

ANISH DAS SARMA

CONTACT INFORMATION

Home

383 King St., #515
San Francisco, CA 94158
Cell: (650) 704 7735

Work

Computer Science Dept.
353 Serra Mall #430
Stanford, CA 94305

Internet

anish.dassarma@gmail.com
<http://infolab.stanford.edu/~anishds>

BACKGROUND

Research Scientist, Yahoo Research

August 2009-present

Stanford University

M.S. in Computer Science

March 2006

Ph.D. in Computer Science, Advisor: **Prof. Jennifer Widom**

January 2010

Indian Institute of Technology (IIT) Bombay

B.Tech. in Computer Science and Engineering

May 2004

Cumulative Performance Index: **9.80/10.0**

FELLOWSHIPS AND HONORS

- Microsoft Graduate Fellowship, *2007-2009*.
- Stanford University School of Engineering Fellowship, *2004-05*.
- IIT-Bombay Dr. Shankar Dayal Sharma Gold Medal, *2004*.

TEACHING AND PROFESSIONAL ACTIVITIES

- Co-taught “CS 245: Database System Principles”, a “mezzanine” (undergraduate and graduate) database class at Stanford in Summer 2008.
- **Journal Reviewing:** Reviewer for ACM Transactions on Database Systems (TODS), Journal of Very Large Data Bases (VLDB), Journal of IEEE Transactions on Knowledge and Data Engineering (TKDE)
- **Program Committees:** (1) MUD 2009 (VLDB workshop on management of uncertain data), (2) MOUND 2010 (ICDE workshop on Management and Mining of Uncertain Data), (3) NTII 2010 (ICDE workshop on New Trends in Information Integration), (4) External reviewer for several major conferences.

RECENT RESEARCH AND WORK EXPERIENCE

- **Yahoo Research, Santa Clara, Research Scientist, 2009-present.**
- **Stanford University, member of Stanford InfoLab, 2004-2009.** My research at Stanford has been in the context of the *Trio* project (<http://infolab.stanford.edu/trio/>) for managing data uncertainty and lineage. My main contributions so far have been in the design and study of data models, in versioning and query processing of data with uncertainty and lineage, and in the integration of uncertain data. I also worked on a number of smaller projects related to Trio, such as the quality estimation in RFID streams, and indexing and statistics for uncertain data.
- **Google Inc., Mountain View, Intern, Summer 2007.** In my internship at Google, we built a completely self-configuring data integration system. We developed a formal framework for handling uncertainty in schema mappings and mediated schemas, which laid the theoretical foundations for our system. The system was able to integrate 50-800 data sources with no human intervention and produce high-quality answers. In continued collaboration with Google researchers, I worked on algorithms and complexity results for query answering over a set of dependent data sources.

- **Microsoft Research (MSR), Redmond, Intern, Summers 2005 and 2006.** I interned twice in the Data Management, Exploration, and Mining (DMX) group at MSR, Redmond. In my first internship, I worked on automatic logical design tuning, and we developed new logical constructs and algorithms for their design based on query workloads. In my second internship, I worked on deduplication. We formally defined and addressed the problem of deduplicating a set of records based on real-world constraints and objectives.
- **IBM India Research Lab Bangalore, Intern, Winter 2006.** In my short (1-month) internship at IBM's research lab in Bangalore, I worked on automatically estimating the complexity of "service requests" based on the semi-structured data that describes them.
- **Google Bangalore R&D Center, Intern, Winter 2005.** In my internship at Google's R&D center and the following months at Stanford, we devised an efficient algorithm for determining near-duplicate web pages, which operate successfully at web scale.
- **Yahoo! Bangalore, Intern, Summer 2004.** I interned at Yahoo's HotJobs team in Bangalore, and work on the problem of automatically extracting structure (such as name, address, previous employment, etc.) from plain-text resumes.

