

Yang Cao

CURRICULUM VITAE – March 2018

IF5.37, Informatics Forum
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB, UK

Tel: +44 (0)754 241 5501
Email: yang.cao@ed.ac.uk
Web: <http://homepages.inf.ed.ac.uk/ycao>

Research Interests	Database systems and theory: query processing, approximation, data quality Web data management: graph query languages, graph querying methods, parallelization
Education	<div><div>University of EdinburghEdinburgh, UK <i>Ph.D.</i>: Database, Computer Science and InformaticsFebruary 2013 – August 2016 Supervisor: Prof. Wenfei Fan(awarded on 29 Nov, 2016)</div><div>Beihang UniversityBeijing, China <i>B.S.</i>: Computer Science and TechnologySeptember 2006 – June 2010 Graduated from the Shen-Yuan Honor School.</div></div>
Employment Record	<div><div>University of EdinburghEdinburgh, UK <i>Research Associate</i>, LFCS, School of InformaticsSeptember 2016 – present</div><div>International Research Center on Big Data at BeihangBeijing, China <i>Research Assistant</i> (working remotely at Edinburgh, UK)February 2014 – April 2016</div></div>
Research Projects	<p>I have been working on three projects described below.</p> <p>(I) BEAS: Making Big Data Small</p> <p>We develop BEAS, a new query evaluation paradigm to answer SQL queries under constrained resources, by reducing queries on big data to computation on small data. Underlying BEAS are two principled approaches:</p> <ul style="list-style-type: none">• <i>bounded evaluation</i> that computes exact answers by accessing a bounded amount of data when possible [1, 3, 7, 8, 11, 12, 14], and• <i>data-driven approximation scheme</i> that answers queries for which exact answers are beyond reach under bounded resources, and offers a deterministic accuracy bound [2]. <p>[<i>Industrial evaluation.</i>] One of our industry collaborators (Huawei Technologies Co., Ltd.) has deployed and tested a prototype system of BEAS [3] using their real-life call-detailed-record (CDR) queries, and found that the performance of 90% of their CDR queries can be improved by 25 times to 5 orders of magnitude for exact answering with bounded evaluation, and data-driven approximation enables flexible trade-offs between query accuracy and evaluation time when approximate answers are allowed.</p> <p>[<i>Publication.</i>] As my main Ph.D. thesis work, the project has produced 2 SIGMOD (one system demo), 2 PODS, 2 VLDB, 1 TODS, 1 ICDE papers and 3 filed US patents.</p> <p>(II) Methods for Querying Big Graph Data</p> <p>I have also worked on methods for querying big graph data, including</p> <ul style="list-style-type: none">• scale independent graph pattern matching by making pattern queries bounded [12];

- parallelizing sequential graph algorithms via partial evaluation and incremental computation, without thinking like a vertex [4, 20, 21] (my contribution includes the characterization and correctness proofs of the auto-parallelization framework);
- trading off structural preservability and query complexity for querying graphs [15, 19];
- approximate graph querying using views [10]; and
- graph querying made easy by query relaxation and explanations [6].

[*Publication.*] This line of research has produced 1 SIGMOD (Best paper award), 1 VLDB, 1 ICDE, 2 CIKM, 1 WWW, 1 TODS, 1 BICOD and 1 Computer Journal (invited). Moreover, it has one invited TODS submission.

(III) Data quality: Data Accuracy and Information Completeness

I have worked also on two novel data quality problems and contribute to 1 SIGMOD and 1 Information Systems papers.

(1) *Data accuracy* belongs to the problem of entity resolution. Given a set I_e of tuples pertaining to an entity e , it aims to find the most accurate values for e from I_e (a target tuple t_e for e from I_e), such that for each attribute A of e , $t_e[A]$ is closest to the true A -value of e [17].

(2) *Relative information completeness* studies the following problem: for a given query Q , can its complete answer be found from an incomplete database D ? That is, the answer to Q in D remains unchanged no matter how D is extended by adding new tuples [16].

