Andrew McGregor

CONTACT
INFORMATION

Department of Computer Science
University of Massachusetts
140 Governors Drive, Amherst, MA, 01003.

Phone: (415) 205-4076
E-mail: mcgregor@cs.umass.edu
Web: www.cs.umass.edu/˜mcgregor

RESEARCH
INTERESTS

Algorithms for processing massive data sets and data streams; computing with noisy or incomplete data; clustering; communication complexity; coding and information theory. My work appears in theoretical computer science, database, data mining, coding theory, and machine learning publications.

EDUCATION

University of Pennsylvania, Philadelphia, USA   **2001 – 2006**

Ph.D., "Processing Data Streams", 2007
M.S.E., Computer Science, 2002

University of Cambridge, Cambridge, UK   **1997 – 2001**

Certificate of Advanced Study in Mathematics (Distinction), 2001
B.A., Mathematics (1st Class Honours), 2000

EMPLOYMENT

University of Massachusetts, Amherst, USA   **January 2009 – Present**

Assistant Professor in the Department of Computer Science.

Microsoft Research, Mountain View, USA   **June 2008 – January 2009**

Postdoctoral researcher in MSR-SVC.

University of California, San Diego, USA   **October 2006 – May 2008**

Postdoctoral researcher in the Information Theory and Applications Center.

Bell Labs, Lucent Technologies, Murray Hill, USA   **Summer 2002, 2004, 2005**

Three internships in the Fundamental Mathematics Group.

DIMACS, New Brunswick, USA   **Summer 2003**

Long term visitor hosted by A. Barg.

University of Glasgow, Glasgow, UK   **Summer 1997, 1999**

Internships in Dept. of Computer Science (1999) and Mechanical Engineering (1997).

Memorial University, St. John's, Canada   **Summer 1998**

Summer internship in the Multimedia and Communications Laboratory.

GRANTS AND
FUNDING

Data Analytics in the Cloud: Exact Answers Fast, Approximate Answers Faster. NEC Research Gift, $40,000. Yanlei Diao (PI) and Andrew McGregor (co-PI).

CAREER: New Directions for Sketching and Stream Computation. National Science Foundation CCF-0953754, $515,634, 04/2010-03/2015. (Single PI) REU Supplement $8,000.

BOOKS AND
THESES

[B1] A. McGregor and S. Muthukrishnan. Data Stream Algorithms. Textbook in Preparation.

[B2]   A. McGregor. Processing Data Streams. Ph.D. Thesis, University of Pennsylvania, 2007. Advisor: Sampath Kannan.

[J1]   T. Tran, L. Peng, Y. Diao, A. McGregor, and A. Liu. CLARO: Modeling and Processing Uncertain Data Streams. *International Journal on Very Large Databases*, 2012.

[J2]   A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. *ACM Transactions in Algorithms* 6(3):1–21, 2010.

[J3]   A. McGregor and O. Milenkovic. On the hardness of approximating stopping and trapping sets in LDPC codes. *IEEE Trans. Inform. Theory*, 56(4):1640–1650, 2010.

[J4]   S. Guha, A. McGregor, and S. Venkatasubramanian. Sub-linear estimation of entropy and information distances. *ACM Transactions in Algorithms*, 5(4):1–16, 2009.

[J5]   S. Guha and A. McGregor. Stream order and order statistics: Quantile estimation in random-order streams. *SIAM Journal of Computing*, 38(5):2044–2059, 2009.

[J6]   J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. Graph distances in the data-stream model. *SIAM Journal on Computing*, 38(5):1709–1727, 2009.

[J7]   T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee. Estimating statistical aggregates on probabilistic data streams. *ACM Trans. Database Syst.*, 33(4):1–30, 2008.

[J8]   S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. *Journal of Machine Learning*, 72(1-2):5–19, 2008.

[J9]   J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. On graph problems in a semi-streaming model. *Theoretical Computer Science*, 348(2-3):207–216, 2005.

[J10]   A. Barg and A. McGregor. Distance distribution of binary codes and the error probability of decoding. *IEEE Trans. Inform. Theory*, 51(12):4237– 4246, 2005.

[C1]   A. McGregor, A. Pavan, S. Tirthapura, and D. Woodruff. Space-Efficient Estimation of Statistics over Sub-Sampled Streams. In *ACM Symposium on Principles of Database Systems*, 2012.

[C2]   K. Ahn, S. Guha, and A. McGregor. Graph Sketching: Sparsification, Spanners, and Subgraphs. In *ACM Symposium on Principles of Database Systems*, 2012.

[C3]   K. Ahn, S. Guha, and A. McGregor. Analyzing Graph Structure via Linear Measurements. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 459–467, 2012.

