

Yanlei Diao

Office: Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003-9264

Phone: (413) 545-1135
Fax: (413) 545-1249
Email: yanlei@cs.umass.edu
URL: <http://www.cs.umass.edu/~yanlei>

Research Interests

Information architectures and data management systems.
Large-scale data analytics, data streams, probabilistic databases,
sensor data management, flash memory databases,
data dissemination, and XML query processing

Education

University of California, Berkeley Berkeley, CA
Ph.D., Electrical Engineering and Computer Science August 2005
Advisor: Professor Michael J. Franklin
Thesis: “Query Processing for Large-Scale XML Message Brokering”

Hong Kong University of Science and Technology Hong Kong, China
M.S., Computer Science June 2000
Advisor: Professor Hongjun Lu
Thesis: “Learning-based Web Query Processing”

Fudan University Shanghai, China
B.S., Computer Science July 1998

Academic Positions

Associate Professor 09/2011 – present
Department of Computer Science
University of Massachusetts Amherst

Assistant Professor 09/2005 – 08/2011
Department of Computer Science
University of Massachusetts Amherst

Research Intern 02/2003 – 08/2003
WebLogic Integration and Liquid Data Groups
BEA Systems, San Jose, CA

Honors and Awards

- **IBM Innovation Awards on Scalable Data Analytics**, 2010
- **National Science Foundation CAREER Award**, 2008
- **Finalist for Microsoft Research New Faculty Fellowship**, 2008
- **Nomination for Distinguished Teaching Award**, University of Massachusetts Amherst, Fall 2007 and Spring 2008
- **Distinguished Faculty Lecture Series**, University of Texas at Austin, December 2005
- **ACM-SIGMOD Dissertation Award Honorable Mention**, 2005
- **First Class Honor**, Fudan University, 1998
- **Intel Fellowship**, Fudan University, 1996-1998
- **GE Fellowship**, Fudan University, 1995-1996

Publications

Book Chapters

- [1] Yanlei Diao and Michael J. Franklin. “High-Performance XML Message Brokering.” *Data Stream Management: Processing High-Speed Data Streams* (Garofalakis, Gehrke, and Rastogi, eds.). Springer Data-Centric Systems and Applications Series, November 2010.
- [2] Fang Yu, Yanlei Diao, Randy Katz, and T. V. Lakshman. “Fast Packet Pattern Matching Algorithms.” *Algorithms for Next Generation Network Architecture* (Graham Cormode and Marina Thottan, eds.). Springer Computer Communications and Networks Series, October 2009.
- [3] Yanlei Diao and Michael J. Franklin. “Publish/Subscribe over Streams.” *Encyclopaedia of Database Systems*. Springer, 2009. 6 pages.
- [4] Yanlei Diao and Michael J. Franklin. “XML Publish/Subscribe.” *Encyclopaedia of Database Systems*. Springer, 2009. 6 pages.

Journal Papers

- [5] Yanming Nie, Richard Cocci, Zhao Cao, Yanlei Diao, and Prashant Shenoy. “SPIRE: Efficient Inference and Compression over RFID Streams.” *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, March 2011, 1096-1107.
- [6] Zhao Cao, Charles Sutton, Yanlei Diao, and Prashant Shenoy. “Distributed Inference and Query Processing for RFID Tracking and Monitoring.” *Volume 4 of the Journal “Proceedings of the VLDB Endowment” (PVLDB)*, 2011. 12 pages.

- [7] Thanh Tran, Andrew McGregor, Yanlei Diao, Liping Peng, and Anna Liu. “Conditioning and Aggregating Uncertain Data Streams: Going Beyond Expectations.” *Volume 3 of the Journal “Proceedings of the VLDB Endowment” (PVLDB)*, September 2010. 12 pages.
- [8] Yanlei Diao, Mehmet Altinel, Michael J. Franklin, Hao Zhang, and Peter Fischer. “Path Sharing and Predicate Evaluation for High-Performance XML Filtering.” *ACM Transactions on Database Systems (TODS)* 28(4): 467-516, December 2003. ACM Press, New York, NY.

