



# Proceedings of the VLDB Endowment

Volume 15, No. 8 – April 2022

Editors in Chief:

**Fatma Özcan, Juliana Freire and Xuemin Lin**

Associate Editors:

**Arun Kumar, Azza Abouzied, Beng Chin Ooi, Boris Glavic, Dan Suciu,  
Divyakant Agrawal, Eugene Wu, Georgia Koutrika, Ioana Manolescu,  
Jeffrey Xu Yu, Julia Stoyanovich, Jun Yang, K. Selçuk Candan,  
Khuzaima Daudjee, Laure Berti-Equille, Lei Chen, Mohamed Mokbel,  
Neoklis Polyzotis, Paolo Papotti, Peter Boncz, Sebastian Schelter,  
Sourav S Bhowmick, Surajit Chaudhuri, Themis Palpanas, Vanessa Braganholo,  
Viktor Leis, Wang-Chiew Tan, Wenjie Zhang, Wook-Shin Han, Xiaofang Zhou**

Publication Editors:

**Lijun Chang and Xin Cao**

PVLDB – Proceedings of the VLDB Endowment

Volume 15, No. 8, April 2022.

All papers published in this issue will be presented at the 48th International Conference on Very Large Data Bases, Sydney, Australia, 2022.

## **Copyright 2022 VLDB Endowment**

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org](mailto:info@vldb.org).

Volume 15, Number 8, April 2022

Pages i – vii and 1519 - 1738

ISSN 2150-8097

Available at: <http://www.pvldb.org> and <https://dl.acm.org/journal/pvldb>

## TABLE OF CONTENTS

### Front Matter

|                                                     |    |
|-----------------------------------------------------|----|
| Copyright Notice .....                              | i  |
| Table of Contents .....                             | ii |
| PVLDB Organization and Review Board – Vol. 15 ..... | iv |

### Research Papers

|                                                                                                                                                              |      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ForBackBench: A Benchmark for Chasing vs. Query-Rewriting .....                                                                                              | 1519 |
| <i>Afnan G Alhazmi, Tom Blount, George Konstantinidis</i>                                                                                                    |      |
| Accurate Summary-based Cardinality Estimation Through the Lens of Cardinality Estimation Graphs .....                                                        | 1533 |
| <i>Jeremy Chen, Yuqing Huang, Mushi Wang, Semih Salihoglu, Kenneth Salem</i>                                                                                 |      |
| Distributed D-core Decomposition over Large Directed Graphs .....                                                                                            | 1546 |
| <i>Xuankun Liao, Qing Liu, Jiaxin Jiang, Xin Huang, Jianliang Xu, Byron Choi</i>                                                                             |      |
| Efficient Maximal Biclique Enumeration for Large Sparse Bipartite Graphs .....                                                                               | 1559 |
| <i>Lu Chen, Chengfei Liu, Rui Zhou, Jiajie Xu, Jianxin Li</i>                                                                                                |      |
| TGL: A General Framework for Temporal GNN Training on Billion-Scale Graphs .....                                                                             | 1572 |
| <i>Hongkuan Zhou, Da Zheng, Israt Nisa, Vassilis N. Ioannidis, Xiang Song, George Karypis</i>                                                                |      |
| Distributed Learning of Fully Connected Neural Networks using Independent Subnet Training .....                                                              | 1581 |
| <i>Binhang Yuan, Cameron Wolfe, Chen Dun, Yuxin Tang, Anastasios Kyrillidis, Chris Jermaine</i>                                                              |      |
| Netherite: Efficient Execution of Serverless Workflows .....                                                                                                 | 1591 |
| <i>Sebastian C Burckhardt, Badrish Chandramouli, Chris Gillum, David A Justo, Konstantinos Kallas, Connor McMahon, Christopher Meiklejohn, Xiangfeng Zhu</i> |      |
| Endure: A Robust Tuning Paradigm for LSM Trees Under Workload Uncertainty .....                                                                              | 1605 |
| <i>Andy Huynh, Harshal Chaudhari, Evimaria Terzi, Manos Athanassoulis</i>                                                                                    |      |
| An I/O-Efficient Disk-based Graph System for Scalable Second-Order Random Walk of Large Graphs .....                                                         | 1619 |
| <i>Hongzheng Li, Yingxia Shao, Junping Du, Bin Cui, Lei Chen</i>                                                                                             |      |
| SNARF: A Learning-Enhanced Range Filter .....                                                                                                                | 1632 |
| <i>Kapil Vaidya, Tim Kraska, Subarna Chatterjee, Eric R Knorr, Michael Mitzenmacher, Stratos Idreos</i>                                                      |      |
| DLCR: Efficient Indexing for Label-Constrained Reachability Queries on Large Dynamic Graphs .....                                                            | 1645 |
| <i>Xin Chen, You Peng, Sibor Wang, Jeffrey Xu Yu</i>                                                                                                         |      |
| QueryFormer: A Tree Transformer Model for Query Plan Representation .....                                                                                    | 1658 |
| <i>Yue Zhao, Gao Cong, Jiachen Shi, Chunyan Miao</i>                                                                                                         |      |
| Index Checkpoints for Instant Recovery in In-Memory Database Systems .....                                                                                   | 1671 |
| <i>Leon Lee, Siphrey Xie, Yunus Ma, Shimin Chen</i>                                                                                                          |      |

