Da Yan

Curriculum Vitae

Rm 116, Ho Sin-Hang Engineering Building
The Chinese University of Hong Kong
Shatin N.T., Hong Kong
(+852) 6431 6958
yanda@cse.cuhk.edu.hk
www.cse.cuhk.edu.hk/∼yanda

Education

2009–2014 **Ph.D. in Computer Science**, *The Hong Kong University of Science & Technology*, Clear Water Bay, Kowloon.

2005–2009 **B.S. in Computer Science**, *Fudan University*, Shanghai, *GPA – 3.7/4.0 (Top 5%)*.

Research Interests

Distributed Systems & Big Data Analytics (*Mainly on Graph Data*)

Spatial Data Management

Uncertain Data Management

Research Experience

2014–Present **Postdoctoral Fellow**, *Department of Computer Science & Engineering*, The Chinese University of Hong Kong, Shatin N.T., Hong Kong.

2007–2009 **Undergraduate Research Assistant**, *ADMIS (Advanced Data Management and Information System) Lab*, Fudan University, Shanghai.

Publications

- Yi Yang, *Da Yan*, Huanhuan Wu, James Cheng, Shuigeng Zhou and John C.S. Lui. **Diversified Temporal Subgraph Pattern Mining.** *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining* (KDD), 2016.
- Cheng Chen, Hejun Wu, *Da Yan* and James Cheng. **SGraph: A Distributed Streaming System For Processing Big Graphs.** *International Conference on Big Data Computing and Communications* (BIGCOM), 2016.
- *Da Yan*, Yingyi Bu, Yuanyuan Tian, Amol Deshpande and James Cheng. **Big Graph Analytics Systems.** *ACM International Conference on Management of Data* (SIGMOD), 2016.
- Qizhen Zhang, *Da Yan* and James Cheng. **Quegel: A General-Purpose System for Querying Big Graphs.** *ACM International Conference on Management of Data* (SIGMOD), 2016.
- *Da Yan*, James Cheng, M. Tamer Özsu, Fan Yang, Yi Lu, John C.S. Lui, Qizhen Zhang and Wilfred Ng. **A General-Purpose Query-Centric Framework for Querying Big Graphs.** *Proceedings of the VLDB Endowment* (PVLDB), 2016.

