

# Arindam Banerjee

Dept of Computer Science & Engineering  
University of Minnesota, Twin Cities  
(612) 625-0041  
banerjee@cs.umn.edu  
<http://www.cs.umn.edu/~banerjee>

## Education

<b>Ph.D.</b> , University of Texas at Austin Dept of Electrical and Computer Engineering	August 2005
<b>M.Tech.</b> , Indian Institute of Technology, Kanpur Dept of Electrical Engineering	May 1999
<b>B.E.</b> , Jadavpur University Dept of Electronics and Telecommunication Engineering	May 1997

## Professional Experience

Assistant Professor	University of Minnesota, Twin Cities	2005–
Visiting Faculty Fellow	University of Texas at Austin	2006
Research Intern	IBM T. J. Watson Research Center	2003,2004
Research Assistant	University of Texas at Austin	2000-2003
Research Intern	Interwoven Inc.	2000,2001
Teaching Assistant	University of Texas at Austin	1999-2000
Research Assistant	IIT Kanpur	1999
Teaching Assistant	IIT Kanpur	1997-1998

## Research Interests

Machine Learning, Data Mining, Information Theory, Convex Analysis, Applications in Text & Web Mining, Social Network Analysis, and Bioinformatics.

## Awards and Honors

- Invited speaker/academic visitor, Max Planck Institute (MPI) for Biological Cybernetics, Germany, August, 2008.
- Invited speaker, Workshop on Algorithms for Modern Massive Datasets (MMDS), Stanford University, June, 2008.
- Invited academic visitor, Institute of Pure and Applied Mathematics (IPAM), University of California, Los Angeles (UCLA), October, 2007.
- Best of SDM Award, SIAM International Conference on Data Mining, April 2007.
- Best Student Paper runner-up, ACM International Workshop on Knowledge Discovery from Sensor Data (SensorKDD), 2007.
- J. T. Oden Faculty Research Fellowship, Visiting Fellow at the Institute for Computational Engineering and Sciences (ICES), University of Texas at Austin, Summer 2006.
- Nominated for Best Dissertation Award, University of Texas at Austin, 2006.
- Best Research Paper Award, University Cooperative Society Research Excellence Awards, University of Texas at Austin, March 2005.
- Best Paper Award, SIAM International Conference on Data Mining, April 2004.

- IBM PhD fellowship for the academic years 2003-2004 and 2004-2005.
- Invited speaker/academic visitor, Toyota Technological Institute (TTI), University of Chicago, December, 2003.
- Selected for the Indian National Mathematical Olympiads and attended the International Mathematical Olympiads camp for two successive years 1992, 1993.
- 1<sup>st</sup> in IBM internal data mining contest, Summer 2003.
- Various travel awards including KDD 2005, KDD 2004, KDD 2003, ISIT 2004 awards; GEC (UT Austin) travel grant; and NIPS 2004 complimentary registration.
- Bronze Medal as a special academic award from Jadavpur University, in Spring 1997.

## Publications

### Book Chapters:

1. “Anomaly Detection in Transportation Corridors using Manifold Embedding,” A. Agovic, A. Banerjee, A. Ganguly, and V. Protopopescu, in *Knowledge Discovery from Sensor Data*, CRC Press, O. Omitaomu and A. Ganguly, editors, 2008, To appear.
2. “Text Clustering with Mixtures of von Mises-Fisher Distributions,” A. Banerjee, I. Dhillon, J. Ghosh, S. Sra, in *Text Mining: Theory, Applications, and Visualization*, Chapman & Hall/CRC Press, M. Sahami and A. Srivastava, editors, 2008, To appear.
3. “Clustering with Balancing Constraints,” A. Banerjee and J. Ghosh, in *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, S. Basu, I. Davidson, and K. L. Wagstaff, editors, CRC Press, 2008, To appear.
4. “Probabilistic Semi-supervised Clustering with Constraints,” S. Basu, M. Bilenko, A. Banerjee, and R. Mooney, in *Semi-supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, editors, MIT Press, 2006.