|                                                                                               |      |
|-----------------------------------------------------------------------------------------------|------|
| MATE: Multi-Attribute Table Extraction.....                                                   | 1684 |
| <i>Mahdi Esmailoghli, Jorge Arnulfo Quiane Ruiz, Ziawasch Abedjan</i>                         |      |
| TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection .....     | 1697 |
| <i>John Paparrizos, Yuhao Kang, Paul Boniol, Ruey Tsay, Themis Palpanas, Michael Franklin</i> |      |
| A Critical Re-evaluation of Neural Methods for Entity Alignment.....                          | 1712 |
| <i>Manuel Leone, Stefano Huber, Akhil Arora, Alberto Garcia-duran, Robert West</i>            |      |
| Analyzing How BERT Performs Entity Matching.....                                              | 1726 |
| <i>Matteo Paganelli, Francesco Del Buono, Andrea Baraldi, Francesco Guerra</i>                |      |

## **PVLDB ORGANIZATION AND REVIEW BOARD - Vol. 15**

### **Editors in Chief of PVLDB**

Fatma Ozcan (Google)  
Juliana Freire (New York University)  
Xuemin Lin (University of New South Wales)

### **Associate Editors of PVLDB**

Arun Kumar (University of California, San Diego)  
Azza Abouzied (NYU Abu Dhabi)  
Beng Chin Ooi (NUS)  
Boris Glavic (Illinois Institute of Technology)  
Dan Suciu (University of Washington)  
Divyakant Agrawal (University of California, Santa Barbara)  
Eugene Wu (Columbia University)  
Georgia Koutrika (ATHENA)  
Ioana Manolescu (INRIA and Institut Polytechnique de Paris)  
Jeffrey Xu Yu (Chinese University of Hong Kong)  
Julia Stoyanovich (New York University)  
Jun Yang (Duke University)  
K. Seçuk Candan (Arizona State University)  
Khuzaima Daudjee (University of Waterloo)  
Laks Lakshmanan (The University of British Columbia)  
Laure Berti-Equille (IRD)  
Lei Chen (Hong Kong University of Science and Technology)  
Mohamed Mokbel (University of Minnesota, Twin Cities)  
Neoklis Polyzotis (Google)  
Paolo Papotti  
Peter Boncz (CWI)  
Sebastian Schelter (University of Amsterdam)  
Sharad Mehrotra (U.C. Irvine)  
Sourav S Bhowmick (Nanyang Technological University)

Surajit Chaudhuri (Microsoft Research)  
Themis Palpanas (University of Paris)  
Vanessa Braganholo (Fluminense Federal University)  
Viktor Leis (Friedrich Schiller University Jena)  
Wang-Chiew Tan (Megagon Labs)  
Wenjie Zhang (University of New South Wales)  
Wook-Shin Han (POSTECH)  
Xiaofang Zhou (Hong Kong University of Science and Technology)

### **Publication Editors**

Lijun Chang (University of Sydney)  
Xin Cao (University of New South Wales)

### **PVLDB Managing Editor**

Wolfgang Lehner (Dresden University of Technology)

### **PVLDB Advisory Committee**

Felix Naumann (HPI)  
Juliana Freire (New York University)  
Xuemin Lin (U of New South Wales)  
Georgia Koutrika (Athena Research Center)  
Jun Yang (Duke University)  
Vanessa Braganholo (Universidade Federal Fluminense)  
Sourav S Bhowmick (Nanyang Technological University)  
Chris Jermaine (Rice University)  
Peter Triantafillou (University of Warwick)  
Xin Luna Dong (Facebook)  
Fatma Ozcan (Google)  
Lei Chen (Hong Kong University of S&T)  
Graham Cormode (University of Warwick)  
Divesh Srivastava (AT&T Labs-Research)  
Wolfgang Lehner (TU Dresden)