Awards & Honors

- | | |
|--|------|
| • Selected for ACM SIGMOD Research Highlight Award | 2017 |
| • ACM SIGMOD Best Paper Award | 2017 |
| • Invited to publish in “Best of SIGMOD 2017” (TODS) | 2017 |
| • Invited to publish in “Best of PODS 2016” (TODS) | 2016 |
| • Invited to publish in “Best of BICOD 2015” (The Computer Journal) | 2015 |
| • Facebook Graduate Fellowship, finalist (<i>34 in total all over the world</i>) | 2014 |
| • Microsoft Research Asia PhD Fellowship (<i>10 in Asia and part of US</i>) | 2012 |
| • International Mathematical Contest in Modeling, FIRST Prize (International) | 2009 |
| • China Mathematical Contest in Modeling, (the ONLY) National NO.1 | 2008 |
| • “CASC Award” first prize, by China Aerospace Science and Technology | 2013 |
| • China National Scholarship for Graduates | 2012 |
| • Microsoft Research Asia Young Scholarship (<i>30 in total within China</i>) | 2009 |

Publications & Patents

Published conference & journal papers

1. **Yang Cao**, Wenfei Fan, Floris Geerts, and Ping Lu “Bounded Query Rewriting Using Views”. *ACM Transaction on Database Systems (TODS)* (**invited**), 2018.
2. **Yang Cao** and Wenfei Fan. “Data Driven Approximation with Bounded Resources”. *International Conference on Very Large Data Bases (VLDB)*, 2017.

3. **Yang Cao**, Wenfei Fan, Yanghao Wang, Tengfei Yuan, Yanchao Li and Laura Yu Chen. “BEAS: Bounded Evaluation of SQL Queries”. *ACM SIGMOD Conference on Management of Data (SIGMOD)* (demo), 2017.
4. Wenfei Fan, Yinghui Wu, Jingbo Xu, Wenyuan Yu, Jiaxin Jiang, Zeyu Zheng, Bohan Zhang, **Yang Cao** and Chao Tian. “Parallelizing Sequential Graph Computations”. *ACM SIGMOD Conference on Management of Data (SIGMOD)* (**Best paper award**), 2017.
5. **Yang Cao**, W. Fan, and T. Yuan. “Is Big Data Analytics Beyond the Reach of Small Companies?”. *Data Analysis and Knowledge Discovery* (**invited**), 1(9), 2017
6. Jia Li, **Yang Cao**, Shuai Ma, “Relaxing Graph Pattern Matching With Explanations”. *ACM International Conference on Information and Knowledge Management (CIKM)*, 2017.
7. **Yang Cao** and Wenfei Fan “An Effective Syntax for Bounded Relational Queries”. *ACM SIGMOD Conference on Management of Data (SIGMOD)*, 2016
8. **Yang Cao**, Wenfei Fan, Floris Geerts, and Ping Lu “Bounded Query Rewriting Using Views”. *ACM Symposium on Principles of Database Systems (PODS)*, 2016
9. **Yang Cao**, Wenfei Fan and Shuai Ma “Virtual Network Mapping: A Graph Pattern Matching Approach”. *The Computer Journal* (**invited**), 2016
10. Jia Li, **Yang Cao** and Xudong Liu “Approximating Graph Pattern Queries Using Views”. *ACM International Conference on Information and Knowledge Management (CIKM)*, 2016
11. Wenfei Fan, Floris Geerts, **Yang Cao**, Ting Deng and Ping Lu. “Querying Big Data by Accessing Small Data”. *ACM Symposium on Principles of Database Systems (PODS)*, 2015
12. **Yang Cao**, Wenfei Fan, Jinpeng Huai, Ruizhe Huang “Making Pattern Queries Bounded in Big Graphs”. *International Conference on Data Engineering (ICDE)*, 2015
13. **Yang Cao**, Wenfei Fan and Shuai Ma “Virtual Network Mapping: A Graph Pattern Matching Approach”. *British International Conference on Databases (BICOD)*, 2015
14. **Yang Cao**, Wenfei Fan, Wenyuan Yu “Bounded Conjunctive Queries”. *International Conference on Very Large Data Bases (VLDB)*, 2014
15. Shuai Ma, **Yang Cao**, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. “Strong Simulation: Capturing Topology in Graph Pattern Matching”. *ACM Transaction on Database Systems (TODS)*, 2014
16. **Yang Cao**, Ting Deng, Wenfei Fan, Floris Geerts. “On the Data Complexity of Relative Information Completeness”. *Information Systems*, 2014
17. **Yang Cao**, Ting Deng, Wenfei Fan, Floris Geerts. “Determining the Relative Accuracy of Attributes”. *ACM SIGMOD Conference on Management of Data (SIGMOD)*, 2013
18. Shuai Ma, **Yang Cao**, Jinpeng Huai, and Tianyu Wo. “Distributed Graph Pattern Matching”. *International World Wide Web Conference (WWW)*, 2012
19. Shuai Ma, **Yang Cao**, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. “Capturing Topology in Graph Pattern Matching”. *International Conference on Very Large Data Bases (VLDB)*, 2012

