, and use this lossy version to approximately evaluate graph queries. Challenges arise when the queries are unknown in advance but are associated with filtering predicates based on edge labels. Surprisingly common, and especially challenging, are labeled-graph streams that have highly skewed and unpredictable label-distributions. This paper introduces Self-Balanced Graph Sketch (SBG-Sketch, for short), a graph sketch for summarizing and querying labeled-graph streams, coping with highly imbalanced labels. SBG-Sketch maintains synopsis for both the edge attributes as well as the topology of the streamed graph. SBG-Sketch allows efficient processing of traversal queries, e.g., reachability queries. Experimental results over a variety of real labeled-graph streams show SBG-Sketch to reduce the estimation errors of state-of-the-art methods by up to 99%.

A Unified Framework of Density-Based Clustering for Semi-Supervised Classification

Jadson Castro Gertrudes (University of Sao Paulo);
Arthur Zimek (University of Southern Denmark);
Jörg Sander (University of Alberta);
Ricardo J. G. B. Campello (James Cook University)

Semi-supervised classification is drawing increasing attention in the era of big data, as the gap between the abundance of cheap, automatically collected unlabeled data and the scarcity of labeled data that are laborious and expensive to obtain is dramatically increasing. In this paper, we introduce a unified framework for semi-supervised classification based on building-blocks from density-based clustering. This framework is not only efficient and effective, but it is also statistically sound. Experimental results on a large collection of datasets show the advantages of the proposed framework.

Feature-based Comparison and Generation of Time Series

Lars Kegel (TU Dresden); Martin Hahmann (TU Dresden); Wolfgang Lehner (TU Dresden)

For more than three decades, researchers have been developping generation methods for the weather, energy, and economic domain. These methods provide generated datasets for reasons like system evaluation and data availability. However, despite the variety of approaches, there is no comparative and cross-domain assessment of generation methods and their expressiveness. We present a similarity measure that analyzes generation methods regarding general time series features. By this means, users can compare generation methods and validate whether a generated dataset is considered similar to a given dataset. Moreover, we propose a feature-based generation method that evolves cross-domain time series datasets. This method outperforms other generation methods regarding the feature-based similarity.

Monday, 9 July, 2018 12:15-13:30 Lunch

Monday, 9 July, 2018 13:30-14:45 Research Session 2: Geospatial Data Management

Session chair: *Erich Schubert*

Finding Shortest Keyword Covering Routes in Road Networks

Vassilis Kaffes (University of the Peloponnese); Alexandros Belesiotis (Athena Research Center); Dimitrios Skoutas (Athena Research Center); Spiros Skiadopoulos (University of Peloponnese)

Millions of users rely on navigation applications to compute an optimal route for their trips. The basic functionality of these applications is to find the minimum cost route between a source and target node in the transportation network. In this paper, we address a variant of this problem, where the computed route is required to contain a set of Points of Interest of specific types. Our approach is based on the concept of keyword skyline. We formally define this concept, and we show how to compute the keyword skyline for the vertices of a given network and how to use it for computing the shortest keyword covering paths. We present different variants of this method, including an approximation algorithm, providing different tradeoffs between preprocessing cost and execution time. Finally, we present an experimental evaluation of our approach using real-world datasets of different sizes, including also a comparison to the current state-of-the-art algorithm for this problem.

TIPP: Parallel Delaunay Triangulation for Large-Scale Datasets

Cuong Nguyen (University of Mississippi); Philip J. Rhodes (University of Mississippi)

Because of the importance of Delaunay Triangulation in science and engineering, researchers have devoted extensive attention to parallelizing this fundamental algorithm. However, generating un- structured meshes for extremely large point sets remains a barrier for scientists working with large scale or high resolution datasets. In this paper, we introduce a novel algorithm – Triangulation of Independent Partitions in Parallel which divides the domain into many independent partitions that can be triangulated in parallel. In contrast to stitching methods, merging our partition triangulations into a single result is easily done, and satisfies the Delaunay criteria. We use C/C++ and MPI (Message Passing Interface) to implement and evaluate our algorithm on a cluster. Experimental results show that our parallel implementation outperforms our serial implementation by roughly 27× for 1 billion triangles. Lastly, we believe we have generated the largest Delaunay mesh to-date, at 16 billion triangles.

GeoSparkViz: A Scalable Geospatial Data Visualization Framework in the Apache Spark Ecosystem

Jia Yu (Arizona State University); Zongsi Zhang (Arizona State University); Mohamed Sarwat (Arizona State University)

Data Visualization allows users to summarize, analyze and reason about data. A map visualization tool first loads the designated geospatial data, processes the data and then applies the map visualization effect. Guaranteeing detailed and accurate geospatial map visualization (e.g., at multiple zoom levels) requires extremely high resolution maps. Classic solutions suffer from limited computation resources and hence take a tremendous amount of time to generate maps for large-scale geospatial data. The paper presents GeoSparkViz a large-scale geospatial map visualization framework. GeoSparkViz extends a cluster computing system (Apache Spark in our case) to provide native support for general cartographic design. The proposed system seamlessly integrates with a Spark-based spatial data management system, GeoSpark. It provides the data scientist a holistic system that allows her to perform data management and visualization on spatial data and reduces the overhead of loading the intermediate spatial data generated during the data management phase to the designated map visualization tool. GeoSparkViz also proposes a map tile data partitioning method that achieves load balancing for the map visualization workloads among all nodes in the cluster. Extensive experiments show that GeoSparkViz can generate a high-resolution (i.e., Gigapixel image) Heatmap of 1.7 billion Open-StreetMaps objects and 1.3 billion NYC taxi trips in ≈4 and 5 minutes on a four-node commodity cluster, respectively.

