Anish Das Sarma

Basic Information

Home
5600 Stevens Creek Blvd., #205
Cupertino, CA 95014
Cell: (650) 704 7735

Internet
anish.dassarma@gmail.com
http://i.stanford.edu/~anishds

Nationality
Indian citizen
U.S. Permanent Resident

Background

Co-Founder and CTO, ClearList Inc. *May 2013-Present*

Senior Research Scientist, Google Research *May 2011-May 2013*

Research Scientist, Yahoo! Research *August 2009-May 2011*

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*
Cumalative Performance Index: **9.80/10.0**

Honors

- Three best-of-conference paper awards (VLDB 2006, SIGMOD 2008, SIGMOD 2012)
- Winner of Yahoo! Hackday Search Innovations Award, 2010
- Microsoft Graduate Fellowship, *2007-2009*.
- Stanford University School of Engineering Fellowship, *2004-05*.
- IIT-Bombay Dr. Shankar Dayal Sharma Gold Medal, *2004*.

Professional Activities

- **Thesis Committees:** Robert Ikeda (Stanford University).
- **Associate Editor:** ACM SIGMOD Record.
- **Chair:** Co-chair ACM SIGSPATIAL GIS 2012 Workshop on SoLoMo Data in the Cloud (SDC).
- **NSF Panel:** Served on NSF panel for faculty grant reviewing.
- **Program Committees:** ICDE 2014 (industrial), WSDM 2014 IEEE Big Data 2013, WWW 2013, GIS 2013 (posters co-chair), SSDBM 2013, SIGMOD 2013 (research/industrial), CIDR 2013, VLDB 2013, SIGSPATIAL GIS 2012, CIKM 2012 (research+poster), SUM 2012, WebDB 2012, IIWeb 2012, SIGMOD 2012, WSDM 2012, VLDB 2012, WWW 2012, ICDE 2012, VLDS 2012, WebDB 2011, CIKM 2011, MUD 2011,VLDB 2011 (industrial track), WWW 2011, VLDB 2011, ICDE 2011, ICDE 2011 Demo, EDBT 2011, SUM 2011, CIKM 2010, SUM 2010, WebDB 2010, MUD 2010 and 2009, MOUND 2010, NTII 2010.
- **Journal Reviewing:** Reviewer for WWW Journal, ACM Transactions on Database Systems (TODS), Journal of Very Large Data Bases (VLDB), Journal of the ACM (JACM), Journal of IEEE Transactions on Knowledge and Data Engineering (TKDE), Information Systems (IS), Journal of Data and Information Quality (JDIQ), IEEE Transactions on Parallel and Distributed Systems (TPDS)

- **Panelist:** New Researcher Symposium, Sigmod 2012.
- **Teaching:** Co-taught "CS 245: Database System Principles", a "mezzanine" (undergraduate and graduate) database class at Stanford in Summer 2008.

Recent Research and Work Experience

- **ClearList Inc., Silicon Valley, Co-Founder and CTO**, *2013-Present.* Stealth-mode startup, leading the development of exciting technology involving information extraction, data integration, and machine learning.
- **Google Research, Mountain View, Sr. Research Scientist**, *2011-2013.* I worked on a wide-range of problems at Google including efficient map visualizations of geographic data (Fusion Tables), schema and entity resolution to generate *triples* (Google's Knowledge Graph), as well as extracting structured content from tables on the Web (Webtables).
- **Yahoo Research, Santa Clara, Research Scientist**, *2009-2011.* The main themes of my research at Yahoo constitute (1) building debugging tools (based on provenance) for information extraction systems, which often produce unexpected or incorrect results due to missing or erroneous data and noisy extraction operators; (2) identifying relationships between entities, such as celebrities that are linked due to a past event, or finding similar restaurants; and (3) identifying and dealing with challenges in large-scale deduplication of Web data. I also work on a variety of other areas such as crowd-sourcing, recommendation systems, search in general, and scheduling problems in distributed networks.
- **Stanford University, member of Stanford InfoLab**, *2004-2009.* My research at Stanford was in the context of the *Trio* project (http://infolab.stanford.edu/trio/) for managing data uncertainty and lineage. My main contributions were 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.