Refereed Major Conference Papers

- [9] Liping Peng, Yanlei Diao, and Anna Liu. “Optimizing Probabilistic Query Processing on Continuous Uncertain Data”. In *Proceedings of the 37th International Conference on Very Large Data Bases (VLDB 2011)*, August 2011. 12 pages. (*Also appeared in Volume 4 of the Journal “Proceedings of the VLDB Endowment” (PVLDB), a new Database journal*)
- [10] Buduo Li, Ed Mazur, Yanlei Diao, Andrew McGregor, and Prashant Shenoy. “A Platform for Scalable On-pass Analytics using MapReduce”. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011)*, 12 pages, June 2011. (*Ranked top 1% of all submissions*)
- [11] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. “Quality-Biased Ranking of Web Documents”. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 10 pages, 2011. (*8.3% for oral presentation*)
- [12] Haopeng Zhang, Yanlei Diao, and Neil Immerman. “Recognizing Patterns in Streams with Imprecise Timestamps”. In *Proceedings of the 36th International Conference on Very Large Data Bases (VLDB 2010)*, September 2010. 12 pages. (*17.3% acceptance rate; also appeared in Volume 3 of the Journal “Proceedings of the VLDB Endowment” (PVLDB), a new Database journal*)
- [13] Exploiting the Interplay Between Memory and Flash Storage In Embedded Sensor Devices. Devesh Agrawal, Boduo Li, Zhao Cao, Deepak Ganesan, Yanlei Diao, Prashant Shenoy. In *Proceedings of the IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2010)*, August 2010. 10 pages. (*23% acceptance rate*)
- [14] Thanh Tran, Liping Peng, Boduo Li, Yanlei Diao, and Anna Liu. “PODS: Modeling and Processing of High-Volume Uncertain Data Streams.” In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2010)*, pages 159-170, June 2010. (*20.8% acceptance rate*)
- [15] Devesh Agrawal, Deepak Ganesan, Ramesh Sitaraman, Yanlei Diao, and Shashi Singh. “Lazy-Adaptive Tree: An Optimized Index Structure for Flash Devices.” In *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB 2009)*, pages 361-372, August 2009. (*16.7% acceptance rate; also appeared in Volume 2 of the Journal “Proceedings of the VLDB Endowment” (PVLDB), a new Database journal*)
- [16] Desislava Petkova, W. Bruce Croft, and Yanlei Diao. “Refining Keyword Queries for XML Retrieval by Combining Content and Structure.” In *Proceedings of 31st European Conference on Information Retrieval (ECIR 2009)*, pages 662-669, April 2009. (*31.9% acceptance rate*)