Monday, 9 July, 2018 14:45-15:15
Coffee Break

Monday, 9 July, 2018 15:15-16:30
Research Session 3: Indexing and Querying Scientific Data

Session chair: *Walid Aref*

Multidimensional Range Queries on Modern Hardware

Stefan Sprenger (Humboldt-Universität zu Berlin);
Patrick Schäfer (Humboldt-Universität zu Berlin);
Ulf Leser (Humboldt-Universität zu Berlin)

Range queries over multidimensional data are an important part of database workloads in many applications. Their execution may be accelerated by using multidimensional index structures (MDIS), such as kd-trees or R-trees. As for most index structures, the usefulness of this approach depends on the selectivity of the queries, and common wisdom told that a simple scan beats MDIS for queries accessing more than 15%-20% of a dataset. However, this wisdom is largely based on evaluations that are almost two decades old, performed on data being held on disks, applying IO-optimized data structures, and using single-core systems. The question is whether this rule of thumb still holds when multidimensional range queries (MDRQ) are performed on modern architectures with large main memories holding all data, multi-core CPUs and data-parallel instruction sets. In this paper, we study the question whether and how much modern hardware influences the performance ratio between index structures and scans for MDRQ. To this end, we conservatively adapted three popular MDIS, namely the R_*-tree, the kd-tree, and the VA-file, to exploit features of modern servers and compared their performance to different flavors of parallel scans using multiple (synthetic and real-world) analytical workloads over multiple (synthetic and real-world) datasets of varying size, dimensionality, and skew. We find that all approaches benefit considerably from using main memory and parallelization, yet to varying degrees. Our evaluation indicates that, on current machines, scanning should be favored over parallel versions of classical MDIS even for very selective queries.

COMPASS: Compact array storage with value index

Haoyuan Xing (Ohio State University);
Gagan Agrawal (Ohio State University)

Efficient array storage is the backbone of scientific data processing. With an explosion of data, rapidly answering queries on array data is becoming increasingly important. Although most of the array storages today support subsetting of an array based on dimensions efficiently, they fall back to full scan while executing value-based filter operations. This has lead to an interest in approximate query processing, but such methods can have substantial inaccuracies. This paper presents COMPASS, an array storage system with integrated value index support. Our approach efficiently encodes arrays as bin-based indices and corresponding residuals describing elements in each bin. Our query processing method uses bin-based indices, with residuals decompressed as needed, to ensure that accuracy is not sacrificed. Our evaluation shows that compared with current array storage systems such as SciDB, our method achieves a smaller storage footprint, but most importantly, can perform filter operations an order of magnitude faster on low selectivity queries. Meanwhile, COMPASS maintains comparable performance on high-selectivity queries or dimension-based subsetting operations.

Distributed Caching for Processing Raw Arrays

Weijie Zhao (University of California, Merced);
Florin Rusu (UC Merced);
Bin Dong (LBNL);
Kesheng Wu (LBNL);
Anna Y. Q. Ho (CalTech);
Peter Nugent (LBNL)

As applications continue to generate multi-dimensional data at exponentially increasing rates, fast analytics to extract meaningful results is becoming extremely important. The database community has developed array databases that alleviate this problem through a series of techniques. In-situ mechanisms provide direct access to raw data in the original format—without loading and partitioning. Parallel processing scales to the largest datasets. In-memory caching reduces latency when the same data are accessed across a workload of queries. However, we are not aware of any work on distributed caching of multi-dimensional raw arrays. In this paper, we introduce a distributed framework for cost-based caching of multi-dimensional arrays in native format. Given a set of files that contain portions of an array and an online query workload, the framework computes an effective caching plan in two stages. First, the plan identifies the cells to be cached locally from each of the input files by continuously refining an evolving R-tree index. In the second stage, an optimal assignment of cells to nodes that collocates dependent cells in order to minimize the overall data transfer is determined. We design cache eviction and placement heuristic algorithms that consider the historical query workload. A thorough experimental evaluation over two real datasets in three file formats confirms the superiority – by as much as two orders of magnitude – of the proposed framework over existing techniques in terms of cache overhead and workload execution time.

Optimizer Time Estimation for SQL Queries

Bhashyam Ramesh (Teradata India Pvt Ltd);
C Jaiprakash (Teradata India Pvt Ltd);
Naveen Sankaran (Teradata India Pvt Ltd);
Jitendra Yasaswi (Teradata India Pvt Ltd)

Predicting the amount of time a SQL query takes to execute can help in prioritizing, optimizing and scheduling the query execution. This also helps in optimal utilization of hardware resources. The total execution time of a query can be split into the time taken for parsing/optimizing a query and the time taken for the actual execution. In this work, we focus on solving the first part of the problem, that is predicting the optimization time of a query. Predicting optimization time can hint the optimizer not to spend too much time optimizing the query in case when parse time is much higher than execution time. Such query execution plans can be cached to speed-up execution in future. If optimization time is much lower than execution time, we can choose not to cache the plan and can make better utilization of execution plan cache. If optimization time is relatively low compared to execution time, optimizer can be hinted to spend more time in parsing to produce better optimized plan which can reduce the execution time. One method towards predicting the parse time is to use some heuristic information by looking at query text. In this work, we take advantage of machine learning techniques by designing a set of features from the SQL query text and use a neural network to predict the parse time. We have tried both regression as well as classification based approaches in this work. We report high accuracy while predicting the elapsed parse time for SQL queries.

