# Nan Tang

*Senior Scientist, Data Analytics Group*
*Qatar Computing Research Institute, HBKU*
*Doha, Qatar*

*+974 66700540*
*+974 44542850*
*ntang@hbku.equ.qa*
*da.qcri.org/ntang/*
*DOB: 28 September 1980*

## Research Interests

My main research interests are to build big data analytics systems, especially **big data curation systems**, to bridge the gap between researchers and practitioners on the problem of discovering, preparing, integrating, and cleaning datasets; these consider to consume at least 80% time of data scientists in large organizations.

## Education

2004/07–2007/12 **Ph.D.**, Systems Engineering & Engineering Management
*The Chinese University of Hong Kong, Hong Kong*
- Thesis: Efficient XPath Query Processing in Native XML Databases
- Co-Supervisors: Jeffrey Xu Yu and Kam-Fai Wong

2001/09–2004/01 **M.Sc.**, Computer Science
*Northeastern University, China*
- Thesis: Parallel XML Databases
- Supervisor: Guoren Wang

1997/09–2001/01 **B.S.**, Computer Science, *Northeastern University, China*

## Professional Experience

2015/04–now *Senior Scientist*, Data Analytics, **Qatar Computing Research Institute**, Qatar.

2017/07–08 *Visiting Scientist*, **CSAIL, MIT**, US. Worked on the Data Civilizer project, with Michael Stonebraker, Samuel Madden, and Armando Solar-Lezama.

2011/12–2015/03 *Scientist*, Data Analytics, **Qatar Computing Research Institute**, Qatar.

2010/02-2012/01 *Research fellow*, **University of Edinburgh**, UK. Worked on data cleaning and graph algorithms, with Wenfei Fan.

2008/02-2010/01 *Scientific staff member*, **CWI** (the national research institute for mathematics and computer science), the Netherlands. Worked on column-store database MonetDB and distributed XQuery processing, with Peter Boncz.

2007/03–08 *Visiting Scholar*, **University of Waterloo**, Canada. Worked on XML indexing and query rewriting, with Tamer Özsu.

## Awards

2015 Selected for **best papers of PVLDB 2015**, for my paper, titled "Lightning Fast and Space Efficient Inequality Joins".

2012 Selected for **best papers of ICDE 2012**, for my paper, titled "Incremental Detection of Inconsistencies in Distributed Data".

2010 The 37th VLDB conference, **the Best Paper Award of VLDB 2010**, for my paper, titled "Towards Certain Fixes with Editing Rules and Master Data".

2009 Selected for **best papers of ICDE 2009**, for my paper, titled "Projective Distribution of Full-Fledged XQuery".

# Research Projects

## Data Curation

**NADEEF**
*A commodity data cleaning system*

- Nadeef (SIGMOD 2013, VLDB demo 2013, SIGMOD demo 2014) is the first commodity data cleaning system that allows the users to specify multiple types of data quality rules, as well as providing arbitrary data repairing algorithms, in a unified program interface.
- BigDansing (SIGMOD 2015) scales Nadeef out to support Big Data over general purpose data processing platforms, ranging from DBMSs to MapReduce-like frameworks, e.g., Spark.
- Inequality join (PVLDB 2016, VLDBJ 2017) is inspired by a costly operation that joins tables on inequality conditions that are widely used in detecting data errors, e.g., finding people who earn higher salary but pay lower tax.

**Trusted Data Cleaning**

Traditionally, integrity constraint based data repairing solutions target at computing a consistent database w.r.t. the given integrity constraints with the minimum cost. However, this "minimality" principle is nothing about the ground truth, hence these solutions are seldomly adopted in real applications, for which *trusted data cleaning* is often preferred.

- *Editing rules* (VLDB 2010 best paper, VLDB Journal 2012) are the first work that repairs data while ensuring the correctness, by interacting with users and by using reference tables.
- *Fixing rules* (SIGMOD 2014) are the first work to do automatic and trusted data repairing.
- Falcon (SIGMOD 2016) is a system that interacts with the user to decide SQL update queries, which actually encode fixing rules, to repair data.
- *Sherlock rules* (ICDE 2015) extend fixing rules by not only repairing data, but also annotating data as correct (proof positive) or wrong (proof negative), using reference tables.
- *Detective rules* (ICDE 2017) differ from Sherlock rules in that they draw evidence from trusted knowledge bases (in RDF format), not from reference tables (in tabular format).
- Katara (SIGMOD 2015, VLDB demo 2015) is an end-to-end data cleaning system powered by knowledge bases and crowdsourcing that interprets table semantics, identifies correct and wrong data, and generates top-$k$ possible repairs.

