Purnamrita Sarkar

Department of Statistics and Data Sciences
The University of Texas at Austin
Email: purna.sarkar@austin.utexas.edu
Webpage: http://psarkar.github.io

Education

M.S+Ph.D. (Aug 2004 – Aug 2010). Machine Learning Department, School of Computer Science, Carnegie Mellon University. Advisor: Andrew W. Moore

B. Tech. (May 2000 – May 2004). Computer Science and Engineering Department, Indian Institute of Technology, Kharagpur

Research interests

Broadly speaking I am interested in large scale machine learning and statistical inference. In recent years, some of the specific topics that have interested me are : "spectral clustering consistency for blockmodels", "variational bayes consistency", "detecting overlapping communities in networks", "estimation and inference in random dot product graphs", "dynamic networks from latent space models".

Work experience

Aug 2014 – present: Assistant Professor, Department of Statistics and Data Sciences, The University of Texas at Austin

Aug 2013 – 2014: Postdoctoral Researcher. Sponsors: Peter J. Bickel, Department of Statistics, and Michael I. Jordan, Department of Statistics and Electrical Engineering and Computer Sciences, University of California, Berkeley

Aug 2010 – Aug 2013: Postdoctoral Researcher, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

Fall 2006: Research Intern, Google Inc. Pittsburgh, PA

Publications

(Student trainee*, postdoctoral scholar‡ at the time of preparation)

Thesis

- P. Sarkar. "Tractable Algorithms for Proximity Search on Large Graphs." Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2010.

Book chapters

- P. Sarkar and A. W. Moore. "Role of Random Walks in Ranking Applications and Social Network Analysis." *Social Network Data Analytics. Ed. Charu Aggarwal, Springer, 2010.*

In preparation

- Q. Lin*, R. Lunde‡ and P. Sarkar. "The network jackknife: successes and failures."
- R. Lunde‡ and P. Sarkar. "Resampling methods for local network statistics."
- R. Lunde‡ , P. Sarkar, R. Ward "Bootstrapping the error of Oja's algorithm."

Submitted/Under revision

- Q. Lin*, R. Lunde‡ and P. Sarkar. "Higher Order Correct Multiplier Bootstraps for Networks." *Preprint on arXiv:2009.06170*.
- S. Mukherjee*, P. Sarkar, and P. Bickel. "Two provably consistent divide and conquer clustering algorithms for large networks." Major Revision. PNAS. **Best student paper award at the 2017 IISA conference.** *Preprint on arXiv:1708.05573*.
- R. Lunde‡ and P. Sarkar. "Subsampling Sparse Graphons Under Minimal Assumptions." *Preprint on arXiv:1907.12528*. Major Revision. Biometrika.
- P. Srivastava*, P. Sarkar, and G. Hanasusanto "A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers" Major Revision. Operations Research. *Preprint on arXiv:1912.07546.* **Honorable mention in the 2020 INFORMS Computing Society Best Student Paper Competition.**

Journal papers

- P. Sarkar, Y. X. R. Wang, and S. Mukherjee. "When random initializations help: a study of variational inference for community detection." *Journal of Machine Learning Research, 2021.*
- T. Li*, L. Lei, S. Bhattacharya, K. Van den Berge*, P. Sarkar, L. Levina and P. J. Bickel "Hierarchical community detection by recursive partitioning." *JASA Theory and Methods, 2020*.
- X. Mao*and P. Sarkar, and D. Chakrabarti. "Estimating Mixed Memberships with Sharp Eigenvector Deviations." *JASA Theory and Methods, 2020*.
- B. Yan*and P. Sarkar. "Covariate Regularized Community Detection in Sparse Graphs." *JASA Theory and Methods, 2020*.
- R. Wang, P. Sarkar, O. Ursu*, A. Kundaje and P. Bickel. "Network modeling of topological domains using Hi-C data." *Annals of Applied Statistics, 2019, Vol 13, number 3, 1511-1536*.
- P. Sarkar and P. J. Bickel. "Role of Normalization in Spectral Clustering for Stochastic Blockmodels." *Annals of Statistics, 2015, Vol 43, number 3, 962-993*.
- P. J. Bickel and P. Sarkar. "Hypothesis Testing for Automated Community Detection in Networks." *Journal of the Royal Statistical Society, Series B, 2015, Volume 78, Issue 1*.
- P. Sarkar, D. Chakrabarti, and M. I. Jordan. "Nonparametric Link Prediction in Large Scale Dynamic Networks." *Electronic Journal of Statistics, 2014, Vol 8, Number 2*.
- A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. "A Scalable Bootstrap for Massive Data." *Journal of the Royal Statistical Society, Series B, 2014, Vol 76, Issue 4*.
- P. Sarkar and A. W. Moore. "Dynamic Social Network Analysis using Latent Space Models." *ACM SIGKDD Explorations, Special Issue on Link Mining, 2005*.
- J. Leskovec, P. Sarkar, and C. Guestrin. "Modeling Link Qualities in a Sensor Network." *Informatica, 29(4): 445-452, 2005*.