[C4]   A. McGregor and P. Valiant. The Shifting Sands Algorithm. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 453–458, 2012.

[C5]   M. Crouch and A. McGregor. Periodicity and Cyclic Shifts via Linear Sketches. In *APPROX-RANDOM*, pages 158–170, 2011.

[C6]   B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy. A platform for scalable one-pass analytics using MapReduce. In *ACM SIGMOD International Conference on Management of Data*, pages 985–996, 2011. **Invited to Special Issue of ACM Trans. Database Syst.**

[C7]   A. McGregor, A. Rudra, and S. Uurtamo. Polynomial Fitting of Data Streams with Applications to Codeword Testing. In *Symposium on Theoretical Aspects of Computer Science*, pages 428–439, 2011.

[C8] M. Cartright, J. Allan, V. Lavrenko, and A. McGregor. Fast Query Expansion Using Approximations of Relevance Models. In *ACM Conference on Information and Knowledge Management*, pages 1573–1576, 2010.

[C9] A. Chakrabarti, G Cormode, R. Kondapally, and A. McGregor. Information Cost Tradeoffs for Augmented Index and Streaming Language Recognition. In *IEEE Symposium on Foundations of Computer Science*, pages 387–396, 2010.

[C10] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. The Limits of Two-Party Differential Privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 81–90, 2010.

[C11] T. Tran, A. McGregor, Y. Diao, L. Peng, and A. Liu. Conditioning and Aggregating Uncertain Data Streams: Going Beyond Expectations. In *Proceedings of the VLDB Endowment*, pages 1302-1313, 2010.

[C12] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing Linear Counting Queries Under Differential Privacy. In *ACM Symposium on Principles of Database Systems*, pages 123–134, 2010.

[C13] S. Chien, K. Ligett, and A. McGregor. Space-Efficient Estimation of Robust Statistics and Distribution Testing. In *Innovations in Computer Science*, pages 251–265, 2010.

[C14] A. Chakrabarti, G. Cormode, and A. McGregor. Annotations in Data Streams. In *International Colloquium on Automata, Languages and Programming*, pages 222–234, 2009.

[C15] A. McGregor, K. Onak, and R. Panigrahy. The Oil Searching Problem. In *European Symposium on Algorithms*, pages 504–515, 2009.

[C16] G. Cormode, A. Deligiannakis, M. Garofalakis, and A. McGregor. Probabilistic Histograms for Probabilistic Data. In *Proceedings of the VLDB Endowment*, pages 526–537, 2009.

[C17] G. Cormode, L. Golab, F. Korn, A. McGregor, D. Srivastava, and X. Zhang. Sampling to Estimate Conditional Functional Dependencies. In *ACM SIGMOD International Conference on Management of Data*, pages 469–482, 2009.

[C18] K. Chaudhuri and A. McGregor. Finding metric structure in information theoretic clustering. In *Conference on Learning Theory*, pages 391–402, 2008.

[C19] S. Guha and A. McGregor. Tight multi-pass stream lower bounds via pass elimination. In *International Colloquium on Automata, Languages and Programming*, pages 760–772, 2008.

[C20] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *ACM Symposium on Principles of Database Systems*, pages 191–200, 2008.

[C21] A. Chakrabarti, G. Cormode, and A. McGregor. Robust lower bounds for communication and stream computation. In *ACM Symposium on Theory of Computing*, pages 641–650, 2008.

[C22] S. Angelov, K. Kunal, and A. McGregor. Sorting and selection with random costs. In *Latin American Theoretical Informatics Symposium*, pages 48–59, 2008.

[C23] P. Indyk and A. McGregor. Declaring independence via the sketching of sketches. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 737–745, 2008.

[C24]   A. McGregor and O. Milenkovic. On the hardness of approximating stopping and trapping sets in LDPC codes. In *IEEE Information Theory Workshop*, pages 248–253, 2007.

[C25]   S. Guha and A. McGregor. Lower bounds for quantile estimation in random-order and multi-pass streaming. In *International Colloquium on Automata, Languages and Programming*, pages 704–715, 2007.

[C26]   M. Chu, S. Kannan, and A. McGregor. Checking and spot-checking of heaps. In *International Colloquium on Automata, Languages and Programming*, pages 728–739, 2007.

[C27]   S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. In *Conference on Learning Theory*, pages 424–438, 2007. **Invited to Special Issue of the Journal of Machine Learning.**

[C28]   T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee. Estimating statistical aggregates on probabilistic data streams. In *ACM Symposium on Principles of Database Systems*, pages 243–252, 2007. **Invited to Special Issue of ACM Transactions on Database Systems.**

[C29]   S. Guha and A. McGregor. Space-efficient sampling. In *AISTATS*, pages 169–176, 2007.

[C30]   A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 328–335, 2007.