Scheduling Data-intensive Scientific Workflows with Reduced Communication

Ilia Pietri (University of Athens);
Rizos Sakellariou (University of Manchester)

Data-intensive scientific workflows, typically modelled by directed acyclic graphs, consist of interdependent tasks that exchange significant amounts of data and are executed on parallel/distributed clusters. However, the energy or monetary costs associated with large data transfers between tasks executing on different nodes may be significant. As a result, there is scope to explore the possibility of trading some communication for computation, aiming to reduce overall communication costs. In this work, we propose a scheduling approach that scales the weight of communication to increase its impact when building the schedule of a scientific workflow; the aim is to assign pairs of tasks with significant data transfers to the same computational node so that the overall communication cost is minimized. The proposed approach is evaluated using simulation and three real-world scientific workflows. The tradeoff between scientific workflow execution time and the size of data transfers is assessed for different weights and a different number of computational nodes.

PARADISO: An Interactive Approach of Parameter Selection for the Mean Shift Algorithm

Daniyal Kazempour (LMU Munich);
Anna Beer (LMU Munich);
Johannes-Y. Lohrer (LMU Munich);
Daniel Kaltenthaler (LMU Munich);
Thomas Seidl (LMU Munich)

Many algorithms have been developed for detecting clusters of various kinds over the past decades. However, just few attempts have been made to provide an interactive setting for the clustering algorithms. In this paper, we present PARADISO, an interactive Mean Shift method. It enables the user to get back to any arbitrary iteration point of the run observing the evolution of the clusters after each iteration and to set different bandwidth parameters. The user gets a clustering result with this method which emerged through multiple bandwidths while the user can see the full chain of effects of the chosen bandwidths over all iterations. Further, our method provides so-called Points-Shifted-Distance plots (PSD plots) for the Mean Shift algorithm which aim to facilitate the choice of a different bandwidth for the user. Beyond the mentioned features, PARADISO provides a visualization method which lets the user see the different bandwidth choices made in form of pathways.

Towards an Efficient and Effective Framework for Evolution of Scientific Databases

Robert E. Schuler (USC Information Sciences Institute);
Carl Kesselman (USC Information Sciences Institute)

Database systems are well suited to scientific data management and analysis workloads, however, a database must evolve to keep pace with changing requirements and adjust to changes in the domain conceptualization as applications mature. Evolving a database (i.e., updating its schema and instance data) is one of the greatest challenges in database maintenance and the difficulties are compounded by the lack of sufficient tools to support scientists. This paper presents a schema evolution framework based on an algebraic approach that introduces extended and higher-level composite relational operators tailored to the task of schema evolution. These higher-level operators simplify the task of evolving a database for non-expert users, while enabling efficient evaluation of schema evolution expressions.

Maximizing Area-Range Sum for Spatial Shapes (MAxRS^3)

Muhammed Mas-ud Hussain (Northwestern University);
Goce Trajcevski (Iowa State University)

We investigate a novel variant of the well-known MaxRS (Maximizing Range Sum) problem – namely, the MAxRS3 (Maximizing Area-Range Sum for Spatial Shapes). The MaxRS problem amounts to detecting a location where a fixed-size rectangle R should be placed, so that it covers a maximum number of points – or sum of weights, if the points are weighted – from a given input set of 2D points. While variants have tackled the settings in which the input set to MaxRS problem consists of polygons instead of points – the solution is still based on (weighted) count. We postulate that in many practical applications it is of interest to determine where to place the input rectangle so that the total area-coverage in its interior is maximized. In this paper, we formalize the MAxRS3 problem and propose (to our knowledge) the first solution to this new problem.

PathGraph: Querying and Exploring Big Data Graphs

Dario Colazzo (U. Paris Dauphine);
Vincenzo Mecca (Università della Basilicata);
Maurizio Nolé (Università della Basilicata);
Carlo Sartiani (Università della Basilicata)

With the widespread diffusion of social networks and the dawn of data-intensive scientific applications, graphs became one of the foundations for modern data management applications. A key role in graph querying and analysis is played by Regular Path Queries, their extensions, and, in particular, GXPath. In this demo we will present PathGraph, a distributed GXPath query processor, and its web-based graphical interface.

Crossing an OCEAN of Queries: Analyzing SQL Query Logs with OCEANLog

Andreas M. Wahl (FAU Erlangen-Nürnberg);
Gregor Endler (FAU Erlangen-Nürnberg);
Peter K. Schwab (FAU Erlangen-Nürnberg);
Sebastian Herbst (FAU Erlangen-Nürnberg);
Julian Rith (FAU Erlangen-Nürnberg);
Richard Lenz (FAU Erlangen-Nürnberg)

SQL queries encapsulate the knowledge of their authors about the usage of the queried data sources. This knowledge also contains aspects that cannot be inferred by analyzing the contents of the queried data sources alone. Due to the complexity of analytical SQL queries, specialized mechanisms are necessary to enable the user-friendly formulation of meta-queries over an existing query log. Currently existing approaches do not sufficiently consider syntactic and semantic aspects of queries along with contextual information. During our demonstration, conference participants learn how to use the latest release of OCEANLog, a framework for analyzing SQL query logs. Our demonstration encompasses several scenarios. Participants can explore an existing query log using domain-specific graph traversal expressions, set up continuous subscriptions for changes in the graph, create time-based visualizations of query results, configure an OCEANLog instance and learn how to choose a decide which specific graph database to use. We also provide them with access to the native meta-query mechanisms of a DBMS to further emphasize the benefits of our graph-based approach.