Major conference papers

- X. Mao*, D. Chakrabarti and P. Sarkar "Consistent Nonparametric Methods for Network Assisted Covariate Estimation" *Thirty-eighth International Conference on Machine Learning (ICML) 2021.*
- Q. Lin*, R. Lunde‡ , P. Sarkar "On the Theoretical Properties of the Network Jackknife ." *Thirty-seventh International Conference on Machine Learning (ICML) 2020.*
- X. Fan*, Y. Yue*, P. Sarkar, Y. X. R. Wang " On hyperparameter tuning in general clustering problems ." *Thirty-seventh International Conference on Machine Learning (ICML) 2020.*
- M. Yin*, Y. X. R. Wang, P. Sarkar "A Theoretical Case Study of Structured Variational Inference for Community Detection ." *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*.
- S. Mukherjee*, P. Sarkar, Y. X. R. Wang, and B. Yan*. "Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues." *The 32nd conference of Neural Information Processing Systems (NeurIPS), 2018.*

- X. Mao* and P. Sarkar, and D. Chakrabarti. "Overlapping Clustering Models, and One (class) SVM to Bind Them All." *The 32nd conference of Neural Information Processing Systems (NeurIPS), 2018.* ***(Spotlight presentation)***
- B. Yan*, P. Sarkar, and X. Cheng "Exact Recovery of Number of Blocks in Blockmodels." *AISTATS, 2018.*
- B. Yan*, M. Yin*, and P. Sarkar "Statistical Convergence Analysis of Gradient EM on General Gaussian Mixture Models." *The 31st conference of Neural Information Processing Systems (NeurIPS), 2017.*
- S. Mukherjee*, P. Sarkar, and L. Lin "On Clustering Network Valued Data." *The 31st conference of Neural Information Processing Systems (NeurIPS), 2017.*
- X. Mao*, P. Sarkar, and D. Chakrabarti "On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations." *The 34th International Conference on Machine Learning (ICML), 2017.*
- B. Yan* and P. Sarkar "On Robustness of Kernel Clustering" *The 30th conference of Neural Information Processing Systems (NeurIPS), 2016.*
- P. Sarkar, D. Chakrabarti, and P. Bickel "Consistency of Common Neighbors for Link Prediction in Stochastic Blockmodels." *The 29th conference of Neural Information Processing Systems (NeurIPS), 2015.*
- B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, S. Madden " Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning." *The 41st International Conference on Very Large Databases (VLDB), 2014.*
- B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. "Crowdsourced Enumeration Queries." *The 29th International Conference on Data Engineering (ICDE), 2013. (**Best paper award**).*
- P. Sarkar, D. Chakrabarti, and M. I. Jordan. "Nonparametric Link Prediction in Dynamic Networks." *The 29th International Conference on Machine Learning (ICML), 2012.*
- A. Kleiner, A. Talwalkar, P. Sarkar and M. I. Jordan. "The Big Data Bootstrap." *The 29th International Conference on Machine Learning (ICML), 2012.*
- P. Sarkar, D. Chakrabarti, and A. W. Moore. "Theoretical Justification of Popular Link Prediction Heuristics." *The 23rd Annual Conference on Learning Theory (COLT), 2010 (**Best student paper award**).*
  - **Invited to the Best Paper Track**, *International Joint Conference on Artificial Intelligence (IJCAI), 2011.*
- P. Sarkar and A. W. Moore. "Fast Nearest-neighbor Search in Disk-resident Graphs." *The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2010.*
- P. Sarkar and A. W. Moore. "Fast Dynamic Reranking in Large Graphs." *The 18th International World Wide Web Conference (WWW), Data Mining Track, 2009.*
- P. Sarkar, A. W. Moore and A. Prakash. "Fast Incremental Proximity Search in Large Graphs." *The 25th International Conference on Machine Learning (ICML), 2008.*
- P. Sarkar and A. W. Moore. "A Tractable Approach to Finding Closest Truncated-commute-time Neighbors in Large Graphs." *The 23rd Conference on Uncertainty in Artificial Intelligence (UAI), 2007.*
- P. Sarkar, S. Siddiqi and G. Gordon. "A Latent Space Approach to Dynamic Embedding of Co-occurrence Data." *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS), 2007.*
- P. Sarkar and A. W. Moore. "Dynamic Social Network Analysis using Latent Space Models." *Advances in Neural Information Processing Systems (NIPS), 2005.*