[C31]   A. McGregor and B. Shepherd. Island hopping and path coloring with applications to WDM network design. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 864–873, 2007. **SIGACT "Significant papers on new areas published in proceedings 2007" (1 of 13).**

[C32]   S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.

[C33]   D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. Spatial scan statistics: approximations and performance study. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 24–33, 2006.

[C34]   S. Guha and A. McGregor. Approximate quantiles and the order of the stream. In *ACM Symposium on Principles of Database Systems*, pages 273–279, 2006.

[C35]   J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. Graph distances in the streaming model: the value of space. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 745–754, 2005.

[C36]   S. Kannan and A. McGregor. More on reconstructing strings from random traces: Insertions and deletions. In *IEEE International Symposium on Information Theory*, pages 297–301, 2005.

[C37]   B. Harb, S. Kannan, and A. McGregor. Approximating the best-fit tree under l_p norms. In *APPROX-RANDOM*, pages 123–133, 2005.

[C38]   A. McGregor. Finding graph matchings in data streams. In *APPROX-RANDOM*, pages 170–181, 2005.

[C39]   J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. On graph problems in a semi-streaming model. In *International Colloquium on Automata, Languages and*

Programming, pages 531–543, 2004. **Invited to Special Issue of Theoretical Computer Science.**

[C40] A. McGregor. A problem in scheduling: Your time starts now.... In *FUN with Algorithms*, pages 34–40, 2004.

[C41] A. Barg and A. McGregor. List decoding of concatenated codes: improved performance estimates. In *IEEE International Symposium on Information Theory*, page 419, 2004.

[C42] T. Batu, S. Kannan, S. Khanna, and A. McGregor. Reconstructing strings from random traces. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 910–918, 2004.

[C43] A. Barg and A. McGregor. More on the reliability function of the binary symmetric channel. In *IEEE International Symposium on Information Theory*, page 115, 2003.

[C44] A. Barg and A. McGregor. Distance distribution of binary codes and the error probability of decoding. In *Workshop in Coding and Cryptography*, pages 51–61, 2003.

[C45] A. McGregor, E. Miranda, and P. Gawthrop. Physical modeling of musical instruments using bond graphs. In *Brazilian Symposium on Computer Music*, 1999.

TEACHING EXPERIENCE	**University of Massachusetts**, Amherst, USA	**2009 – Present**

CMPSCI 240 Reasoning About Uncertainty (Spring 2010, Fall 2011).
CMPSCI 611 Advanced Algorithms (Fall 2009, Fall 2010).
CMPSCI 696 Independent Study (Spring 2010, Spring 2011).
CMPSCI 711 Randomized Algorithms (Spring 2009).
CMPSCI 711 Data Stream Algorithms (Spring 2012).
CMPSCI 891M Theory Seminar (Spring 2010, Spring 2011).

University of Pennsylvania, Philadelphia, USA   **2002 – 2003**

Teaching assistant for CSE 334 Advanced Algorithms and CIS 502 Algorithms.

DEPARTMENT SERVICE   **University of Massachusetts**, Amherst, USA   **2009 – Present**

Graduate Admissions (2009-10)
Annual Faculty Review committee (2009-10, 2011-12)
Hiring Committee (2010-11)
Graduate Committee (2010-11)
Professionalism Seminar (Job Hunting) Panelist (2009, 2010)
Organizer of the Distinguished Lecturer Series (2011-2012)

Previous Departments   **2000 – 2008**

Organizer and Co-Creator of the UCSD ITA Seminar (2007 – 08)
Organizer of the Penn Algorithms and Complexity Seminar (2003 – 06)
Graduate Student Representative at Penn C.I.S. Faculty Meetings (2003 – 04)
Organizer and Co-Creator of the Bi-weekly Penn C.I.S. Departmental Social (2002 – 04)
Organizer of the Penn Theory Lunch (2001 – 02)
President of the Archimedeans, University of Cambridge Maths Society (2000 – 01)

INVITED TALKS   Invited Lecture Courses and Tutorials:

L'Ecole de Printemps d'Informatique Théorique (2012)

Stream Processing Workshop, Johns Hopkins Applied Physics Laboratory (2011)
Coding, Complexity, and Sparsity Workshop, University of Michigan (2011)
Stream Processing Workshop, Johns Hopkins Applied Physics Laboratory (2010)
Barbados Workshop on Computational Complexity (2009)

Invited Sessions and Workshops:

NII Shonan Meeting "Large-scale Distributed Computation" (2012)
Bertinoro Workshop on Sublinear Algorithms (2011)
5th Bertinoro Workshop on Randomized Algorithms and Graphs (2010)
Barbados Workshop on Computational Complexity (2009)
Dagstuhl Seminar 08341 on Sublinear Algorithms (2008)
NIPS Workshop "Representations & Inference on Distributions" with P. Indyk (2007)
Allerton Conference on Communication, Control, and Computing (2007)
UCSD, Information Theory and Applications Workshop (2007)
AMS & MAA Joint Mathematics Meeting (2007)
IIT Kanpur Workshop on Algorithms for Data Streams (2006)
Bertinoro Workshop on Space-Conscious Algorithms (2006)
Dagstuhl Seminar 05291 on Sublinear Algorithms (2005)
DIMACS Workshop on Algebraic Coding Theory and Information Theory (2003)

Seminars (Universities):

Aarhus University, Dartmouth College, Dortmund University of Technology, Hong Kong University of Science and Technology, King's College London, Massachusetts Institute of Technology, McGill University, Purdue University, Tsinghua University, University of California (Merced & San Diego), University of Cambridge, University of Edinburgh, University of Illinois (Urbana Champaign), University of Maryland, University of Michigan, Université Paris Sud, University of Pennsylvania, University of Utah, Washington University of St. Louis, and Williams College.

Seminars (Industry):

Alcatel-Lucent Bell Labs, AT&T Labs, Google NYC, IBM Almaden, Microsoft SVC, and Yahoo! Research.

STUDENT
SUPERVISION

Thesis Committees:

Milad Ebtehaj, 2011 (University of Massachusetts, Advisor: Iqbal Agha)
Chao Li, 2011 (University of Massachusetts, Advisor: Gerome Miklau)
Thanh Tran, 2010 (University of Massachusetts, Advisor: Yanlei Diao)
Valerio Grossi, 2009 (University of Pisa, Advisor: Franco Turini)

Ph.D. Student Supervision:

Mark McCartin-Lim
Michael Crouch

Undergraduate Supervision (REU and Honors Thesis):

John Brattin
Marco Leandro Carmosino
Nicolas Scarrci
Vinay Shah (**CRA Outstanding Undergraduate Research**, Honorable Mention)
Daniel Stubbs

Synthesis Project Supervision:

Marc Cartright
Sam Huston
Md. Ashraful Alam
Van Dang

Bodou Li
Kriste Krstovski
Daniel Barowy

COMMUNITY INVOLVEMENT

Organizer:

IIT Kanpur Workshop on Algorithms for Data Stream (2009)
DIMACS/DyDAn Workshop on Streaming, Coding, and Compressive Sensing: Unifying Theory and Common Applications to Sparse Signal/Data Analysis and Processing (2009)

Panels:

NSF (2010, 2011, 2012)

Program Committees:

32th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (2013)
44th ACM Symposium on Theory of Computing (2012)
5th International Frontiers of Algorithmics Workshop (2011)
19th ACM Conference on Information and Knowledge Management (2010)
6th IEEE/ACM International Conference on Distributed Computing in Sensor Systems (2010)
21st ACM-SIAM Symposium on Discrete Algorithms (2010)
17th ACM Conference on Information and Knowledge Management (2008)
4th IEEE/ACM International Conference on Distributed Computing in Sensor Systems (2008)

Conference Reviewing:

AAIM, APPROX, COLT, ESA, FSTTCS, FOCS, ICALP, ISAAC, ISIT, ITCS, MFCS, PODC, PODS, RANDOM, SODA, STOC, and STACS.

Journal Reviewing:

SIAM Journal on Computing, Journal of Algorithms, IEEE Transactions on Information Theory, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Communications, Theory of Computing Systems, Discrete Applied Mathematics, VLDB Journal, and ACM Transactions on Algorithms.

Collation and Editing of "Open Problems in Data Streams and Related Topics":

`www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf`

Collation and Editing of "Open Problems In Data Streams, Property Testing, and Related Topics " with P. Indyk, I. Newman, and K. Onak:

`www.cs.umass.edu/ mcgregor/papers/11-openproblems.pdf`

Research Blog: `polylogblog.wordpress.com`

HONORS AND AWARDS

Special Issues for Selected Papers:
Trans. Database Syst. for Best Papers from SIGMOD 2011
Journal of Machine Learning for Best Papers from COLT 2007
Transactions on Database Systems for Best Papers from PODS 2007
Theoretical Computer Science for Best Papers from ICALP 2004

NSF Career Award (2010)
SIGACT's "Significant papers on new areas" (1 of 13 from 2007)
St. Andrew's Society of the State of New York Scholarship (2001)

Brown Prize in Pure Mathematics (2001)
Gonville and Caius Junior (1998) and Senior (2000) Scholarship for Mathematics (2000)
Bodey Prize in Applied Mathematics (1999)
Daily Telegraph & BAAS Young Science Writer of the Year (1998)

REFERENCES   Available on request.