PUBLICATIONS

MAJOR REFEREED CONFERENCE PAPERS

1. Anish Das Sarma, Aditya Parameswaran, Hector Garcia-Molina, and Jennifer Widom. *Synthesizing View Definitions from Data*, To appear in Proceedings of the International Conference on Database Theory (ICDT), Lausanne, Switzerland, March 2010.
2. Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, Rina Panigrahy, *Ranking Mechanisms in Twitter-Like Forums*, To appear in Proceedings of the International Conference on Web Search and Data Mining (WSDM), New York, USA, February 2010.
3. Anish Das Sarma, Luna Dong, Alon Halevy, *Bootstrapping Pay-As-You-Go Data Integration Systems*, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Vancouver, Canada, June 2008.
4. Anish Das Sarma, Martin Theobald, Jennifer Widom, *Exploiting Lineage for Confidence Computation in Uncertain and Probabilistic Databases*, Proceedings of the 24th International Conference on Data Engineering (ICDE), Cancun, Mexico, April 2008.
5. Surajit Chaudhuri, Anish Das Sarma, Venkatesh Ganti, Raghav Kaushik, *Leveraging Aggregate Constraints for Deduplication*, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Beijing, China, June 2007.
6. Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, *Detecting Near-Duplicates for Web Crawling*, Proceedings of the 16th International World Wide Web (WWW) Conference, Banff, Canada, May 2007.
7. Omar Benjelloun, Anish Das Sarma, Alon Halevy, Jennifer Widom, *ULDBs: Databases with Uncertainty and Lineage*, Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), pp.953-964, Seoul, Korea, September 2006.
8. Anish Das Sarma, Omar Benjelloun, Alon Halevy, Jennifer Widom, *Working Models for Uncertain Data*, Proceedings of the 22nd International Conference on Data Engineering (ICDE), Atlanta, Georgia, April 2006.

JOURNAL ARTICLES AND BOOK CHAPTERS

1. Anish Das Sarma, Omar Benjelloun, Alon Halevy, Shubha Nabar, Martin Theobald, Jennifer Widom, *Representing Uncertain Data: Models, Properties, and Algorithms*, In VLDB Journal, 18(5), 989-1019, October 2009. (Special issue on uncertain and probabilistic databases.)
2. Anish Das Sarma, Luna Dong, Alon Halevy, *Data modeling in Dataspace Support Platforms*, In Conceptual Modeling: Foundations and Applications, Essays in Honor of John Mylopoulos, Springer Festschrift, LNCS 5600, 2009.
3. Anish Das Sarma, Luna Dong, Alon Halevy, *Uncertainty In Data Integration*, In C. Aggarwal, editor, Managing and Mining Uncertain Data, Springer, 2009.
4. Omar Benjelloun, Anish Das Sarma, Alon Halevy, Martin Theobald, Jennifer Widom, *Databases with Uncertainty and Lineage*, VLDB Journal, 17(2), 243-264, March 2008. (Special issue on Best papers of VLDB '06.)
5. Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Jennifer Widom, *An Introduction to ULDBs and the Trio System*, IEEE Data Engineering Bulletin, Special Issue in Probabilistic Databases, 29(1):5-16, March 2006.

WORKSHOP, DEMONSTRATION, AND “VISIONARY” PAPERS

1. Daisy Zhe Wang, Luna Dong, Anish Das Sarma, Alon Halevy, Michael J. Franklin, *Functional Dependency Generation and Applications in Pay-As-You-Go Data Integration Systems*, In Proceedings of WebDB, Rhode Island, June 2009.
2. Anish Das Sarma, Jeffrey Ullman, Jennifer Widom, *Schema Design for Uncertain Databases*, Proceedings of the Alberto Mendelzon Workshop on Foundations of Data Management, Peru, May 2009.
3. Laure Berti-Equille, Anish Das Sarma, Xin Luna Dong, Amelie Marian, Divesh Srivastava, *Sailing the Information Ocean with Awareness of Currents: Discovery and Application of Source Dependence*, Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR), Pacific Grove, California, January 2009.
4. Anish Das Sarma, Parag Agrawal, Shubha Nabar, Jennifer Widom, *Towards Special-Purpose Indexes and Statistics for Uncertain Data*, Proceedings of the Workshop on Management of Uncertain Data (MUD), Auckland, New Zealand, August 2008.
5. Michi Mutsuzaki, Martin Theobald, Ander de Keijzer, Jennifer Widom, Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Raghotham Murthy, Tomoe Sugihara, *Trio-One: Layering Uncertainty and Lineage on a Conventional DBMS*, Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR), Pacific Grove, California, January 2007. (Demonstration description)
6. Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha Nabar, Tomoe Sugihara, Jennifer Widom, *Trio: A System for Data, Uncertainty, and Lineage*, *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pp.1151-1154, Seoul, Korea, September 2006. (Demonstration description)
7. Anish Das Sarma, Shawn R. Jeffery, Michael J. Franklin, Jennifer Widom, *Estimating Data Stream Quality for Object-Detection Applications*, Proceedings of the 3rd International ACM SIGMOD Workshop on Information Quality in Information Systems, Chicago, Illinois, June 2006.
8. Shantanu Biswas, Y. Narahari, Anish Das Sarma, *A Decomposition Based Approach for Design of Supply Aggregation and Demand Aggregation Exchanges*, International Workshop on Theory Building and Formal Methods in Electronic/Mobile Commerce (TheFormEMC), Madrid, Spain, September 2004. Published in LNCS, 3236:58-71, October 2004.