### Journal Publications:

1. “Anomaly Detection: A Survey,” V. Chandola, A. Banerjee, V. Kumar, *ACM Computing Surveys*, 2008, To appear.
2. “Anomaly Detection in Transportation Corridors using Manifold Embedding,” A. Agovic, A. Banerjee, A. Ganguly, and V. Protopopescu, *Intelligent Data Analysis*, To appear.
3. “Meta-prediction of Phosphorylation Sites with Weighted Voting and Restricted Grid Search Parameter Selection,” J. Wan, S. Kang, C. Tang, J. Yan, Y. Ren, J. Liu, X. Gao, A. Banerjee, L. Ellis, T. Li, *Nucleic Acids Research*, doi: 10.1093/nar/gkm848, 2008.
4. “A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation,” A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D. Modha, *Journal of Machine Learning Research*, 8 (Aug), 1919-1986, 2007.
5. “Scalable Clustering Algorithms with Balancing Constraints,” A. Banerjee, J. Ghosh, *Data Mining and Knowledge Discovery*, 13(3), 365-395, November, 2006.
6. “A Clustering Based Approach to Perceptual Image Hashing,” V. Monga, A. Banerjee, and B. Evans, *IEEE Transactions on Information Forensics and Security*, 1(1), 68-79, March 2006.
7. “On the Optimality of Conditional Expectation as a Bregman Predictor,” A. Banerjee, X. Guo, and H. Wang, *IEEE Transactions on Information Theory*, 51(7), 2664-2669, 2005.
8. “Clustering with Bregman Divergences,” A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, *Journal of Machine Learning Research*, 6 (Oct), 1705-1749, 2005.

9. "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions," A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, *Journal of Machine Learning Research*, 6 (Sep), 1345-1382, 2005.
10. "Frequency Sensitive Competitive Learning for Balanced Clustering on High Dimensional Hyperspheres," A. Banerjee, and J. Ghosh, *IEEE Transactions on Neural Networks*, 15(3), 702-719, May 2004.

Conference Publications:

1. "Bayesian Co-clustering," H. Shan and A. Banerjee, *IEEE International Conference on Data Mining (ICDM)*, 2008.
2. "Multiplicative Mixture Models for Overlapping Clustering," Q. Fu and A. Banerjee, *IEEE International Conference on Data Mining (ICDM)*, 2008.
3. "I/O Scalable Bregman Co-clustering," K. Hsu, A. Banerjee, and J. Srivastava, *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2008.
4. "Latent Dirichlet Conditional Naive Bayes Models," A. Banerjee and H. Shan, *IEEE International Conference on Data Mining (ICDM)*, 2007.
5. "Multi-way Clustering on Relation Graphs," A. Banerjee, S. Basu, S. Merugu, *SIAM International Conference on Data Mining (SDM)*, 2007 [**Best of SDM Award**].
6. "An Analysis of Logistic Models: Exponential Family Connections and Online Performance," A. Banerjee, *SIAM International Conference on Data Mining (SDM)*, 2007.
7. "Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning," A. Banerjee and S. Basu, *SIAM International Conference on Data Mining (SDM)*, 2007.
8. "On Bayesian Bounds," A. Banerjee, *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 81-88, 2006.
9. "Model-based Overlapping Clustering," A. Banerjee, C. Krumpelman, S. Basu, R. Mooney, and J. Ghosh, *Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining (KDD)*, 532-537, August 2005.
10. "A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation," A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha, *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, 509-514, August 2004.
11. "An Objective Evaluation Criterion for Clustering," A. Banerjee and J. Langford, *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, 515-520, August 2004.
12. "An Information Theoretic Analysis of Maximum Likelihood Mixture Estimation for Exponential Families," A. Banerjee, I. Dhillon, J. Ghosh and S. Merugu, *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 57-64, July 2004.
13. "Optimal Bregman Prediction and Jensen's Equality," A. Banerjee, X. Guo and H. Wang, *Proceedings of the International Symposium on Information Theory (ISIT)*, 169, June 2004.
14. "Clustering with Bregman Divergences," A. Banerjee, S. Merugu, I. Dhillon and J. Ghosh, *Proceedings of the 4th SIAM International Conference on Data Mining (SDM)*, 234-245, April 2004 [**Best Algorithms Paper Award**].
15. "Active Semi-supervision for Pairwise Constrained Clustering," S. Basu, A. Banerjee and R. Mooney, *Proceedings of the 4th SIAM International Conference on Data Mining (SDM)*, 333-344, April 2004.
16. "Generative Model-based Clustering of Directional Data," A. Banerjee, I. Dhillon, J. Ghosh and S. Sra, *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD)*, 19-28, August 2003.
17. "Competitive Learning Mechanisms for Scalable, Incremental and Balanced Clustering of Streaming Texts," A. Banerjee, and J. Ghosh, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, July 2003.