  In an earlier internship at Google Bangalore's R&D center (Winter 2006) and the subsequent months at Stanford, we devised an efficient algorithm for determining near-duplicate web pages, which operated successfully at web scale.
- **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.

Publications

REFEREED CONFERENCE PAPERS

1. Anish Das Sarma, Hongrae Lee, Hector Gonzalez, Jayant Madhavan, Alon Halevy, *Efficient Spatial Sampling of Large Geographical Tables*, In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Arizona, USA, May 2012.

2. Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, Cong Yu., *Finding Related Tables*, In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Arizona, USA, May 2012.

3. Lujun Fang, Anish Das Sarma, Cong Yu, Philip Bohannon, *REX: Explaining Relationships between Entity Pairs*, Very Large Databases (VLDB), September 2012.

4. Anish Das Sarma, Sreenivas Gollapudi, Rina Panigrahy and Li Zhang, *Understanding Cyclic Trends in Social Choices*, In Proceedings of the conference on Web Search and Data Mining (WSDM), Seattle, USA, 2012.

5. Foto Afrati, Anish Das Sarma, David Menestrina, Aditya Parameswaran, Jeffrey Ullman, *Fuzzy Joins Using MapReduce*, In Proceedings of the conference on International Conference on Data Engineering (ICDE), Washington, USA, April 2012.

6. Xiaodan Wang, Anish Das Sarma, Christopher Olston, Randal Burns, *CoScan: Cooperative Scan Sharing in the Cloud*, In Proceedings of the Synmposium on Cloud Computing (SoCC), Portugal, 2011.

7. Aditya Parameswaran, Anish Das Sarma, Hector Garcia-Molina, Alkis Polyzotis, Jennifer Widom, *Human-Assisted Graph Search: It's Okay to Ask Questions*, Very Large Databases (VLDB), 2011.

8. Anish Das Sarma, Luna Dong, Alon Halevy, *Data Integration with Dependent Sources*, International conference on extending database technology (EDBT), 2011.

9. Anish Das Sarma, Alpa Jain, Cong Yu, *Dynamic Relationship and Event Discovery*, Web-Search and Data Mining Conference (WSDM), 2011.

10. Christopher Olston, Anish Das Sarma, *Ibis: A Provenance Manager for Multi-Layer Systems*, Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR), Pacific Grove, California, January 2011.

11. Parag Agrawal, Anish Das Sarma, Jeffrey Ullman, Jennifer Widom, *LAV Integration of Uncertain Data*, In Proceedings of the 36th International Conference on Very Large Data Bases (VLDB), Singapore, September 2010.

12. Anish Das Sarma, Martin Theobald, Jennifer Widom, *Data Modifications and Versioning in Trio*, In Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany, June 2010.

13. Anish Das Sarma, Alpa Jain, Divesh Srivastava. *I4E: Interactive Investigation of Iterative Information Extraction*, In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Indianapolis, Indiana, June 2010.

14. 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.

15. Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, Rina Panigrahy, *Ranking Mechanisms in Twitter-Like Forums*, Proceedings of the International Conference on Web Search and Data Mining (WSDM), New York, USA, February 2010.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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

23. Anish Das Sarma, Luna Dong, Alon Halevy, *Uncertainty in Data Integration and Dataspace Support Platforms*, Book chapter, In Schema Matching and Mapping, ISBN: 978-3-642-16517-7, 2010.

24. 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.)

25. 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.

26. Anish Das Sarma, Luna Dong, Alon Halevy, *Uncertainty In Data Integration*, In C. Aggarwal, editor, Managing and Mining Uncertain Data, Springer, 2009.

27. 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.)

28. 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, POSTERS, AND DEMONSTRATION PAPERS

29. Foto N. Afrati, Anish Das Sarma, Semih Salihoglu, Jeffrey D. Ullman, *Vision Paper: Towards an Understanding of the Limits of Map-Reduce Computation*, In Cloud Futures, Berkeley, California, USA, May 2012.

30. Anish Das Sarma, Alpa Jain, Philip Bohannon, *Building a Generic Debugger for Information Extraction Pipelines*, CIKM 2011 poster paper.