**Data Discovery and Integration**

With thousands of disparate data sources, data discovery and integration are more important than ever to help analysts consolidate valuable sources and solve analytical tasks.

- Data Civilizer (CIDR 2017, SIGMOD demo 2017) is a system that finds, integrates, and cleans datasets to facilitate large companies to do analytical tasks.

**Entity Resolution and Consolidation**

- Data Currency (ICDE 2013) identifies the most current values for duplicate records.
- SynthesizER (SIGMOD demo 2017, PVLDB 2017) studies the problem of synthesizing concise and interpretable entity matching rules from examples, using *program synthesis*.
- DeepER (under submission to ICDE 2018) leverages the idea of distributed representations and representation learning from *deep learning* to improve the accuracy of entity resolution.

## Data Visualization

**ML-powered Automatic Data Visualization**

The current data visualization tools (e.g., Tableau and Microsoft Excel) allow users to easily create visualizations, *only if* they know their data well, such as which attributes to use, which chart (line or bar) is appropriate. It is still hard for non-experts to produce great visualizations.

- DeepEye (under submission to ICDE 2018) targets at automating the above process by training *binary classifiers* to recognize good/bad visualizations, and a supervised *learning-to-rank* model to select top-$k$ visualizations.

## Data Streams

**Graph Stream Management**

- TCM (SIGMOD 2016) is a graphical sketch for graph streams. Traditional sketches have 1-dimensional structures (or simply counters). TCM is 2-dimensional that keeps all connections of the graph stream, which supports richer graph queries than existing sketches.

## Teaching and Mentoring Experience

### Internship Mentor, QCRI

| | | |
|---|---|---|
| Jinsong Guo | Ph.D. from University of Oxford, UK, on *Estimating data errors* | 2017/03-2017/09 |
| Dong Deng | Ph.D. from Tsinghua University, China, on *Data Civilizer* | 2016/06-2016/08 |
| Sourav Medya | Ph.D. from UC Santa Barbara, US, on *Mining data streams* | 2016/06-2016/08 |
| Qing Chen | Master from Fudan University, China, on *Graph stream summarization* | 2015/07-2016/04 |
| Jian He | Master from Tsinghua University, China, on Falcon | 2014/11-2015/02 |
| Matteo Interlandi | Ph.D. from University of Modena, Italy, on *Sherlock rules* | 2014/03-2014/05 |
| Chu Xu | Ph.D. from University of Waterloo, Canada, on Katara | 2013/05-2014/07 |
| Jiannan Wang | Ph.D. from Tsinghua University, China, on *Fixing rules* | 2012/12-2013/02 |
| Yu Tang | Master from Hong Kong University, HK, on Nadeef *dashboard* | 2012/11-2013/01 |
| Amr Ebaid | Ph.D. from Purdue University, US, on Nadeef | 2012/04-2013/01 |
| Ahmed Eldawy | Ph.D. from University of Minnesota, US, on Nadeef | 2012/01-2012/05 |
| Michele Dallachiesa | Ph.D. from University of Trento, Italy, on Nadeef | 2012/01-2012/05 |

### Teaching, University of Edinburgh, UK (Tutorials)

| | |
|---|---|
| Applied Databases | 2010/09-11 |

### Teaching, The Chinese University of Hong Kong, Hong Kong (Tutorials)

| | |
|---|---|
| Digital Logical and Systems | 2006/09-12, 2007/09-12 |
| Fundamentals of Information Systems | 2004/09-12, 2006/01-05 |
| Information Systems Design & Analysis | 2005/01-05 |

## Selected Professional Activities and Services

| | | |
|---|---|---|
| **PC Member** | ACM SIGMOD International Conference on Management of Data (SIGMOD) | 2015, 2017, 2018 |
| | International Conference on Very Large Data Bases (PVLDB) | 2015 |
| | IEEE International Conference on Data Engineering (ICDE) | 2013, 2018 |
| | International Conference on Extending Database Technology (EDBT) | 2017 |
| | SIAM International Conference on Data Mining (SDM) | 2017 |
| | ACM Conference on Information and Knowledge Management (CIKM) | 2011, 2012 |
| **Journal Reviewer** | VLDB Journal | 2009, 2010, 2011 |
| | IEEE Trans. on Knowledge and Data Engineering (TKDE) | 2007, 2011, 2012, 2016 |
| | ACM Transactions on Knowledge Discovery from Data (TKDD) | 2012 |
| | ACM Transactions on Database Systems (TODS) | 2013 |
| | ACM Transactions on the Web (TWEB) | 2012, 2015 |