- *Da Yan*, James Cheng, Yi Lu and Wilfred Ng. **Effective Techniques for Message Reduction and Load Balancing in Distributed Graph Computation.** *World Wide Web Conference* (WWW), pages 1307–1317, 2015.
- *Da Yan*, Zhou Zhao and Wilfred Ng. **Efficient Processing of Optimal Meeting Point Queries in Euclidean Space and Road Networks.** *Knowledge and Information Systems* (KAIS), 42(2):319–351, 2015.
- *Da Yan*, James Cheng, Zhou Zhao and Wilfred Ng. **Efficient Location-based Search of Trajectories with Location Importance.** *Knowledge and Information Systems* (KAIS), 45(1):215–245, 2015.
- Yi Lu, James Cheng, *Da Yan* and Huanhuan Wu. **Large-Scale Distributed Graph Computing Systems: An Experimental Evaluation.** *Proceedings of the VLDB Endowment* (PVLDB), 8(3):281–292, 2015.
- Huanhuan Wu, James Cheng, Yi Lu, Yiping Ke, Yuzhen Huang, *Da Yan* and Hejun Wu. **Core Decomposition in Large Temporal Graphs.** *IEEE BigData*, 2015.
- *Da Yan*, Zhou Zhao, Wilfred Ng and Steven Liu. **Probabilistic Convex Hull Queries over Uncertain Data.** Accepted and to appear in *IEEE Transactions on Knowledge and Data Engineering* (TKDE), 2014.
- *Da Yan*, James Cheng, Yi Lu and Wilfred Ng. **Blogel: A Block-Centric Framework for Distributed Computation on Real-World Graphs.** *Proceedings of the VLDB Endowment* (PVLDB), 7(14):1981–1992, 2014.
- *Da Yan*, James Cheng, Kai Xing, Yi Lu, Wilfred Ng and Yingyi Bu. **Pregel Algorithms for Graph Connectivity Problems with Performance Guarantees.** *Proceedings of the VLDB Endowment* (PVLDB), 7(14):1821–1832, 2014.
- Zhou Zhao, *Da Yan* and Wilfred Ng. **Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases.** *IEEE Transactions on Knowledge and Data Engineering* (TKDE), 26(5):1171–1184, 2014.
- Zhou Zhao, *Da Yan*, Wilfred Ng and Shi Gao. **A Transfer-Learning Based Framework of Crowd-Selection on Twitter.** In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD), pages 1514–1517, 2013.
- *Da Yan*, James Cheng, Wilfred Ng and Steven Liu. **Finding Distance-Preserving Subgraphs in Large Road Networks.** In *IEEE International Conference on Data Engineering* (ICDE), pages 625–636, 2013.
- *Da Yan*, Zhou Zhao and Wilfred Ng. **Leveraging Read Rates of Passive RFID Tags for Real-Time Indoor Location Tracking.** In *ACM Conference on Information and Knowledge Management* (CIKM), pages 375–384, 2012.
- *Da Yan*, Zhou Zhao and Wilfred Ng. **Monochromatic and Bichromatic Reverse Nearest Neighbor Queries on Land Surfaces.** In *ACM Conference on Information and Knowledge Management* (CIKM), pages 942–951, 2012.
- Zhou Zhao, *Da Yan* and Wilfred Ng. **Mining Probabilistically Frequent Sequential Patterns in Uncertain Databases.** In *International Conference on Extending Database Technology* (EDBT), 2012.

- Zhou Zhao, *Da Yan* and Wilfred Ng. **A Probabilistic Convex Hull Query Tool for Animal Tracking.** In *International Conference on Extending Database Technology* (EDBT), pages 74–85, 2012.
- *Da Yan*, Raymond Chi-Wing Wong and Wilfred Ng. **Efficient Methods for Finding Influential Locations with Adaptive Grids.** In *ACM Conference on Information and Knowledge Management* (CIKM), pages 1475–1484, 2011.
- *Da Yan*, Zhou Zhao and Wilfred Ng. **Efficient Algorithms for Finding Optimal Meeting Point on Road Networks.** *Proceedings of the VLDB Endowment* (PVLDB), pages 968–979, 2011.
- *Da Yan* and Wilfred Ng. **Robust Ranking of Uncertain Data.** In *Database Systems for Advanced Applications* (DASFAA), pages 254–268, 2011. **The Best Paper Award.**

Projects

I am the Co-Principal Investigator (Co-PI) of the following project funded by the Innovation and Technology Fund (ITF) from 01/07/2015 to 31/12/2016 with an amount of HK$1,410,550.

A Scalable System for Companies to Manage and Analyze Their Temporal Network.
Project Reference: ITS/246/14

I also led the following projects on distributed graph computing systems (also summarized in `http://www.cse.cuhk.edu.hk/systems/graph`). I was the main contributor to the system design and wrote all the system codes, while I also implemented critical application codes and coordinated the work among other members.

Blogel: `http://www.cse.cuhk.edu.hk/blogel`.
Blogel is a block-centric graph computing framework, where programmers think like a block when developing their applications. Blogel naturally handles the three adverse characteristics of real graphs: (1)skewed degree distribution, (2)large diameter, and (3)(relatively) high density. Experiments on large real-world graphs verified that Blogel is able to achieve orders of magnitude performance improvements over the state-of-the-art graph computing systems (e.g., Giraph, GraphLab, GPS, GraphX, Giraph++, GraphChi, etc.).