## Review Board

Abolfazl Asudeh (University of Michigan)  
Aécio Santos (New York University)  
Ahmed Eldawy (University of California, Riverside)  
Alexander Hall (RelationalAI)  
Alexander J Ratner (University of Washington)  
Aline Bessa (New York University)  
Alkis Simitsis (Athena Research Center)  
Altigran da Silva (Universidade Federal do Amazonas)  
AnHai Doan (University of Wisconsin-Madison)  
Anna Fariha (Microsoft)  
Anton Dignös (Free University of Bozen-Bolzano)  
Antonio Cavalcante Araujo Neto (University of Alberta)  
Arijit Khan (Nanyang Technological University)  
Arvind Arasu (Microsoft)  
Babak Salimi (University of California, San Diego)  
Bailu Ding (Microsoft Research)  
Bertram Ludaescher (University of Illinois)  
Bolong Zheng (Huazhong University of Science and Technology)  
Brandon Haynes (Gray Systems Lab, Microsoft)  
Byron Choi (Hong Kong Baptist University)  
Carlo Curino (Microsoft -- GSL)  
Carlos Scheidegger (The University of Arizona)  
Carsten Binnig (TU Darmstadt)  
Ce Zhang (ETH)  
Cheng Long (Nanyang Technological University)  
Chengfei Liu (Swinburne University of Technology)  
Chuan Lei (Instacart)  
Chunbin Lin (Amazon AWS)  
Curtis Dyreson (Utah State University)  
Dan Kifer (Pennsylvania State University)  
Dana M Van Aken (Carnegie Mellon University)  
Daniel Deutch (Tel Aviv University)  
Daniel Oliveira (UFF, Brazil)  
David Koop (Northern Illinois University)  
Davide Mottin (Aarhus University)  
Dong Xie (Penn State University)  
Eduardo Ogasawara (CEFET-RJ)  
Eleni Tzirita Zacharitou (TU Berlin)  
Fabio Porto (LNCC)  
Faisal Nawab (University of California at Irvine)  
Fan Zhang (Guangzhou University)  
Fatemeh Nargesian (University of Rochester)  
Fei Chiang (McMaster University)  
Florin Rusu (UC Merced)  
Floris Geerts (University of Antwerp)  
Fotis Psallidas (Microsoft)  
George Fletcher (Eindhoven University of Technology)  
George Papadakis (University of Athens)  
Gerhard Weikum (Max-Planck-Institut für Informatik)  
Germain Forestier (University of Haute Alsace)  
Guoliang Li (Tsinghua University)  
Haipeng Dai (Nanjing University)  
Harish Doraiswamy (Microsoft Research India)  
Heiko Mueller (DeepReason.ai)  
Herodotos Herodotou (Cyprus University of Technology)