In-Database Analytics with ibmdbpy

Edouard Fouché (Karlsruhe Institute of Technology);
Alexander Eckert (IBM Deutschland R&D);
Klemens Böhm (Karlsruhe Institute of Technology)

The increasing size of the available data and database volumes rep- resents a real challenge for the data management community. In general, current approaches in data mining require the data to be first extracted from an underlying database. From a practical point of view, this presents many drawbacks. In this short article, we present a possible solution to bridge the gap between data repositories and end user analysis. We demonstrate the interestingness of this approach with ibmdbpy, an open source Python interface developed by IBM for database administration and data analytics.

Visual Querying of Large Multilayer Graphs

Erick Cuenca (Lirmm, University of Montpellier);
Arnaud Sallaberry (Lirmm, University of Montpellier);
Dino Ienco (IRSTEA);
Pascal Poncelet (Lirmm, University of Montpellier)

Many real world data can be represented by a network with a set of nodes linked each other by multiple relations. Such a rich graph is called multilayer graph. In this demo, we present a tool for Visual Querying of Large Multilayer Graphs that allows to visually draw the query, retrieve result patterns and finally navigate and browse the results considering the original multilayer graph database. Our approach does not only provide a graphical user interface for the graph engine but the query processing is fully integrated.

Federated Database System for Scientific Data

Sangchul Kim (Seoul National University);
Bongki Moon (Seoul National University)

Much like traditional databases, scientific data are managed in multiple separate databases by different sources and organizations. When such distributed data are analyzed together for more comprehensive understanding and prediction, it is necessary to access data via multiple simultaneous connections or collected in a single location. The inevitable consequence is, however, that a significant overhead is incurred due to differences in schemas, data transformation, and extraneous cost for storing intermediate data. This demo presents SDF, Scientific Database in Federation, which facilitates data sharing and exchange in order to support complex analytics with minimal integration overhead. SDF is currently implemented in SciDB using user-defined operators, providing two connection models, master-to-master and cluster-to-master, for a shared-nothing architecture.

Monday, 9 July, 2018 19:00-21:00
Welcome Reception (F6 – University Club)

Tuesday, 10 July, 2018 09:00 - 10:15
Research Session 4: Privacy and Encryption

Session chair: *John Wu*

Towards Meaningful Distance-Preserving Encryption

Christine Tex (Karlsruhe Institute of Technology);
Martin Schäler (Karlsruhe Institute of Technology);
Klemens Böhm (Karlsruhe Institute of Technology)

Mining complex data is an essential and at the same time challenging task. Therefore, organizations pass on their encrypted data to service providers carrying out such analyses. Thus, encryption must preserve the mining results. Many mining algorithms are distance-based. Thus, we investigate how to preserve the results for such algorithms upon encryption. To this end, we propose the notion of distance-preserving encryption (DPE). This notion has just the right strictness – we show that we cannot relax it, using formal arguments as well as experiments. Designing a DPE scheme is challenging, as it depends both on the data set and the specific distance measure in use. We propose a procedure to engineer DPE-schemes, dubbed DisPE. In a case study, we instantiate DisPE for SQL query logs, a type of data containing valuable information about user interests. In this study, we design DPE schemes for all SQL query distance measures from the scientific literature. We formally show that one can use a combination of existing secure property-preserving encryption schemes to this end. Finally, we discuss on the generalizability of our findings using two other data sets as examples.

Publishing Spatial Histograms Under Differential Privacy

Soheila Ghane (Melbourne University);
Lars Kulik (University of Melbourne);
Kotagiri Ramamohanarao (University of Melbourne)

Studying trajectories of individuals has received growing interest. The aggregated movement behaviour of people provides important insights about their habits, interests, and lifestyles. Understanding and utilizing trajectory data is a crucial part of many applications such as location based services, urban planning, and traffic monitoring systems. Spatial histograms and spatial range queries are key components in such applications to efficiently store and answer queries on trajectory data. A spatial histogram maintains the sequentiality of location points in a trajectory by a strong sequential dependency among histogram cells. This dependency is an essential property in answering spatial range queries. However, the trajectories of individuals are unique and even aggregating them in spatial histograms cannot completely ensure an individual's privacy. A key technique to ensure privacy for data publishing is ϵ-differential privacy as it provides a strong guarantee on an individual's provided data. Our work is the first that guarantees ϵ-differential privacy for spatial histograms on trajectories, while ensuring the sequentiality of trajectory data, i.e., its consistency. Consistency is key for any database and our proposed mechanism, PriSH, synthesizes a spatial histogram and ensures the consistency of published histogram with respect to the strong dependency constraint. In extensive experiments on real and synthetic datasets, we show that (1) PriSH is highly scalable with the dataset size and granularity of the space decomposition, (2) the distribution of aggregate trajectory information in the synthesized histogram accurately preserves the distribution of original histogram, and (3) the output has high accuracy in answering arbitrary spatial range queries.