Pregel+: `http://www.cse.cuhk.edu.hk/pregelplus`.
Pregel+ improves the message passing model of Google's Pregel by introducing two effective message reduction techniques: (1)vertex mirroring and (2)a new request-respond. These two techniques address the communication bottleneck and the corresponding imbalanced workload of existing Pregel-like systems. Extensive experiments over various large real graphs show that Pregel+ is significantly more efficient than the state-of-the-art graph computing systems, especially for processing power-law graphs and dense graphs.

Quegel: `http://www.cse.cuhk.edu.hk/quegel`.
Quegel allows users to write Pregel-style algorithms for efficient online graph querying. Quegel supports both interactive querying by a small number of users or high-throughput batch processing of a large number of graph queries, and processes light-workload graph queries on demand using a novel superstep-sharing execution model that better utilizes the cluster resources. Quegel further provides convenient interfaces for constructing indexes to significantly improve query performance, which existing graph-parallel systems do not support.

GraphD: `http://www.cse.cuhk.edu.hk/systems/graphd`.
GraphD is designed to process very large graphs with a small cluster, for which existing in-memory Pregel-like systems would run out of memory. As an illutration, on our 15-machine cluster, it takes merely 250 seconds in each superstep to compute PageRank on the ClueWeb graph with 978,408,098 nodes and 42,574,107,469 links. The key idea is to stream edges and messages on local disks, whose cost is hidden inside the cost of message transmission by parallelism. The performance of GraphD is comparable to that of an in-memory Pregel-like system running with sufficient memory space.

LWCP: `http://www.cse.cuhk.edu.hk/pregelplus/ft.html`.
Lightweight checkpointing (LWCP) is a novel fault tolerance mechanism for Pregel-like systems. It avoids writing edges and messages into every checkpoint. Topology data are handled by incremental checkpointing, while messages are generated from checkpointed vertex states online to continue the processing. LWCP is tens of times faster than writing a conventional checkpoint, and can also be integrated with log-based recovery to reduce recovery time.

Invited Talks

Jun 2, 2016 **Systems for Big Graph Analytics**, *City University of Hong Kong*, Kowloon Tong, Hong Kong.

Feb 29, 2016 **Systems and Algorithms for Big Data Analytics**, *University of California, Riverside (UCR)*, California, USA.

Feb 22, 2016 **Systems and Algorithms for Big Data Analytics**, *Michigan Technological University*, Houghton, Michigan.

Feb 9, 2016 **Systems and Algorithms for Big Data Analytics**, *University of Notre Dame*, Notre Dame, Indiana.

Nov 25, 2015 **Systems and Algorithms for Big Data Analytics**, *Hong Kong Baptist University (HKBU)*, Kowloon Tong, Hong Kong.

Nov 14, 2014 **Distributed Graph Computing: Algorithms and Systems**, *Sun Yat-Sen University (SYSU)*, Guangzhou, China.

May 23, 2014 **Distributed Graph Computing Beyond Pregel**, *Alibaba Xixi Park (Taobao City)*, Hangzhou, China.

May 8, 2014 **Distributed Graph Computing: Algorithms and Systems**, *National Institute of Informatics (NII)*, Tokyo, Japan.

Dec, 2012 **Query Processing in Spatial Databases**, *East China Normal University (ECNU)*, Shanghai, China.

Program Committee Membership

2016 IEEE International Conference on Parallel and Distributed Systems (ICPADS 2016)

2016 The First International Workshop on Graph Analytics and Query Processing (GAP 2016)

Journal Review

2016 IEEE Transactions on Parallel and Distributed Systems (TPDS)

2014, 2016 IEEE Transactions on Knowledge and Data Engineering (TKDE)

2015 World Wide Web Journal (WWWJ)

2014-2016 Knowledge and Information Systems (KAIS)

External Review

2013, 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD)

2013, 2014 International Conference on Very Large Data Bases (VLDB)

2013, 2015 IEEE International Conference on Data Engineering (ICDE)

2010, 12, 14 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)