- [17] Thanh Tran, Charles Sutton, Richard Cocci, Yanming Nie, Yanlei Diao, and Prashant Shenoy. “Probabilistic Inference over RFID Streams in Mobile Environments.” In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*, pages 1096-1107, March 2009. (17% acceptance rate)
- [18] Yanlei Diao, Boduo Li, Anna Liu, Liping Peng, Charles Sutton, Thanh Tran, and Michael Zink. “Capturing Uncertainty in High-Volume Stream Processing.” In *Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR 2009)*, January 2009. 11 pages. (27.1% acceptance rate)
- [19] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. “Efficient Pattern Matching over Event Streams”. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, pages 147-160, June 2008. (18% acceptance rate)
- [20] Yanlei Diao, Deepak Ganesan, Gaurav Mathur, and Prashant Shenoy. “Re-thinking Data Management for Storage-centric Sensor Networks.” In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR 2007)*, pages 22-31, January 2007. (44% acceptance rate)
- [21] Fang Yu, Zhifeng Chen, Yanlei Diao, T.V. Lakshman, and Randy H. Katz. “Fast and Memory-Efficient Regular Expression Matching.” *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS 2006)*, pages 93-102, December 2006. (19% acceptance rate)
- [22] Eugene Wu, Yanlei Diao, and Shariq Rizvi. “High-Performance Complex Event Processing over Streams.” In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2006)*, pages 407-418, June 2006. (13% acceptance rate)
- [23] Yanlei Diao, Shariq Rizvi, and Michael J. Franklin. “Towards an Internet-Scale XML Dissemination Service.” In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004)*, pages 612-623, August 2004. (16% acceptance rate)
- [24] Yanlei Diao, Daniela Florescu, Donald Kossmann, Michael J. Carey, and Michael J. Franklin. “Implementing Memoization in a Streaming XQuery Processor.” In *Proceedings of the 2nd International XML Database Symposium (XSym 2004)*, pages 35-50, August 2004. (25% acceptance rate)
- [25] Yanlei Diao and Michael J. Franklin. “Query Processing for High-Volume XML Message Brokering.” In *Proceedings of the 29th International Conference on Very Large Data Bases (VLDB 2003)*, pages 261-272, September 2003. (14% acceptance rate)
- [26] Yanlei Diao, Hongjun Lu, Songting Chen, and Zengping Tian. “Toward Learning Based Web Query Processing.” In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB 2000)*, pages 317-328, September 2000. (15% acceptance rate)
- [27] Yanlei Diao, Hongjun Lu, and Dekai Wu. “A Comparative Study of Classification Based Personal E-mail Filtering.” In *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2000)*, pages 408-419, April 2000. (29% acceptance rate)

Other Refereed Papers (including Short Papers, Workshop Papers, and System Demonstrations in Published Proceedings)

- [28] Ed Mazur, Boduo Li, Yanlei Diao, and Prashant Shenoy. “Towards Scalable One-Pass Analytics using MapReduce.” In *Proceedings of the 1st International Workshop on Data Intensive Computing in the Clouds, May 2011*. 10 pages.
- [29] Zhao Cao, Yanlei Diao, and Prashant Shenoy. “Architectural Considerations for Distributed RFID Tracking and Monitoring.” In *Proceedings of the 5th International Workshop on Networking Meets Databases (NetDB 2009)*, October 2009. 6 pages.
- [30] Daniel Gyllstrom, Jagrati Agrawal, Yanlei Diao, and Neil Immerman. “On Supporting Kleene Closure over Event Streams.” In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, pages 1391-1393, April 2008. (31% acceptance rate)
- [31] Richard Cocci, Thanh Tran, Yanlei Diao, and Prashant Shenoy. “Efficient Data Interpretation and Compression over RFID Streams.” In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, pages 1445-1447, April 2008. (31% acceptance rate)
- [32] Richard Cocci, Yanlei Diao, and Prashant Shenoy. “SPIRE: Scalable Processing of RFID Event Streams.” In *Proceedings of the 5th RFID Academic Convocation*, April 2007. 6 pages.
- [33] Daniel Gyllstrom, Eugene Wu, Hee-Jin Chae, Yanlei Diao, Patrick Stahlberg, and Gordon Anderson. “SASE: Complex Event Processing over Streams.” System Demo. In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR 2007)*, pages 407-411, January 2007.
- [34] Yanlei Diao, Peter Fischer, Michael J. Franklin, and Raymond To. “YFilter: Efficient and Scalable Filtering of XML Documents.” System Demo. In *Proceedings of the 18th International Conference on Data Engineering (ICDE 2002)*, pages 341-342, February 2002.
- [35] Songting Chen, Yanlei Diao, Hongjun Lu and Zengping Tian. “FACT: A Learning Based Web Query Processing System.” System Demo. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 2000)*, page 587, May 2000.

Non-Refereed Invited Work

- [36] Yanlei Diao and Michael J. Franklin. “High-Performance XML Filtering: An Overview of YFilter.” *IEEE Data Engineering Bulletin* 26(1): 41-48, March, 2003.