## Selected Publications

### Journal Publications

1. Rohit Singh, Vamsi Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Armando Solar-Lezama, and Nan Tang. *Synthesizing Entity Matching Rules by Examples.* PVLDB, 2017.
2. Zuhair Khayyat, William Lucia, Meghna Singh, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Panos Kalnis. *Fast and Scalable Inequality Joins.* VLDB Journal, 2017 (Special issue: *Best Papers of VLDB 2015, invited*).
3. Shuang Hao, Nan Tang, Guoliang Li, Jian He, Na Ta, and Jianhua Feng. *A Novel Cost-Based Model for Data Repairing.* IEEE Transaction on Knowledge and Data Engineering (TKDE), 2017.
4. Jiannan Wang, and Nan Tang. *Dependable Data Repairing with Fixing Rules.* ACM Journal of Data and Information Quality (JDIQ), 2017.

5. Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. *Data Cleaning: Where are we and what needs to be done? [Experiments and Analyses].* PVLDB, 2016.
6. Zuhair Khayyat, William Lucia, Meghna Singh, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, Panos Kalnis. *Lightning Fast and Space Efficient Inequality Joins.* PVLDB, 2015.
7. Wenfei Fan, Floris Geerts, Nan Tang, Wenyuan Yu. *Conflict Resolution with Data Currency and Consistency.* ACM Journal of Data and Information Quality (JDIQ), 2014 (*invited*).
8. Wenfei Fan, Shuai Ma, Nan Tang, Wenyuan Yu. *Interaction between Record Matching and Data Repairing.* ACM Journal of Data and Information Quality (JDIQ), 2014.
9. Wenfei Fan, Jianzhong Li, Nan Tang, Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data.* IEEE Transaction on Knowledge and Data Engineering (TKDE), 2014 (Special issue: *Best Papers of ICDE 2012, invited*).
10. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu. *Adding Regular Expressions to Graph Reachability and Pattern Queries.* Frontiers of Computer Science, 2012 (Special issues: New Topics in Database, *invited*).
11. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data.* VLDB Journal, 2012 (Special issue: *Best Papers of VLDB 2010, invited*).
12. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data.* In PVLDB, 2010 (*The Best Paper award*).
13. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu, Yunpeng Wu. *Graph Pattern Matching: From Intractable to Polynomial Time.* In PVLDB, 2010.
14. Ying Zhang, Nan Tang, Peter Boncz. *Projective Distribution of Full-Fledged XQuery.* IEEE Transaction on Knowledge and Data Engineering (TKDE), 2010 (Special issue: *Best Papers of ICDE 2009, invited*).
15. Kam-Fai Wong, Jeffrey Xu Yu, and Nan Tang. *Answering XML queries using path-based indexes: A survey.* World Wide Web Journal, 2006.

## Conference Publications

1. Saravanan Thirumuruganathan, Laure Berti-Equille, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz and Nan Tang. *UGuide – User-Guided Discovery of FD-Detectable Errors.* SIGMOD, 2017.
2. Raul Castro Fernandez, Dong Deng, Essam Mansour, Abdulhakim Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker and Nan Tang. *A Demo of the Data Civilizer System.* SIGMOD demo, 2017.
3. Rohit Singh, Vamsi Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Armando Solar-Lezama and Nan Tang. *Generating Concise Entity Matching Rules.* SIGMOD demo, 2017.
4. Shuang Hao, Nan Tang, Guoliang Li and Jian Li. *Cleaning Relations using Knowledge Bases.* ICDE, 2017.
5. Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani and Nan Tang. *The Data Civilizer System.* CIDR, 2017.
6. Nan Tang, Qing Chen, and Prasenjit Mitra. *Graph Stream Summarization.* SIGMOD, 2016.
7. Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. *Interactive and Deterministic Data Cleaning.* SIGMOD, 2016.
8. Divy Agrawal, M. Lamine Ba, Laure Berti-Equille, Sanjay Chawla, Ahmed Elmagarmid, Hossam Hammady, Yasser Idris, Zoi Kaoudi, Zuhair Khayyat, Sebastian Kruse, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Mohammed J. Zaki. *Rheem: Enabling Multi-Platform Task Execution.* SIGMOD demo, 2016.