2013 IEEE International Conference on Data Mining (ICDM)

2013 International Conference on Extending Database Technology (EDBT)

2013–15 ACM International Conference on Information and Knowledge Management (CIKM)

2013–15 International Conference on Database Systems for Advanced Applications (DASFAA)

2013, 2014 IEEE International Conference on Big Data (IEEE Big Data)

2013, 2014 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

2013 International Conference on Conceptual Modeling (ER)

2015 International Conference on Web-Age Information Management (WAIM)

2014, 2015 Asia-Pacific Web Conference (APWeb)

2010, 11, 13 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)

2013, 2014 International Workshop on Social Network Mining and Analysis (SNAKDD)

2010, 2012 International Conference on Database and Expert Systems Applications (DEXA)

2012, 2013 International Conference on Web Information System Engineering (WISE)

2014 IEEE International Conference on Big Data Science and Engineering (BDSE)

2011 Information Systems Journal (IS)

Teaching Experience

Fall 2012 **COMP4431: Multimedia Computing**, *The Hong Kong University of Science & Technology*, Teaching Assistant.

Spring 2011 **COMP231: Database Management Systems**, *The Hong Kong University of Science & Technology*, Teaching Assistant.

Fall 2010 **COMP570: Introduction to Advanced Algorithmic Techniques**, *The Hong Kong University of Science & Technology*, Teaching Assistant.

Spring 2010 **COMP152: Object-Oriented Programming and Data Structures**, *The Hong Kong University of Science & Technology*, Teaching Assistant.

Fall 2009 **COMP104: Programming Fundamentals and Methodology**, *The Hong Kong University of Science & Technology*, Teaching Assistant.

Awards

2015 HKIS-Towngas 2015 Young Scientist Award in Physical/Mathematical Science

2013 ICDE 2013 Student Travel Award

2011 DASFAA 2011 Best Paper Award

2009 Fudan Excellent Graduate

2008–2009 Fudan Undergraduate Scholarship

2008 Excellent League Member Award, School of Information Science and Engineering, Fudan University

2007–2008 SCSK Corporation Scholarship, Fudan University

2007–2008 Fudan Excellent Student

2007–2008 Fudan Excellent Undergraduate Scholarship

2006–2007 Fudan Undergraduate Scholarship

2005–2006 Fudan Undergraduate Scholarship

2004 China Utility Model Patent: Retractable Washing Line (ZL 2004 2 0020359.7)

Referees

- **Wilfred Ng** (Ph.D. Supervisor). Associate Professor. *Department of Computer Science & Engineering. The Hong Kong University of Science & Technology.* Email: wilfred@cs.ust.hk. Homepage: `http://www.cs.ust.hk/~wilfred`.
- **James Cheng**. Assistant Professor. *Department of Computer Science & Engineering. The Chinese University of Hong Kong.* Email: jcheng@cse.cuhk.edu.hk. Homepage: `www.cse.cuhk.edu.hk/~jcheng`.
- **John C.S. Lui**. Choh-Ming Li Professor of Computer Science and Engineering. Fellow of ACM, Fellow of IEEE, Croucher Senior Research Fellow. *Department of Computer Science & Engineering. The Chinese University of Hong Kong.* Email: cslui@cse.cuhk.edu.hk. Homepage: `www.cs.cuhk.hk/~cslui`.
- **Raymond Chi-Wing Wong**. Associate Professor. *Department of Computer Science & Engineering. The Hong Kong University of Science and Technology.* Email: raywong@cse.ust.hk. Homepage: `www.cse.ust.hk/~raywong`.

English Scores

GRE 660 (Verbal) + 800 (Quantitative) + 5.0 (Analytical Writing)

TOEFL 107

CET6 660

CET4 654

Hobbies

2006–2008 **Member of Classical Guitar Orchestra**, *Fudan Guitar Club*.

2006–2007 **Leader of the Secretary Department**, *Fudan Guitar Club*.