Grants

- Lead PI on NSF DMS - 2109155: "Learning with confidence: Bootstrapping error estimates for stochastic iterative algorithms," 2021. Co-PI: Rachel Ward, Department of Mathematics. Status: Recommended for funding.

- Senior personnel on NSF grant 2019844- AI Institute: A Vision for the Next Decade of Foundational Machine Learning.
- Co PI on NSF 1934932: HDR TRIPODS: UT Austin Institute on the Foundations of Data Science, 2019. PI-Sujay Sanghavy, ECE, other Co-PI's Rachel Ward, Mathematics, Adam Klivans, Computer Science.
- Lead PI on Collaborative Research Grant NSF DMS 1713082: "Inference for Network Models with Covariates: Leveraging Local Information for Statistically and Computationally Efficient Estimation of Global Parameters", 2017.
- Unrestricted gift funding of 7000$ from Adobe Research.

Awards

- Honorable mention in the 2020 INFORMS Computing Society Best Student Paper Competition.
- Best student paper award at IISA 2017
- Best paper award at ICDE 2013,
- Best student paper award at COLT 2010.

Postdocs and Students

- Robert Lunde, Postdoctoral Scholar.
  - Postdoctoral scholar at the Department of Statsitics, University of Michigan 2021-2022.
  - Tenure-track assistant professor at the Department of Mathematics and Statistics, Washington University, St. Louis, starting Fall, 2022.
- Xueyu Mao, Ph.D. student. Applied Scientist, Amazon
- Prateek Srivastava, Ph.D. student. (coadvised with Grani Hanasusanto, Department of Operations Research & Industrial Engineering, U. T. Austin) To start as Research Scientist, Amazon
- Bowei Yan, Ph.D. student. Research Scientist, Nuro
- Qiaohui Lin, Continuing Ph.D. student. (coadvised with Peter Mueller, Department of Statistics and Data Sciences, U. T. Austin)

Talks

- New Directions in Statistical Inference on Networks and Graphs, Banff International Research Station, 2021 "Canceled."
- Networks 2021, A joint Sunbelt and Netsci conference "Resampling methods in networks.'
- Joint Statistical Meetings (JSM 2021) "Resampling methods in networks."
- International Indian Statistical Association annual conference (IISA 2021) "Trading off Accuracy for Speedup: Multiplier Bootstraps for Subgraph Counts"
- The 26th Pfizer/ASA/UConn Distinguished Statistician Colloquium, 2021 "Discussant"
- CM Statistics, 2020 "Resampling methods in networks."
- Department of Mathematics, University of Maryland, College Park, 2020 "Resampling methods in networks."

- ASA Statistical Learning and Data Science Section webinar " Introductory Overview Lectures in Social Network"
- INFORMS, 2020 "A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers."
- Interdisciplinary Statistical Research Unit seminar, Indian Statistical Institute, 2019 "Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues."
- U. C. Berkeley– Neyman Seminar, 2019 "Overlapping clustering models and one (class) SVM to bind them all."
- U. C. Davis, Peter Hall conference, 2019 "Overlapping clustering models and one (class) SVM to bind them all."
- IEEE Data Science Workshop (DSW 2019) "Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues"
- Joint Statistical Meetings (JSM 2019) "Overlapping clustering models and one (class) SVM to bind them all."
- International Indian Statistical Association annual conference (IISA 2019) "Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues."
- Department of Statistics, Boston University, 2019. Unable to attend.
- Statistical Scalability program, Isaac Newton Institute of Mathematical Sciences, 2018. Declined.
- CMStatistics, 2018. Declined.
- International Indian Statistical Association annual conference, (IISA 2016) "On mixed memberships and matrix factorization"
- Royal Statistical Society International Conference, Manchester, 2016, Methods & Theory: Journal of the Royal Statistical Society Series B Editors' Invited Session, "Hypothesis Testing for Automated Community Detection in Networks"
- Graph limits and statistics workshop, Isaac Newton Institute of Mathematical Sciences, 2016 "Convex Relaxation for Community Detection with Covariates"
- Department of Statistics and Data Sciences, UT Austin, 2014 "Towards a theory for network clustering",
- ORIE, Cornell University, 2014 "Towards a theory for network clustering"
- ORIE, University of Texas at Austin, 2014 "Towards a theory for network clustering"
- Department of Statistics, Ohio State University, 2014 "Towards a theory for network clustering"
- Department of Statistics, Purdue University, 2014 "Towards a theory for network clustering"
- ORFE, Princeton University, 2014 "Towards a theory for network clustering"
- Department of Statistics, UCLA, 2014 "Towards a theory for network clustering"
- Department of Computer Science and Engineering, Ohio State University, 2014 "Link prediction: theory and practice"
- Department of Computer Science, University of Illinois at Urbana Champaign, 2014 "Link prediction: theory and practice"
- Department of Computer Science, University of California, Santa Cruz, 2014 "Link prediction: theory and practice"
- Workshop on Large Graphs: Modeling, Algorithms, and Applications, IMA, 2011 "Nonparametric Link Prediction".