9. Divy Agrawal, Sanjay Chawla, Ahmed Elmagarmid, Zoi Kaoudi, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Mohammed J. Zaki. *RHEEM: Road to Freedom in Big Data Analytics*. EDBT (vision paper), 2016.
10. Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. *KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing*. SIGMOD, 2015.
11. Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. *KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing*. VLDB demo, 2015.
12. Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Sam Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Si Yin. *BigDansing: A System for Big Data Cleansing*. SIGMOD, 2015.
13. Nan Tang. *Big RDF Data Cleaning*. The 6th International Workshop on Data Engineering meets the Semantic Web (DESWeb), in conjunction with ICDE, 2015 (*invited*).
14. Matteo Interlandi and Nan Tang. *Proof Positive and Negative in Data Cleaning*. ICDE, 2015.
15. Nan Tang. *Big Data Cleaning*. APWeb, 2014 (*invited as Distinguished Lecture Series*).
16. Jiannan Wang and Nan Tang. *Towards Dependable Data Repairing with Fixing Rules*. SIGMOD, 2014.
17. Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF/ER: Generic and Interactive Entity Resolution*. SIGMOD demo, 2014.
18. Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. *NADEEF: A Commodity Data Cleaning System.* SIGMOD, 2013.
19. Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF: A Generalized Data Cleaning System.* VLDB demo, 2013.
20. Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. *Inferring Data Currency and Consistency for Conflict Resolution.* ICDE, 2013.
21. Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data.* ICDE, 2012.
22. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *CerFix: A System for Cleaning Data with Certain Fixes.* VLDB demo, 2011.
23. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Interaction between record matching and data repairing.* SIGMOD, 2011.
24. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu. *Adding Regular Expressions to Graph Reachability and Pattern Queries.* ICDE, 2011.
25. Nan Tang, Lefteris Sidirourgos, Peter Boncz. *Space-Economical Q-Gram Index for Exact String Matching.* CIKM, 2009.
26. Ying Zhang, Nan Tang, Peter Boncz. *Efficient Distribution of Full-Fledged XQuery.* ICDE, 2009.
27. Nan Tang, Jeffrey Xu Yu, Hao Tang, M. Tamer Özsu and Peter Boncz. *Materialized View Selection in XML Databases.* DASFAA, 2009.
28. Nan Tang, Jeffrey Xu Yu, M. Tamer Özsu, Byron Choi, and Kam-Fai Wong. *Multiple materialized view selection for XPath query rewriting.* ICDE, 2008.
29. Nan Tang, Jeffrey Xu Yu, M. Tamer Özsu, and Kam-Fai Wong. *Hierarchical indexing approaches to support XPath queries.* ICDE, 2008.
30. Nan Tang, Guoren Wang, Jeffrey Xu Yu, Kam-Fai Wong, and Ge Yu. *Win: An effcient data placement strategy for parallel XML databases.* In Proc. 11th International Conference on Parallel and Distributed System (ICPADS), 2005.

## Miscellaneous Writings

1. Mourad Ouzzani and Nan Tang, contributors to the chapter "The Data Civilizer System", for *Michael Stonebraker ACM A.M. Turing Award Book*, under revision (*invited*).

2. Wenfei Fan, Floris Geerts, Shuai Ma, Nan Tang, and Wenyuan Yu. *Data Quality Problems beyond Consistency and Deduplication.* In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman, Edinburgh, UK, 2013 (*invited*).
3. George Beskales, Gautam Das, Ahmed K. Elmagarmid, Ihab F. Ilyas, Felix Naumann, Mourad Ouzzani, Paolo Papotti, Jorge Quiane-Ruiz, and Nan Tang. *The Data Analytics Group at the Qatar Computing Research Institute.* SIGMOD Record, 2012.

### Patents

1. *Dependable Data Repairing with Fixing Rules.* Qatar Computing Research Institute, HBKU (PCT/EP2013/052476).
2. *Towards dependable Data Repairing with Fixing Rules.* Qatar Computing Research Institute, HBKU (PCT/EP2014/052494).
3. KATARA*: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing.* Qatar Computing Research Institute, HBKU (PCT/GB2014/051670).
4. NADEEF*: A holistic and extensible data cleaning platform.* Qatar Computing Research Institute, HBKU (PCT/EP2012/062446).
5. *Generalized Data Cleaning using SAT-Solvers.* Qatar Computing Research Institute, HBKU (PCT/EP2012/062445).

## Invited Talks

2016/10 "Data Cleaning Techniques", at Harvard University, US

2016/03 "Graph Stream Summarization", at MIT, US

2015/12 "Trusted Data Cleaning", at KAUST, Saudi Arabia

2015/04 "Big RDF Data Cleaning", at the 6th International Workshop on Data Engineering meets the Semantic Web (DESWeb), in conjunction with ICDE 2015

2014/09 "Big Data Cleaning" at Asia-Pacific Web Conference 2014, Distinguished Lecturer series