18. "Semi-supervised Clustering by Seeding," S. Basu, A. Banerjee and R. Mooney, *Proceedings of the 19th International Conference on Machine Learning (ICML)*, 19-26, July 2002.
19. "Frequency Sensitive Competitive Learning for Clustering on High-dimensional Hyperspheres," A. Banerjee, and J. Ghosh, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1590-1595, May 2002.
20. "On Scaling Up Balanced Clustering Algorithms," A. Banerjee and J. Ghosh, *Proceedings of the 2nd SIAM International Conference on Data Mining (SDM)*, 333-349, April 2002.

#### Workshop & Other Publications:

1. "A Social Query Model for Decentralized Search," A. Banerjee and S. Basu, *2nd ACM International Workshop on Social Network Mining and Analysis (SNAKDD)*, August, 2008.
2. "Social Topic Models for Community Extraction," N. Pathak, C. Delong, K. Erickson, and A. Banerjee, *2nd ACM International Workshop on Social Network Mining and Analysis (SNAKDD)*, August, 2008.
3. "Anomaly Detection in Transportation Corridors using Manifold Embedding," A. Agovic, A. Banerjee, A. Ganguly, and V. Protopopescu, *1st ACM International Workshop on Knowledge Discovery from Sensor Data (Sensor-KDD)*, August, 2007.
4. "Clustering Algorithms for Perceptual Image Hashing," V. Monga, A. Banerjee and B. Evans, *Proceedings of IEEE Digital Signal Processing Workshop*, August, 2004.
5. "Rate Distortion, Bregman Divergences and Maximum Likelihood Mixture Estimation," A. Banerjee, I. Dhillon, J. Ghosh and S. Merugu, *The Learning Workshop at Snowbird*, April, 2004.
6. "Mean Model Clustering," A. Banerjee and J. Ghosh, *The Learning Workshop at Snowbird*, April, 2003.
7. "Characterizing Visitors to a Website Across Multiple Sessions," A. Banerjee and J. Ghosh, *Proceedings of the National Science Foundation(NSF) Workshop on Next Generation Data Mining*, pp. 218-227, Nov 2002.
8. "Clickstream Clustering using Weighted Longest Common Subsequence," A. Banerjee and J. Ghosh, *Proceedings of the 1st SIAM International Conference on Data Mining (SDM): Workshop on Web Mining*, pp. 33-40, April 2001.
9. "Concept-based Clustering of Clickstream Data," A. Banerjee and J. Ghosh, In *Proceedings of the 3rd International Conference on Information Technology*, pp. 145-160, Dec 2000.
10. "Computerized Tumor Boundary Detection Using Genetic Algorithm," A. Banerjee, *Proceedings of the National Conference on Applications of Signal Processing*, Sept 1998.

#### Technical Reports:

1. "Bayesian Co-clustering," H. Shan and A. Banerjee, *Technical Report TR-08-022*, Department of Computer Science & Engineering, University of Minnesota, Twin Cities, 2008.
2. "A Social Query Model for Decentralized Search," A. Banerjee and S. Basu, *Technical Report TR-08-017*, Department of Computer Science & Engineering, University of Minnesota, Twin Cities, 2008.
3. "Social Topic Models for Community Extraction," N. Pathak, C. Delong, K. Erickson, and A. Banerjee, *Technical Report TR-08-005*, Department of Computer Science & Engineering, University of Minnesota, Twin Cities, 2008.
4. "Outlier Detection: A Survey," V. Chandola, A. Banerjee, and V. Kumar, *Technical Report TR 07-017*, Department of Computer Science & Engineering, University of Minnesota, Twin Cities, 2007.
5. "Clustering with Bregman Divergences," A. Banerjee, S. Merugu, I. Dhillon and J. Ghosh, *Technical Report TR-03-19*, Department of Computer Sciences, University of Texas at Austin, 2003.
6. "Expectation Maximization for Clustering on Hyperspheres," A. Banerjee, I. Dhillon, J. Ghosh and S. Sra, *Technical Report TR-03-07*, Department of Computer Sciences, University of Texas at Austin, 2003.