Holger Pirk (Imperial College)  
Hongzhi Yin (The University of Queensland)  
Huiping Cao (New Mexico State University)  
Immanuel Trummer (Cornell)  
Ioana Manolescu (INRIA and Institut Polytechnique de Paris)  
Ippokratis Pandis (Amazon)  
Ishtiyaque Ahmad (University of California, Santa Barbara)  
Jae-Gil Lee (KAIST)  
Jana Giceva (TU Munich)  
Jeffrey Xu Yu (Chinese University of Hong Kong)  
Jens Teubner (TU Dortmund University)  
Jia Zou (Arizona State University)  
Jian Pei (Simon Fraser University)  
Jianguo Wang (Purdue University)  
Jiannan Wang (Simon Fraser University)  
Jianxin Li (Deakin University)  
Jianye Yang (Central South University)  
Jiwon Seo (Hanyang University)  
Johannes Gehrke (Microsoft)  
Jorge Arnulfo Quijane Ruiz (TU Berlin)  
Joseph Near (University of Vermont)  
Junhu Wang (Griffith University)  
Kaiping Zheng (National University of Singapore)  
Kangfei Zhao (The Chinese University of Hong Kong)  
Karima Echihabi (Mohammed VI Polytechnic University)  
Katja Hose (Aalborg University)  
Kenneth A Ross (Columbia University)  
Kostas Zoumpatianos (Snowflake Computing)  
Lei Zou (Peking University)  
Leopoldo Bertossi (Universidad Adolfo Ibáñez)  
Li Xiong (Emory University)  
Lianke Qin (University of California, Santa Barbara)  
Lijun Chang (The University of Sydney)  
Lin Ma (Carnegie Mellon University)  
Long Yuan (Nanjing University of Science and Technology)  
Lu Qin (UTS)  
Luciano Barbosa (Universidade Federal de Pernambuco)  
Marcelo Arenas (Universidad Católica & IMFD)  
Maria Luisa Sapino (U. Torino)  
Matteo Lissandrini (Aalborg University)  
Matthias Boehm (Graz University of Technology)  
Matthias Renz (University of Kiel)  
Max Heimerl (Snowflake)  
Maximilian Schleich (University of Washington)  
Meihui Zhang (Beijing Institute of Technology)  
Melanie Herschel (Universität Stuttgart)  
Michael Abebe (University of Waterloo)  
Min Xie (Instacart)  
Mirella M Moro (Universidade Federal de Minas Gerais)  
Mohamed Sarwat (Arizona State University)  
Mohammad Dashti (MongoDB)  
Mohammad Javad Amiri (University of Pennsylvania)  
Mohammad Sadoghi (University of California, Davis)  
Muhammad Aamir Cheema (Monash University)

Nikita Bhutani (Megagon Labs)  
 Oliver A Kennedy (University at Buffalo, SUNY)  
 Panos K. Chrysanthis (University of Pittsburgh)  
 Paolo Missier (Newcastle University)  
 Parth Nagarkar (NMSU)  
 Paul Groth (University of Amsterdam)  
 Peng CHENG (East China Normal University)  
 Peter Pietzuch (Imperial College London)  
 Pierangela Samarati (Universita delgi Studi di Milano)  
 Pinar Karagoz (METU, Turkey)  
 Pinar Tozun (IT University of Copenhagen)  
 Prithu Banerjee (UBC)  
 Raoni Lourenço (New York University)  
 Raul Castro Fernandez (UChicago)  
 Ravi Ramamurthy (Microsoft)  
 Raymond Chi-Wing Wong (Hong Kong University of Science and Technology)  
 Renata Borovica-Gajic (University of Melbourne)  
 Reynold Cheng (The University of Hong Kong)  
 Rui Mao (Shenzhen University)  
 Ruoming Jin (Kent State University)  
 Sai Wu (Zhejiang University)  
 Sainyam Galhotra (University of Chicago)  
 Sanjay Krishnan (University of Chicago)  
 Sanjib Kumar Das (Google)  
 Sayan Ranu (IIT Delhi)  
 Sebastian Link (University of Auckland)  
 Semih Salihoglu (University of Waterloo)  
 Senjuti Basu Roy (New Jersey Institute of Technology)  
 Sergey Melnik (Google)  
 Shantanu Sharma (New Jersey Institute of Technology)  
 Shaoxu Song (Tsinghua University)  
 Sheng Wang (New York University)  
 Shimin Chen (Chinese Academy of Sciences)  
 Shumo Chu (University of California, Santa Barbara)  
 Shweta Jain (University of Illinois, Urbana-Champaign)  
 Sibo Wang (The Chinese University of Hong Kong)  
 Srinivasan Keshav (University of Cambridge)  
 Steffen Zeuch (DFKI GmbH)  
 Steven E Whang (KAIST)  
 Subarna Chatterjee (Harvard University)  
 Sudip Roy (Google)  
 Supun C Nakandala (University of California, San Diego)  
 Tamer Özsu (University of Waterloo)  
 Tarique A Siddiqui (Microsoft Research)  
 Thomas Heinis (Imperial College)  
 Thomas Neumann (TUM)  
 Tianzheng Wang (Simon Fraser University)  
 Tien Tuan Anh Dinh (Singapore University of Technology and Design)  
 Tilmann Rabl (HPI, University of Potsdam)  
 Ting Yu (Qatar Computing Research Institute)  
 Torben Bach Pedersen (Aalborg University)  
 Torsten Grust (Universität Tübingen)  
 Umar Farooq Minhas (Microsoft Research)  
 Vasiliki Kalavri (Boston University)  
 Verena Kantere (National Technical University of Athens)  
 Victor Zakhary (Oracle)  
 Vivek Narasayya (Microsoft Research)  
 Vraj Shah (University of California, San Diego)  
 Walid G Aref (Purdue)  
 Wasay Abdul (Harvard)  
 Wei Wang (Hong Kong University of Science and Technology (Guangzhou))  
 Wei Lu (Renmin university of china)  
 Weiren Yu (University of Warwick)  
 Wen Hua (The University of Queensland)  
 Wolfgang Lehner (TU Dresden)  
 Xi He (University of Waterloo)  
 Xiang Lian (Kent State University)  
 Xiao Qin (IBM Research)  
 Xiaofei Zhang (University of Memphis)  
 Xiaokui Xiao (National University of Singapore)  
 Xiaolan Wang (Megagon Labs)  
 Xiaoyang Wang (Zhejiang Gongshang University)  
 Xin Huang (Hong Kong Baptist University)  
 Yael Amsterdamer (Bar-Ilan university)  
 Yanyan Shen (Shanghai Jiao Tong University)  
 Ye Yuan (Northeastern University)  
 Yeye He (Microsoft Research)  
 Yi Chen (NJIT)  
 Yi Lu (MIT)  
 Yikai Zhang (Chinese University of Hong Kong)  
 Yinan Li (Microsoft Research)  
 Ying Zhang (University of Technology Sydney)  
 Yongxin Tong (Beihang University)  
 Yuanyuan Zhu (Wuhan University)  
 Yue Wang (Shenzhen Institute of Computing Sciences, Shenzhen University)  
 Yufei Tao (Chinese University of Hong Kong)  
 Yuliang Li (Megagon Labs)  
 Yuncheng Wu (National University of Singapore)  
 Yunjun Gao (Zhejiang University)  
 Yuval Moskovitch (University of Michigan)  
 Zhifeng Bao (RMIT University)  
 Zhongle Xie (Zhejiang University)  
 Zi Huang (University of Queensland)  
 Ziawasch Abedjan (Leibniz Universität Hannover)  
 Zohar Karnin (Amazon)  
 Zsolt István (IT University of Copenhagen)