Declarative Cartography under Fine-Grained Access Control

Thomas Jensen (June, Danske Bank);
Marcos Antonio Vaz Salles (University of Copenhagen);
Michael Vindahl Bang (June, Danske Bank)

Visualization of spatial data is of increasing importance in science and society, but opens up justified concerns about data privacy and security. A classic methodology for cartography through generalization is data selection; however, data selection can be challenging under security constraints for two main reasons. First, individual records are kept in the visualization, so a data security approach such as access control needs to be put in place to avoid leakage of information about protected records to unauthorized parties. Second, it can be computationally hard to pick out records from a large spatial dataset so as to create an aesthetically pleasing visualization respecting user constraints and optimization goals. The latter expense can get compounded by the need to additionally respect access control restrictions. This paper presents a way to integrate label-based access control into an existing technique for declarative cartography termed global selection. Through a set of theorems and new algorithms, we demonstrate that we can reuse derivation and resolution of record conflicts when computing global selections across access roles in a security hierarchy. In experiments with realistic datasets, the runtime of the best among these new methods achieves an improvement of up to 2x-5x compared with repeatedly computing the global selection in medium-to-large security hierarchies.

Tuesday, 10 July, 2018 10:15-10:45
Coffee Break

Tuesday, 10 July, 2018 10:45-12:00
Research Session 5: Complexity and Scale

Session chair: *Marcos António Vaz Salles*

Numerically Stable Parallel Computation of (Co-) Variance

Erich Schubert (Heidelberg University);
Michael Gertz (Heidelberg University)

With the advent of big data, we see an increasing interest in computing correlations in huge data sets with both many instances and many variables. Essential descriptive statistics such as the variance, standard deviation, covariance, and correlation can suffer from a numerical instability known as "catastrophic cancellation" that can lead to problems when naively computing these statistics with a popular textbook equation. While this instability has been discussed in the literature already 50 years ago, we found that even today, some high-profile tools still employ the instable version. In this paper, we study a popular incremental technique originally proposed by Welford, which we extend to weighted covariance and correlation. We also discuss strategies for further improving numerical precision, how to compute such statistics online on a data stream, with exponential aging, with missing data, and a batch parallelization for both high performance and numerical precision. We demonstrate when the numerical instability arises, and the performance of different approaches under these conditions. We showcase applications from the classic computation of variance as well as advanced applications such as stock market analysis with exponentially weighted moving models and Gaussian mixture modeling for cluster analysis that all benefit from this approach.

NoSingles: a Space-Efficient Algorithm for Influence Maximization

*Diana Popova (University of Victoria);
Naoto Ohsaka (University of Tokyo);
Ken-ichi Kawarabayashi (National Institute of Informatics);
Alex Thomo (University of Victoria)*

Algorithmic problems of computing influence estimation and influence maximization have been actively researched for decades. We developed a novel algorithm, NoSingles, based on the Reverse Influence Sampling method proposed by Borgs et al. in 2013. NoSingles solves the problem of influence maximization in large graphs using much smaller space than the existing state-of-the-art algorithms while preserving the theoretical guarantee of the approximation of $(1 - 1/e - \epsilon)$ of the optimum, for any $\epsilon > 0$. The NoSingles data structure is saved on the hard drive of the machine, and can be used repeatedly for playing out "what if" scenarios (e.g. trying different combination of seeds and calculating the influence spread). We also introduce a variation of NoSingles algorithm, which further decreases the running time, while preserving the approximation guarantee. We support our claims with extensive experiments on large real-world graphs. Savings in required space allow to successfully run NoSingles on a consumer-grade laptop for graphs with tens of millions of vertices and hundreds of millions of edges.

Point Pattern Search in Big Data

*Fabio Porto (LNCC);
João Guilherme Rittmeyer (LNCC);
Eduardo Ogasawara (CEFET-RJ);
Alberto Krone-Martins (University of Lisbon);
Patrick Valduriez (INRIA); Dennis Shasha (NYU)*

Consider a set of points P in space with at least some of the pairwise distances specified. Given this set P, consider the following three kinds of queries against a database D of points : (i) pure constellation query: find all sets S in D of size |P| that exactly match the pairwise distances within P up to an additive error ϵ; (ii) isotropic constellation queries: find all sets S in D of size |P| such that there exists some scale factor f for which the distances between pairs in S exactly match f times the distances between corresponding pairs of P up to an additive ϵ; (iii) non-isotropic constellation queries: find all sets S in D of size |P| such that there exists some scale factor f and for at least some pairs of points, a maximum stretch factormi, j > 1 such that (f ×mi, j×dist (pi,pj))+ϵ > dist(si,sj) >(f × dist (pi,pj)) - ϵ. Finding matches to such queries has applications to spatial data in astronomical, seismic, and any domain in which (approximate, scale-independent) geometrical matching is required. Answering the isotropic and non-isotropic queries is challenging because scale factors and stretch factors may take any of an infinite number of values. This paper proposes practically efficient sequential and distributed algorithms for pure, isotropic, and non-isotropic constellation queries. As far as we know, this is the first work to address isotropic and non-isotropic queries.