Other Publications

- [37] Zhao Cao, Charles Sutton, Yanlei Diao, and Prashant Shenoy. Distributed Inference and Query Processing for RFID Tracking and Monitoring. Technical Report UM-CS-2010-024, University of Massachusetts Amherst, 2010. 12 pages.
- [38] Ravishankar Rajamony and Yanlei Diao. Improved Memory Management for XML Data Stream Processing. Technical Report UM-CS-2010-040, University of Massachusetts Amherst, 2010. 14 pages.
- [39] Yanming Nie, Richard Cocci, Zhao Cao, Yanlei Diao, and Prashant Shenoy. “SPIRE: Efficient Data Interpretation and Compression over RFID Streams.” Technical Report UM-CS-2009-035, University of Massachusetts Amherst, 2009. 14 pages.

- [40] Yanlei Diao, Neil Immerman, and Daniel Gyllstrom. “SASE+: An Agile Language for Kleene Closure over Event Streams”. Technical Report UM-CS-2007-03, University of Massachusetts Amherst, 2007.
- [41] Yanlei Diao. “Query Processing for Large-Scale XML Message Brokering.” PhD Dissertation, University of California, Berkeley, August 2005. **ACM-SIGMOD Dissertation Award Honorable Mention**. http://www.cs.umass.edu/~yanlei/publications/Yanlei_thesis.pdf.

Research Support

- Kevin Fu (PI), Yanlei Diao (co-PI), Charles Ross, Deepak Ganesan, Wayne Bureson. “MRI: Acquisition of an RFID Testbed Using Renewable Energy for Object Identification and Habitat Monitoring.” **National Science Foundation** CNS-0923313, \$450,010, 10/2009-09/2012.
- Yanlei Diao (PI) and Anna Liu. “III-COR-small: Capturing Data Uncertainty in High-Volume Stream Processing.” **National Science Foundation** IIS-0812347, \$441,594, 09/2008-08/2011.
 - **REU Supplement**, \$16,000, 06/2010-05/2011.
- Yanlei Diao (PI). “CAREER: Efficient, Robust RFID Stream Processing for Tracking and Monitoring.” **National Science Foundation** IIS-0746939, \$588,879, 06/2008-05/2013.
 - **REU Supplement**, \$16,000, 07/02/2009-06/30/2010.
- Deepak Ganesan (PI), Yanlei Diao (co-PI), Prashant Shenoy. “NeTS-NOSS: STONES: Storage-centric Networked Embedded Systems.” **National Science Foundation** CNS-0626873, \$500,000, 09/2006-08/2010.
- Yanlei Diao (PI). “Streaming Analytics in Large-Scale Monitoring Applications.” **IBM Scalable Data Analytics for A Smarter Planet Innovation Award**. \$20,000.
- Yanlei Diao (PI) and Andrew McGregor. “Data Analytics in the Cloud: Exact Answers Fast, Approximate Answers Faster.” **NEC Labs Research Award**. \$40,000.
- Yanlei Diao (PI) and Prashant Shenoy. “Streamlining Large-Scale Data Analysis: Exact Answers Fast, Approximate Answers Faster.” **Google Research Award**, \$50,000.
- Yanlei Diao (PI). “In-Network Event Processing over Distributed Streams.” **Cisco Research Gift**, \$89,916.

Invited Talks

- “Career Paths in Databases”
DataBase Mentoring (DB Me) Workshop in conjunction with SIGMOD 2010, June 11 2010
- “Addressing New Challenges in Data Stream Processing”
University of Wisconsin Madison, September 1, 2010
AT&T Research Center, August 3 2010
IBM Watson Research Center, August 2 2010

Microsoft Research Silicon Valley, May 11 2010
Yahoo! Research, May 10 2010
Portland State University, May 7 2010
Massachusetts Institute of Technology, April 29 2010
University of Pittsburgh and Carnegie Mellon University, April 19 2010
University of California San Diego, April 14 2010
University of California Los Angeles, November 17 2009
University of California Irvine, November 16 2009
University of California Berkeley, November 13 2009
IBM Almaden Research Center, July 28 2009
HP Labs, July 27 2009