31. 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 the Web and Databases workshop (WebDB), Rhode Island, June 2009.

32. Anish Das Sarma, Jeffrey Ullman, Jennifer Widom, *Schema Design for Uncertain Databases*, Proceedings of the Alberto Mendelzon Workshop on Foundations of Data Management (AMW), Peru, May 2009.

33. 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.

34. 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)

35. 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)

36. 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.

37. 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

38. Foto N. Afrati, Anish Das Sarma, Semih Salihoglu, Jeffrey D. Ullman, *Upper and Lower Bounds on the Cost of a Map-Reduce Computation*, Technical Report, June 2012.

39. Vibhor Rastogi, Ashwin Machanavajjhala, Laukik Chitnis, Anish Das Sarma, *Finding Connected Components on Map-reduce in Logarithmic Rounds* , Technical Report, 2012.

40. Anish Das Sarma, Ankur Jain, Ashwin Machanavajjhala, *CBLOCK: An Automatic Blocking Mechanism for Large-Scale De-duplication Tasks*, Technical Report, 2010.

41. Ioannis Antonellis, Anish Das Sarma, and Shaddin Dughmi. *Succinct Coverage Oracles*, Technical Report, December 2009.

42. Anish Das Sarma, *Managing Uncertain Data*, Ph.D. Thesis, November 2009.

43. Anish Das Sarma, Shubha U. Nabar, Jennifer Widom, *Representing Uncertainty: Uniqueness, Equivalence, Minimization and Approximation*, Technical Report, Stanford University, December 2005.

Patents (Approved/Filed)

1. Anish Das Sarma, Hongrae Lee, Hector Gonzalez, Jayant Madhavan, Alon Halevy, *Efficient Spatial Sampling of Large Geographical Tables.*

2. Anish Das Sarma, Lujun Fang, Alon Halevy, Hongrae Lee, Fei Wu, Cong Yu, Reynold Xin, *Finding Related Web Tables.*

3. Anish Das Sarma, Alpa Jain, Cong Yu, *Discovering dynamic relations between entities.*

4. Lujun Fang, Anish Das Sarma, Cong Yu, Philip Bohannon, *Explaining and Displaying Relationships between Entity Pairs.*

5. Christopher Olston, Anish Das Sarma, *Ibis: A Provenance Manager for Multi-Layer Systems.*

6. Anish Das Sarma, Alpa Jain, Philip Bohannon, *PROBER: Ad-Hoc Debugging of Extraction and Integration Pipelines.*

7. Anish Das Sarma, Alpa Jain, Philip Bohannon, *SoS: Socializing over Search.*

8. Anish Das Sarma, Alpa Jain, Divesh Srivastava. *I4E: Interactive Investigation of Iterative Information Extraction.*

9. Anish Das Sarma, Ankur Jain, Ashwin Machanavajjhala, *CBLOCK: An Automatic Blocking Mechanism for Large-Scale De-duplication Tasks.*

10. Surajit Chaudhuri, Anish Das Sarma, Venkatesh Ganti, Shriraghav Kaushik, *Leveraging constraints for deduplication.*

11. Sanjay Agrawal, Anish Das Sarma, Vivek Narasayya, *Automated logical database design tuning.*

Invited Talks

- *How to be a good advisor/advisee?* Given at SIGMOD New Researcher Symposium, Arizona, USA, May 2012 (Invited Panelist).
- *Why is this data here?.* Given at the U. of California Santa Cruz, CA, USA, February 2010.
- *Uncertainty in Information Extraction and Integration.* Given at Yahoo! Labs, Bangalore, India, December, 2009.
- *Managing Uncertain Data.* Given at Stanford University InfoLab Seminar, Jan. 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.
- *The Role of Uncertainty in Data Integration.* Guest Lecture, given in the Stanford course "CS345C: Data Integration", May 2008.
- *The Trio System for Data, Uncertainty, and Lineage: Overview and Demo.* Given at InfoLab/Hitachi Workshop, Stanford, March 2008.
- *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.

Selected Sports and Cultural Achievements

- Internationally rated chess player (FIDE: 2071). Represented Stanford in the 2004 Pan-American intercollegiate 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.