## LETTER FROM THE EDITORS IN CHIEF

We are pleased to present the eighth issue of PVLDB, Volume 15. This issue contains 17 papers in total including 10 regular research papers, 2 scalable data science (SDS) papers, and 5 experiments analysis & benchmark (EA&B) papers. A broad range of topics are covered in this issue including distributed database systems, machine learning & applied AI for data management, graph data management, database performance, database engines, data quality, and data mining.

For the first paper in this issue, Chen et al. present experimental studies of summary-based cardinality estimators in graph database management systems. Next, Alhazmi et al. present a benchmark for chasing vs. query-rewriting. Liao et al. propose efficient solutions for distributed D-core decomposition problem over large directed graphs. Chen et al. study the maximal biclique enumeration problem on large sparse bipartite graphs and propose efficient algorithms. Zhou et al. propose a general framework for temporal GNN training on billion-scale graphs. Yuan et al. study the distributed learning problem of fully connected neural networks using independent subnet training. Burckhardt et al. introduce Netherite, a novel architecture for executing serverless workflows on an elastic cluster. Huynh et al. present Endure, a new paradigm for tuning LSM trees in the presence of workload uncertainty. Li et al. introduce an I/O-efficient disk-based graph system for scalable second-order random walk of large graphs. Vaidya et al. present Sparse Numerical Array-Based Range Filters (SNARF), a learned range filter that efficiently supports range queries for numerical data. Chen et al. propose DLCR to support efficient label-constrained reachability queries on large dynamic graphs. Zhao et al. propose a tree transformer model for query plan representation. Lee et al. investigate index checkpoints for instant recovery in in-memory database systems. Esmailoghli et al. introduce MATE, a table discovery system that leverages a novel hash-based index. Paparrizos et al. present an end-to-end benchmark suite for univariate time-series anomaly detection. Leone et al. comprehensively re-evaluate neural methods for entity alignment. Paganelli et al. perform a multi-facet analysis of the components of pre-trained and fine-tuned BERT architectures applied to entity matching.

All the papers in this issue will be presented at the 48th International Conference on Very Large Data Bases, 2022, in Sydney. We sincerely thank all the authors for submitting their work and all the reviewers for their outstanding service in reviewing the submissions. We hope that the reader will find this volume enjoyable.

Fatma Özcan, Juliana Freire, and Xuemin Lin  
Editors-in-Chief of PVLDB Volume 15  
Program Chairs for VLDB 2022