- “An Optimized Index Structure for Flash Devices”
Non-volatile Memories Workshop, April 13 2010
- “SASE+: Expressing and Processing Complex Event Patterns over Streams”
Worcester Polytechnic Institute, Department Colloquium, April 25, 2008
New England Database Day, Cambridge, MA, February 4, 2008
Renming University and Beihuang University, Beijing, China, January 4, 2008
Microsoft Research Center, Redmond, WA, October 12, 2007
StreamBase Systems, Lexington, Massachusetts, February 16, 2007
- “SPIRE: Scalable Processing of RFID Event Streams”
International Workshop on RFID Data Management, April 7, 2008
- “High-Performance Complex Event Processing over Streams”
Hong Kong Baptist University, Hong Kong, January 16, 2007
Fudan University, China, September 7, 2006
- “Query Processing for Large-Scale XML Message Brokering”
Tsinghua University and Peking University, September 5-6, 2006
Distinguished Faculty Lecture Series, University of Texas at Austin, December 12-14, 2005
AT&T Research Lab, December 15, 2005
- “XML Filtering, Transformation, and Routing with YFilter”
Cisco Systems, January 5, 2007
Database Seminar, University of Wisconsin, Madison, December 17, 2004
IBM Almaden Research Center, September 23, 2004
Oracle Corporation, December 9, 2003
BEA Systems, May 22, 2003

Software

SASE 1.0 (<http://avid.cs.umass.edu/sase/index.php?page=code>): A stream processing system developed at the University of Massachusetts Amherst that provides fast pattern matching over event streams. The release of SASE 1.0 contains the core stream processing engine of the SASE system.

YFilter 1.0 (http://yfilter.cs.umass.edu/code_release.htm): A high-performance filtering system that allows users or applications to submit queries to be continuously executed against streaming XML messages. This release supports queries written in a subset of XPath 1.0.

Teaching

Assistant Professor

University of Massachusetts Amherst

Information Systems (CMPSCI 445), Fall 2007, S2008, F2009

Database Design and Implementation (CMPSCI 645), S2006 (with G. Miklau), S2007, S2009, S2011

Advanced Topics in Database Systems (CMPSCI 745), F2008, F2010

Hot Topics in Databases (CMPSCI 691TD), Spring 2009, S2010

Advanced Database Systems (CMPSCI 691LL), F2006

Independent Study Courses: CMPSCI 701, F2007, S2008; CMPSCI496, S2008, S2010

Graduate Student Instructor

University of California, Berkeley

Introduction to Database Systems (CS 186), Fall 2000

Teaching Assistant

Hong Kong University of Science and Technology

Database Management Systems (COMP231), Spring 2000

Principles of Database Design (COMP332), Fall 1999

Computer and Programming Fundamentals II (COMP103), Spring 1999

Computer and Programming Fundamentals I (COMP102), Fall 1998

Professional Service & Activities

- **Organizing Committee Member**

Co-Chair, New Researcher Symposium, ACM International Conference on Management of Data (SIGMOD), 2010-2011

Chair, New England Database (NEDB) Summit, 2011

Organizer, Gong Show at the Conference on Innovative Data Systems Research (CIDR), 2011

Member, Steering Committee, International Workshop on Data Management for Sensor Networks (DMSN), 2010

Co-Chair, International Workshop on Data Management for Sensor Networks (DMSN), 2008-2009

Publicity Chair, ACM SIGMOD Conference on Management of Data (SIGMOD), 2009

Co-organizer, New England Database Society (NEDS) Seminar Series, 2009-present

- **Program Committee Member**

ACM International Conference on Management of Data (SIGMOD), 2011

ACM International Conference on Information and Knowledge Management (CIKM), 2010