Tuesday, 10 July, 2018 12:00-12:30
Local Stakeholder Session

Session chair: *Johann Gamper*

ITOA: Predictive Analytics to enhance IT operation

Susanne Greiner, Data Scientist at Würth Phoenix

Susanne Greiner, Data Scientist at Würth Phoenix, will illustrate how IT Operation Analytics [ITOA] can help to analyze the configuration of complex server, DB and application environments and potential changes dynamically. Bottlenecks can be found more easily and root cause detection becomes faster and more reliable.In this context AI-based and integrated automated (monitoring) solutions play an important role. They are able to provide holistic visibility into critical application performance and the ability to alert only on IT and business anomalies.

Challenges for Databases in the Open Data Hub Südtirol

Peter Moser, IDM Südtirol / Alto Adige, Ecosystem ICT & Automation

The Open Data Hub (http://opendatahub.bz.it) is a project aiming to provide an access point to South Tyrol's relevant data. This talk provides a short glance into the Open Data Hub, and in particular, focuses on two problems that need to be faced in the development. First, how to find unified views upon different data sources during data integration, and second, how to create fast temporal and spatial SQL queries to serve stored data through web-services in short time.

Tuesday, 10 July, 2018 12:30-13:30
Lunch

Tuesday, 10 July, 2018 15:00-23:00
Excursion and Conference Banquet
Best paper award

Wednesday, 11 July, 2018 09:00-10:30
30 Years of SSDBM, SSDBM 2019 and Keynote by Christian S. Jensen

Session chair: *Johann Gamper*

30 Years of SSDBM and SSDBM 2019

Arie Shoshani (Lawrence Berkeley National Laboratory)

Keynote II: Data-Intensive Vehicle Routing

Christian S. Jensen (Aalborg University)

As the society-wide digitalization unfolds, important societal processes are being captured at an unprecedented level of detail, in turn enabling us to better understand and improve those processes. Vehicular transportation is one such process, where the availability of vehicle trajectories holds the potential to enable better routing. The speaker argues that with massive trajectory data available, the traditional vehicle routing paradigm, Dijkstra's paradigm, where a road network is modeled as a graph and where travel costs such as travel times are assigned to edges, is obsolete. Instead, new and data-intensive paradigms that thrive on data are called for. The talk will cover several such paradigms: a path-based paradigm, where travel costs are associated with paths and not just graph edges; an on-the-fly paradigm, where high-resolution travels costs are not pre-computed but are computed from purposefully selected trajectories during vehicle routing; and a cost-oblivious paradigm, where routing is done without the use of travel costs. These paradigms present new challenges and opportunities to research in routing.

Bio: Christian S. Jensen is Obel Professor of Computer Science at Aalborg University, Denmark, and he was recently with Aarhus University for three years and spent a one-year sabbatical at Google Inc., Mountain View. His research concerns data management and data-intensive systems, and its focus is on temporal and spatio-temporal data management. Christian is an ACM and an IEEE Fellow, and he is a member of Academia Europaea, the Royal Danish Academy of Sciences and Letters, and the Danish Academy of Technical Sciences. He is Editor-in-Chief of ACM Transactions on Database Systems.

Wednesday, 11 July, 2018 10:30-11:00
Coffee Break

Wednesday, 11 July, 2018 11:00-12:15
Research Session 6: Social and Collaborative Data Management

Session chair: *Ulf Leser*

ERMrest: A web service for collaborative data management

Karl Czajkowski (USC Information Sciences Institute); Carl Kesselman (USC Information Sciences Institute); Robert E. Schuler (USC Information Sciences Institute); Hongsuda Tangmunarunkit (USC Information Sciences Institute)

The foundation of data oriented scientific collaboration is the ability for participants to find, access and reuse data created during the course of an investigation, what has been referred to as the FAIR principles. In this paper, we describe ERMrest, a collaborative data management service that promotes data oriented collaboration by enabling FAIR data management throughout the data life cycle. ERMrest is a RESTful web service that promotes discovery and reuse by organizing diverse data assets into a dynamic entity relationship model.We present details on the design and implementation of ERMrest, data on its performance and its use by a range of collaborations to accelerate and enhance their scientific output.

Metadata-Driven Error Detection

Larysa Visengeriyeva (TU Berlin); Ziawasch Abedjan (TU Berlin)

Scientific data often originates from multiple sources and human agents. The integration of data from different sources must also resolve data quality problems that might occur because of inconsistency or different quality assurance levels of the sources. To identify various data quality problems in a dataset, it is necessary to use several error detection methods. Existing error detection solutions are usually tailored towards one specific type of data errors, such as rule violations or outliers, requiring the application of multiple strategies. Using all possible error detection methods is also not satisfying, as some systems might perform poorly on a particular dataset by producing a large number of false positives and missing some results. However, it is not trivial to assess the effectiveness of each strategy upfront. We propose two new holistic approaches for effectively combining off-the-shelf error detection systems. Our approaches are learning-based and incorporate metadata extracted from the dataset at hand. We empirically show, using four real-world datasets, that our method of combining error-detecting strategies achieves an average F1 score 15% higher than multiple heuristics-based baselines.