Submissions under review

20. Wenfei Fan, **Yang Cao**, Jingbo Xu, Wen yuan Yu, Yinghui Wu, Chao Tian, Jiaxin Jiang, and Bohan Zhang “From Think Parallel to Think Sequential”. *ACM SIGMOD Highlight* (**invited**), 2018. (Under review)
21. “Parallelizing Sequential Graph Computations”. *ACM Transaction on Database Systems* (**TODS**) (**invited**), 2018. (Under review)

Ph.D dissertation

22. “Querying Big Data with Bounded Data Access”. University of Edinburgh, 2016

U.S. patents

23. **Yang Cao**, Wenfei Fan, Jinpeng Huai. “Making Graph Pattern Queries Bounded in Big Graphs”. U.S. patent (US20170308620A1), October 2017.
24. Wenfei Fan, **Yang Cao**, Floris Geerts, Ting Deng, Ping Lu. “Querying Big Data By Accessing Small Data” U.S. patent (US20170277750A1), September 2017.
25. Wenfei Fan, **Yang Cao**, Floris Geerts, Ping Lu, Yu Chen, Demai Ni “Bounded Query Rewriting Using Views” U.S. patent (pending), 2017

Professional Activities

Program Committee Member

- International Conference on Extending Database Technology (**EDBT**), 2018

Invited Journal Reviewer

- ACM Journal of Data and Information Quality (**JDIQ**)
- The International Journal on Very Large Data Bases (The **VLDB Journal**)

External Reviewer

- International Conference on Very Large Data Bases (**VLDB**), 2016; external reviewer
- International Conference on Very Large Data Bases (**VLDB**), 2015; external reviewer

Tutorials & Talks

Talks

- “Data Driven Approximation with Bounded Resources”
VLDB Conference
Munich, Germany, August 2017
- “BEAS: Bounded Evaluation of SQL Queries”
Annual workshop of the National Basic Research Program of China (973 Program) on Fundamental theory of Big Data Computation in Cyberspace
Beijing, China, January 2017
- “An Effective Syntax for Bounded Relational Queries”
SIGMOD Conference
San Francisco, USA, June 2016
- “Data Driven Approach to Querying Big Data”
1st Microsoft Research Asia Ph.D Forum
Beijing, China, September 2015

- “Querying Big Data by Accessing Small Data”
PODS Conference
Melbourne, Victoria, Australia, June 2015
- “Theory and Algorithms for Querying Big Relations”
Beihang University
Beijing, China, May 2015
- “Making Pattern Queries Bounded in Big Graphs”
ICDE Conference
Seoul, Korea (South), April, 2015
- “Bounded Conjunctive Queries”
VLDB Conference
Hangzhou, China, September 2014
- “Bounded Conjunctive Queries”
Annual workshop of the National Basic Research Program of China (973 Program) on Fundamental theory of Big Data Computation in Cyberspace
Beijing, China, April 2014
- “Determining the Relative Accuracy of Attributes”
SIGMOD Conference
New York, USA, June 2013
- “Virtual Machine Live-Migration and Virtual Network Mapping”
Microsoft Research Asia Young Scholar Forum
Beijing, China, September 2009
- “How to Do Mathematical Modeling?”
“Higher Education Press” Cup Award Ceremony for National Mathematical Modeling
Chongqing, China, December 2008

Tutorials

- “Mathematical Modeling” (Summer School Course)
Beihang University, Beijing, China, Summer 2011/ Summer 2010/ Summer 2009

Students Mentoring

I have been mentoring and co-supervising the following students on a project basis:

- Yanghao Wang (MSc student, University of Edinburgh, supervisor: Wenfei Fan)
[*My role.*] I co-supervised on Mr. Wang’s master thesis project based on *Data-driven approximation* (see Project (I) “BEAS: Making Big Data Small”).
[*Outcome.*] Mr. Wang has been awarded an MSc by Research Degree *with Distinction*.
- Jia Li (PhD student, Beihang University, supervisor: Shuai Ma)
[*My role.*] I co-supervised on Ms. Li’s PhD thesis work on novel methods for querying big graphs (based on Project (II) “Methods for querying big data graphs”).
[*Outcome.*] 2 CIKM papers as a large part of Ms. Li’s PhD thesis.
- Tengfei Yuan (PhD student, University of Edinburgh, supervisor Wenfei Fan)
[*My role.*] I co-supervised Mr. Yuan on prototyping BEAS (see Project (I))
[*Outcome.*] BEAS@PostgreSQL (one of the BEAS prototypes); 1 SIGMOD demo paper.