TECHNICAL REPORTS

1. Anish Das Sarma. *Managing Uncertain Data*, Ph.D. Thesis, November 2009.
2. Anish Das Sarma, Luna Dong, Alon Halevy, *Data Integration with Dependent Sources*, Technical Report, December 2008.
3. Parag Agrawal, Anish Das Sarma, Jeffrey Ullman, Jennifer Widom, *LAV Integration of Uncertain Data*, Technical Report, Stanford University, August 2008.
4. Anish Das Sarma, Martin Theobald, Jennifer Widom, *Data Modifications and Versioning in Trio*, Technical Report, Stanford University, March 2008.
5. Anish Das Sarma, Shubha U. Nabar, Jennifer Widom, *Representing Uncertainty: Uniqueness, Equivalence, Minimization and Approximation*, Technical Report, Stanford University, December 2005.

TALKS

- *Schema Design for Uncertain Databases*. Given at the Alberto Mendelzon Workshop (AMW), Arequipa, Peru, May 2009.
- *Managing Uncertain Data*. Given at Stanford University in Jan. 2009, UC Irvine and Yahoo Research in Feb. 2009, UF Gainesville, AT&T Research, and MSR-Redmond in Mar. 2009, CMU in Apr. 2009, Google, UC Berkeley, and HP Labs in May 2009.
- *Trio: A System for Data, Uncertainty, and Lineage* (Plenary talk). Given at the Dagstuhl Seminar on Uncertainty Management in Information Systems, Germany, October 2008.
- *Towards Special-Purpose Indexes and Statistics for Uncertain Data*. Given at MUD workshop on Management of Uncertain Data, co-located with VLDB, Auckland, New Zealand, August 2008.
- *Bootstrapping Pay-As-You-Go Data Integration Systems*. Given at SIGMOD, Vancouver, Canada, and Stanford, June 2008.
- *The Role of Uncertainty in Data Integration*. Guest Lecture, given in the Stanford course “CS345C: Data Integration”, May 2008.
- *Robust Stratified Sampling Plans for Low Selectivity Queries*. Given at ICDE, Cancun, Mexico, and Stanford, April 2008. Presented the paper for the authors (Shantanu Joshi and Christopher Jermaine).
- *Exploiting Lineage for Confidence Computation in Uncertain and Probabilistic Databases*. Given at ICDE, Cancun, Mexico, and Stanford, April 2008.
- *The Trio System for Data, Uncertainty, and Lineage: Overview and Demo*. Given at InfoLab/Hitachi Workshop, Stanford, March 2008.
- *Leveraging Aggregate Constraints for Deduplication*. Given at SIGMOD, Beijing, China, and Stanford in June 2007.
- *Detecting Near-Duplicates for Web Crawling*. Given at WWW, Banff, Canada & Stanford in May 2007.
- *Service Request Complexity Estimation*. Given at IBM IRL Bangalore in January 2007.
- *Trio: A System for Integrated Management of Data, Uncertainty, and Lineage*. Given at IBM IRL Bangalore in December 2006.
- *ULDBs: Databases with Uncertainty and Lineage*. Given at MSR Redmond and University of Washington in August 2006, and at VLDB, Seoul, Korea in September 2006.
- *Working Models for Uncertain Data*. Given at ICDE, Atlanta, Georgia in April 2006.

SELECTED SPORTS AND CULTURAL ACHIEVEMENTS

- Internationally rated chess player (FIDE: 2071). Represented Stanford in the 2004 Pan-American inter-collegiate chess championship, and secured 4th position.
- Awarded the 2004 IIT-Bombay Institute Sports Citation, the highest institute sports award, for contributions in chess and table-tennis.
- Indian percussion instrument, *Tabla* player, *senior level*.

REFERENCES

Available upon request.