## Professional & Other Activities

### Chair:

- Technical Chair, NASA Conference on Intelligent Data Understanding (CIDU), 2008.

### Program Committee:

- International Conference on Machine Learning (ICML'07, ICML'08)
- Advances in Neural Information Processing Systems (NIPS'06, NIPS'07, NIPS'08)
- National Conference on Artificial Intelligence (AAAI'06, AAAI'07)
- ACM International Conference on Knowledge Discovery and Data Mining (KDD'08)
- SIAM Conference on Data Mining (SDM'06, SDM'07)
- IEEE International Conference on Data Mining (ICDM'05, ICDM'07)

### Reviewing:

- Journal Reviewing: Journal of Machine Learning Research, Machine Learning Journal, IEEE Transactions on Information Theory, IEEE Transactions on Neural Networks, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on System, Man and Cybernetics, Data Mining and Knowledge Discovery, Statistical Analysis and Data Mining, Communications of the ACM, Neurocomputing, Applied Intelligence.
- Conference Reviewing: ACM International Conference on Knowledge Discovery and Data Mining (KDD'03, KDD'04, KDD'05), IEEE International Conference on Data Mining (ICDM'04), SIAM Conference on Data Mining (SDM'04, SDM'05), European Conference on Machine Learning (ECML'03).
- Book Reviewing: “Text Mining: Predictive Methods for Analyzing Unstructured Information” by Sholom Weiss, Nitin Indurkha, Tong Zhang, Fred Damerau, published October 2004 by Springer-Verlag.
- Grant Reviewing: NASA grant proposals for the Intelligent Systems (IS) Project.
- Workshop Reviewing: ACM Workshop on Knowledge Discovery from Sensor Data (Sensor-KDD'07), AAAI Spring Symposium on Social Information Processing (2008).

### Panels:

- Invited panelist on “Text mining: The discipline that never was” organized by Prabhakar Raghavan, head of Yahoo! Research, at the 11th ACM International Conference on Knowledge Discovery and Data Mining (KDD'05).

### Presentations:

- Invited Presentations: Max Planck Institute for Biological Cybernetics (2008), Thomson-Reuters (2008), Stanford University (2008), SRI International (2006), Oak Ridge National Labs (2006), University of Florida at Gainesville (2005), StonyBrook University (2005), IBM T. J. Watson Research Center (2003,2004), Toyota Technological Institute, Chicago (2003).
- Conference Presentations: SIAM Conference on Data Mining (SDM'07, SDM'04,SDM'02), International Conference on Machine Learning (ICML'06), ACM International Conference on Knowledge Discovery and Data Mining (KDD'03,KDD'04,KDD'05), IEEE International Symposium on Information Theory (ISIT'04), International Conference on Information Technology (2000), National Conference on Applications of Signal Processing (1998).

## Advising

1. Students: Amrudin Agovic (co-advised with Maria Gini), Varun Chandola (co-advised with Vipin Kumar), Qiang Fu, Hanhuai Shan, Nisheeth Srivastava, Roman Briskine (MS'07), Charles Curtsinger (UROP'07).
2. PhD Thesis Committees: Stefan Atev, Nathaniel Bird, Bret Borghetti, Mete Celik, Zachary Crockett, Steven Damer, Seongwook Jeong, Bridget McInnes, Faraz Mirzaei, Fikri Goksu, Gyorgy Simon, Shilad Sen, Nikolas Trawny.

## Teaching

### Academic Courses Created:

- CSci 8980: Topics in Machine Learning, first offered Spring 2006.
- CSci 5525: Machine Learning, first offered Fall 2006.
- CSci 8980: Advanced Topics in Graphical Models, first offered Fall 2007.

## Departmental Service

2008-09 Graduate Admissions Committee  
2007-08 Graduate Admissions Committee  
2006-07 Strategic Planning Committee  
2005-06 Graduate Admissions Committee

## Technology Transfer

- Member of startup Neonyoyo Inc., developed recommendation systems for wireless and internet applications. The company was formed and later acquired by Interwoven Inc. in 2000.
- Project with Oak Ridge National Labs, developed data mining package for anomaly detection in truck weigh-station data.