Selecting Representative and Diverse Spatio-Textual Posts over Sliding Windows

Dimitris Sacharidis (TU Vienna); Paras Mehta (FU Berlin); Dimitrios Skoutas (Athena Research Center); Kostas Patroumpas (University of Piraeus); Agnès Voisard (FU Berlin)

Thousands of posts are generated constantly by millions of users in social media, with an increasing portion of this content being geotagged. Keeping track of the whole stream of this spatio-textual content can easily become overwhelming for the user. In this paper, we address the problem of selecting a small, representative and diversified subset of posts, which is continuously updated over a sliding window. Each such subset can be considered as a concise summary of the stream's contents within the respective time interval, being dynamically updated every time the window slides to reflect newly arrived and expired posts. We define the criteria for selecting the contents of each summary, and we present several alternative strategies for summary construction and maintenance that provide different trade-offs between information quality and performance. Furthermore , we optimize the performance of our methods by partitioning the newly arriving posts spatio-textually and computing bounds for the coverage and diversity of the posts in each partition. The proposed methods are evaluated experimentally using real-world datasets containing geotagged tweets and photos.

Wednesday, 11 July, 2018 12:15-13:30
Lunch

Wednesday, 11 July, 2018 13:30-14:45
Research Session 7: GPU-Assisted Scientific Data Analysis

Session chair: *Florin Rusu*

GPU-Based Parallel Indexing for Concurrent Spatial Query Processing

Zhila Nouri (University of South Florida);
Yi-Cheng Tu (University of South Florida)

In most spatial database applications, the input data is very large. Previous work has shown the importance of using spatial indexing and parallel computing to speed up such tasks. In recent years, GPUs have become a mainstream platform for massively parallel data processing. On the other hand, due to the complex hardware architecture and programming model, developing programs optimized towards high performance on GPUs is non-trivial, and traditional wisdom geared towards CPU implementations is often found to be ineffective. Recent work on GPU-based spatial indexing focused on parallelizing one individual query at a time. In this paper, we argue that current one-query-at-a-time approach has low work efficiency and cannot make good use of GPU resources. To address such challenges, we present a framework named G-PICS for parallel processing of large number of concurrent spatial queries over big datasets on GPUs. G-PICS is motivated by the fact that many spatial query processing applications are busy systems in which a large number of queries arrive per unit of time. G-PICS encapsulates an efficient parallel algorithm for constructing spatial trees on GPUs and supports major spatial query types such as spatial point search, range search, within-distance search, k-nearest neighbors, and spatial joins. While support for dynamic data inputs missing from existing work, G-PICS provides an efficient parallel update procedure on GPUs. With the query processing, tree construction, and update procedure introduced, G-PICS shows great performance boosts over best-known parallel GPU and parallel CPU-based spatial processing systems.

Massively-Parallel Break Detection for Satellite Data

Malte von Mehren (University of Copenhagen);
Fabian Gieseke (University of Copenhagen);
Jan Verbesselt (Wageningen University);
Sabina Rosca (Wageningen University);
Stéphanie Horion (University of Copenhagen);
Achim Zeileis (Universität Innsbruck)

The field of remote sensing is nowadays faced with huge amounts of data. While this offers a variety of exciting research opportunities, it also yields significant challenges regarding both computation time and space requirements. In practice, the sheer data volumes render existing approaches too slow for processing and analyzing all the available data. This work aims at accelerating BFAST, one of the state-of-the-art methods for break detection given satellite image time series. In particular, we propose a massively-parallel implementation for BFAST that can effectively make use of modern parallel compute devices such as GPUs. Our experimental evaluation shows that the proposed GPU implementation is up to four orders of magnitude faster than the existing publicly available implementation and up to ten times faster than a corresponding multi-threaded CPU execution. The dramatic decrease in running time renders the analysis of significantly larger datasets possible in seconds or minutes instead of hours or days. We demonstrate the practical benefits of our implementations given both artificial and real datasets.

Order-Independent Constraint-Based Causal Structure Learning for Gaussian Distribution Models using GPUs

Christopher Schmidt (Hasso Plattner Institute); Johannes Huegle (Hasso Plattner Institute); Matthias Uflacker (Hasso Plattner Institute)

Learning the causal structures in high-dimensional datasets allows deriving advanced insights from observational data, thus creating the potential for new applications. One crucial limitation of state-of-the-art methods for learning causal relationships, such as the PC algorithm, is their long execution time. While, in the worst case, the execution time is exponential to the dimension of a given dataset, it is polynomial if the underlying causal structures are sparse. To address the long execution time, parallelized extensions of the algorithm have been developed addressing the Central Processing Unit (CPU) as the primary execution device. While modern multicore CPUs expose a decent level of parallelism, coprocessors, such as Graphics Processing Units (GPUs), are specifically designed to process thousands of data points in parallel, providing superior parallel processing capabilities compared to CPUs. In our work, we leverage the parallel processing power of GPUs to address the drawback of the long execution time of the PC algorithm and develop an efficient GPU-accelerated implementation for Gaussian distribution models. Based on an experimental evaluation of various high-dimensional real-world gene expression datasets, we show that our GPU-accelerated implementation outperforms existing CPU-based versions, by factors up to 700.

Wednesday, 11 July, 2018 14:45-15:15
Coffee Break

Wednesday, 11 July, 2018 15:15-16:05
Research Session 8: Exploratory Data Analyses

Session chair: *Michael Shekelyan*

Learning Interesting Attributes for Automated Data Categorization

Koninika Pal (TU Kaiserslautern); Sebastian Michel (TU Kaiserslautern)

This work proposes and evaluates a novel approach to determining interesting attributes, in order to categorize entities accordingly. Once identified, such categories are of immense value to allow constraining (filtering) a user's current view to subsets of entities. We show how a classifier is trained that is able to tell whether or not a categorical attribute can act as a constraint, in the sense of human-perceived interestingness. The training data is harnessed from Wikipedia tables, treating the presence or absence of a table as an indication that the attribute used as a filter constraint is reasonable or not. For learning the classification model, we review four well-known statistical measures (features) for categorical attributes—entropy, unalikeability, peculiarity, and coverage. We additionally propose three new statistical measures to capture the distribution of data, tailored to our main objective. The learned model is evaluated by relevance assessments obtained through a user study, reflecting the applicability of the approach as a whole and, further, demonstrates the superiority of the proposed diversity measures over existing measures like information entropy.