International Conference on Very Large Data Base (VLDB), 2007, 2008, 2009, 2010

– VLDB PhD Workshop, 2008

IEEE International Conference on Data Engineering (ICDE), 2007, 2008, 2009, 2010

Conference on Innovative Data Systems Research (CIDR), 2009, 2011

International Workshop on Networking Meets Databases (NetDB), 2008

International Workshop on RFID Data Management (RFDM), 2008

International Workshop on Data Management for Sensor Networks (DMSN), 2007-2010
International Workshop on Scalable Stream Processing (SSPS), 2007
International XML Database Symposium (XSym), 2006
International Workshop on Peer-to-Peer Information Management (P2PIM), 2006
ACM International Conference on Management of Data (SIGMOD), Demonstrations, 2005

- **Journal Reviewer**

ACM Transactions on Database Systems (TODS)
International Journal on Very Large Data Bases (VLDB Journal)
ACM Transactions on Information Systems (TOIS)
ACM Transactions on Internet Technology (TOIT)
IEEE Transactions on Knowledge & Data Engineering (TKDE)
Journal of Computer Science and Technology (JCST)

- **Panelist**

International Workshop on Data Management for Sensor Networks (DMSN), 2010
International Workshop on RFID Data Management (RFDM), 2008
National Science Foundation, 2007

- **Diversity Efforts**

MACAA Aspiration Award Ceremony, in Cambridge MA, November 2010
Dot Diva Launch Event, Cambridge MA, September 2010
DataBase Mentoring (DB Me) Workshop at SIGMOD 2010
Speaker at the Women in Science Club at UMass Boston, April 2010
Member of Organization for Technical Women in Computing (Systers)

Departmental Service

- **Member**

Undergraduate Committee, 2009-2010
Graduate Committee, 2007-2008, 2009-2010
Faculty Recruiting committee, 2008-2009, 2010-2011
Diversity Committee, 2005-2008
Personnel Committee, 2006-2007, 2010-2011
Graduate Admissions Committee, 2005-2006

- **Panelist**

Professionalism Seminar on Job Hunting, Fall 2005

- **Speaker**

Lab Description Seminar, Fall 2005, Fall 2006, Fall 2007, Fall 2009
Professionalism Seminar on Job Hunting, Fall 2005
CS Women Luncheon Meeting, Fall 2005

Advising

- **Current PhD Students**

Thanh Tran
Liping Peng
Boduo Li
Haopeng Zhao

- **Current MS Student(s)**

Abhishek Roy

- **Graduated MS Students**

Edward Mazur (2011), Google.

Thesis: *Towards Scalable One-Pass Analytics using MapReduce*. UMass Amherst.
Richard Cocci (2008), State Street Corporation, Boston.

Thesis: *Efficient Data Interpretation and Compression over RFID Streams*.
Ravishankar Rajamony (2008), Goldman Sachs.

Thesis: *Improved Memory Management in XML Data Stream Processing*.

Daniel Gyllstrom (2007), UMass Amherst.

Thesis: *On Supporting Kleene Closure over Event Streams*.

- **Graduate Student Researchers**

Hee-Jin Chae (Summer 2006)

Jagrati Agrawal (Spring 2007-Spring 2008)

Junghee Jo (Fall 2008-Spring 2009)

- **Undergraduate Student Researchers**

Andrew Huang (Spring 2008)

Jacob Mitchell (Fall 2009)

Alex Jackson (Spring 2010)

Eugene Wu (Fall 2006), University of California Berkeley

- **Dissertation Committees**

Xiaoyong Liu (2006)

Jiwoon Jeon (2006)

Mo Liu (2009-2010), Worcester Polytechnic Institute

- **Masters Thesis Committees**

Desislava Petkova (2008)

- **Visiting Students**

Yanming Nie (2007-2008), Northwestern Polytechnical University, China

Zhao Cao (2008-2010), Beijing Institute of Technology, China