- Complex Network Modeling Workshop, SAMSI, 2010 "Theoretical Justification of Popular Link Prediction Heuristics"
- Joint Statistical Meetings, 2010 "Probabilistic Modeling of Dynamic Networks using Latent Space Models"
- Google Research, New York, 2007 "A Tractable Approach to Finding Closest Truncated hitting- time Neighbors in Large Graphs."
- The Snowbird Workshop, 2007 "A Tractable Approach to Finding Closest Truncated-hitting time Neighbors in Large Graphs."
- Social Net Mid-Year Workshop, SAMSI, 2006 "Dynamic Social Network Analysis using Latent Space Models."

Teaching

- SDS 321: Introduction to Probability and Statistics
  - Level: undergraduate
  - Years: 2015, 2016, 2017
- SDS 383C: Statistical Modeling I
  - Level: graduate
  - Years: 2015, 2016
- SDS 384: Theoretical Statistics
  - Level: graduate
  - Years: 2018, 2019, 2020, 2021
- SDS 385: Stat Models for Big Data
  - Level: graduate
  - Years: 2018, 2019, 2020, 2021 (upcoming)
- DSC 383 : Advanced Predictive Models for Complex Data
  - Level: Masters, online
  - Years: 2021 (upcoming)

Service

Departmental:

- Scientific board, Machine Learning Laboratory
- Online MS in Data Science (Option III) Oversight/Curriculum Committee, 2019-2021
- Faculty search committee : 2017,2019
- Chair search committee: 2018
- Co-organized SDS seminars from 2014-2017.
- Thesis committee:
  - Abhinandan Dalal, Masters in Statistics, Indian Statistical Institute
  - Xinjie Fan, PhD candidate, SDS
  - Novin Ghaffari, PhD candidate, SDS

- Yuege Xie, PhD candidate, Oden Institute
- Mackenzie M. Johnson, PhD candidate, Integrative Biology
- Shuying Wang, PhD candidate, SDS
- Xi Chen, PhD candidate, ORIE
- Mingzhang Yin, PhD candidate, SDS
- Su Chen, PhD candidate, SDS
- M. Narayana Prasad, PhD candidate, ORIE
- Oscar M. Padilla, PhD candidate, SDS

College:

- Promotion review committee for lecturers: 2018
- Served as a member of the selection committee for "FRA for Natural Sciences" 2017 – 2018

External:

- Major Conference Program Committees- ICML 2008, ICML 2010, KDD 2011.
- Senior program committee member/ Area chair- AISTATS 2014 and 2015, NeurIPS (previously NIPS) 2016-2021.
- Journal, Conference, and Book Reviewing - International Conference of Machine Learning (ICML), Neural Information Processing Systems (NeurIPS), ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Conference of Learning Theory (COLT), Journal of Machine Learning Research (JMLR), Annals of Statistics (AOS), Journal of the Royal Statistical Society, Series B (JRSSB) and C (JRSSC), The Journal of American Statistical Association: theory and methods (JASA), Statistical Science, Statistica Sinica, Journal of Computational and Statistical Graphics (JCGS)
- Grant Proposal Reviewing- NSF Data Mining proposal review panel, 2012. NSF Statistics proposal review panel 2018. Invited to review for NSF Harnessing the Data Revolution Institutes 2021 (Declined), Invited to review for NSF statistics review panel 2021 (Declined).