Efficient Anti-community Detection in Complex Networks

Sebastian Lackner (Heidelberg University);
Andreas Spitz (Heidelberg University);
Matthias Weidemüller (Heidelberg University);
Michael Gertz (Heidelberg University)

Modeling the relations between the components of complex systems as networks of vertices and edges is a commonly used method in many scientific disciplines that serves to obtain a deeper understanding of the systems themselves. In particular, the detection of densely connected communities in these networks is frequently used to identify functionally related components, such as social circles in networks of personal relations or interactions between agents in biological networks. Traditionally, communities are considered to have a high density of internal connections, combined with a low density of external edges between different communities. However, not all naturally occurring communities in complex networks are characterized by this notion of structural equivalence, such as groups of energy states with shared quantum numbers in networks of spectral line transitions. In this paper, we focus on this inverse task of detecting anti-communities that are characterized by an exceptionally low density of internal connections and a high density of external connections. While anti-communities have been discussed in the literature for anecdotal applications or as a modification of traditional community detection, no rigorous investigation of algorithms for the problem has been presented. To this end, we introduce and discuss a broad range of possible approaches and evaluate them with regard to efficiency and effectiveness on a range of real-world and synthetic networks. Furthermore, we show that the presence of a community and anti-community structure are not mutually exclusive, and that even networks with a strong traditional community structure may also contain anti-communities.

CONFERENCE OFFICERS

General Chair
Johann Gamper (Free University of Bozen-Bolzano)

Program Chair
Michael Böhlen (University of Zurich)

Proceeding Chair
Dimitris Sacharidis (TU Vienna)

Demo Chair
Peer Kröger (LMU Munich)

Local Arrangements Chairs
Anton Dignös (Free University of Bozen-Bolzano)
Patrick Ohnewein (IDM Südtirol / Alto Adige)

Web Chairs
Theodoros Chondrogiannis (University of Konstanz)
Luca Miotto (IDM Südtirol / Alto Adige)

Honorary Chair
Arie Shoshani (Lawrence Berkeley National Laboratory)

Steering Committee
Arie Shoshani (Lawrence Berkeley National Laboratory), chair
Torben Bach Pedersen (Aalborg University)
Magdalena Balazinska (University of Washington)
Amarnath Gupta (University of California San Diego)
Ioana Manolescu (Inria-Paris)

COMMITTEES

Program Committee
Amr Magdy (University of California Riverside)
Andreas Züfle (George Mason University)
Barbara Catania (University of Genova)
Bertram Ludaescher (University of Illinois)
Bongki Moon (Seoul National University)
Christophe Claramunt (Naval Academy, France)
Christos Doulkeridis (University of Pireaus)
Elio Masciari (CNR Italy)
Eric Stephan (Pacific Northwest National Laboratory)
Erich Schubert (University of Heidelberg)
Filippo Furfaro (University of Calabria)
Florin Rusu (UC Merced)
Gagan Agrawal (Ohio State University)
Giovanna Guerrini (University of Genova)
Goce Trajcevski (Northwestern University)
Hannes Mühleisen (CWI)
Istavan Csabai (Eötvös University)
Jagan Sankaranarayanan (Google)
Jeffrey Yu (Chinese University of Hong Kong)
Joerg Sander (University of Alberta)
Kurt Stockinger (ZHAW)
Lukasz Golab (University of Waterloo)
Manolis Terrovitis (IMIS Athena RC)
Manos Athanassoulis (Harvard SEAS)
Markus Schneider (University of Florida)
Martin Schäler (Karlsruhe Institute of Technology)
Matthias Schubert (LMU Munich)
Michael Gertz (Heidelberg University)
Nesime Tatbul (Intel Labs and MIT)
Peiquan Jin (University of Science and Technology of China)
Periklis Andritsos (University of Toronto)
Peter Baumann (Jacobs University Bremen)
Qiang Zhu (University of Michigan - Dearborn)
Qing Liu (New Jersey Institute of Technology)
Saiful Islam (Griffith University)
Thomas Brinkhoff (Jade University Oldenburg)
Thomas Heinis (Imperial College)
Torben Pedersen (Aalborg University)
Vassilis Tsotras (University of California, Riverside)
Wolfgang Lehner (TU Dresden)
Yi-Cheng Tu (University of South Florida)
Yongluan Zhou (University of Copenhagen)

Demo Committee
Arthur Zimek (University of Southern Denmark)
Ilkcan Keles (Aalborg University)
Mateusz Pawlik (University of Salzburg)
Pavlos Paraskevopoulos (George Mason University)
Thi Thao Nguyen (Aalborg University)
Michael Shekelyan (Free University of Bozen-Bolzano)

Best Paper Award Committee
Bongki Moon (Seoul National University), chair
John Wu (Berkeley Lab)
Periklis Andritsos (University of Toronto)

NOTES