Algorithms for Big Data

Chandra Chekuri

CS498ABD, Spring 2019

Project Details

- Project constitutes 25% of the grade
- The due date is **10am Friday, May 3rd.**
- Pick a paper or set of papers or reading material and send private note on Piazza to course staff by **10am Monday, April 1st.** Earlier is better.

Aim: The goal of the project is to give you an opportunity to examine a paper/topic in depth. It should be relevant to the contents of the course and should be of some current interest. An ambitious, and an achievable outcome, is to do original research inspired by the project.

Guidelines: Projects can be done individually or in groups of up to two. You should pick a paper (or two or more) from the list below or from your own search, confirm the choice with the course staff, and write a report and submit it via Gradescope by the deadline mentioned above. Moses Charikar, outlines quite well, the scope and contents of a good project and report. See here.

Potential Papers:

Several of the papers below are taken from Moses Charikar's project page. This list is far from exhaustive on these topics. As mentioned above you are free to come up with your own paper(s) and consult with us.

Hashing

- Fast Cross-Polytope Locality-Sensitive Hashing, Kennedy, R. Ward, 2016.
- Practical and Optimal LSH for Angular Distance, A. Andoni et al., NIPS 2015.
- Optimal Data-Dependent Hashing for Approximate Near Neighbors, Andoni, Razenshteyn, STOC 2015.
- LSH Forest: Practical Algorithms Made Theoretical, Andoni, Razenshteyn, Nosatzki, SODA 2017.
- Hashing for statistics over k-partitions, Dahlgaard et al., FOCS 2015.
- Non-Empty Bins with Simple Tabulation Hashing, Aamand and Thorup, SODA 2019.

Sketching, Streaming, Sparsification

- Sketching and Embedding are Equivalent for Norms, Andoni, Krauthgamer, Razenshteyn, STOC 2015.
- On Fully Dynamic Graph Sparsifiers I. Abraham et al., 2016.
- An Optimal Algorithm for 鈥$_1$-Heavy Hitters in Insertion Streams and Related Problems, A. Bhattacharyya et al., PODS 2016
- Graphical Model Sketch, Kveton et al., 2016.
- Heavy hitters via cluster-preserving clustering, K.G. Larsen et al., 2016.
- BPTree: an 鈥$_2$ heavy hitters algorithm using constant memory, V. Braverman etl., 2016.
- Beating CountSketch for Heavy Hitters in Insertion Streams, V. Braverman et al., STOC 2016.
- L_p Row Sampling by Lewis Weights, M.B. Cohen, R. Peng, STOC 2015.
- A Unified Approach for Clustering Problems on Sliding Windows, V. Braverman et al., 2015.
- Sketching for M-estimators: a unified approach to robust regression, K.L. Clarkson, D.P. Woodruff, SODA 2015.
- Single Pass Spectral Sparsification in Dynamic Streams, M. Kapralov et al., FOCS 2014.
- Counting Arbitrary Subgraphs in Data Streams, D.M. Kane et al., ICALP 2012
- HyperLogLog Hyper Extended: Sketches for Concave Sublinear Frequency Statistics, Edith Cohen, KDD 2017.
- Streaming Symmetric Norms via Measure Concentration, Blasiok et al. STOC 2017
- Optimal streaming and tracking distinct elements with high probability Blasiok. SODA 2018.
- Perfect L_p Sampling in a Data Stream , Jayaram and Woodruff, FOCS 2018.
- Nearly Optimal Distinct Elements and Heavy Hitters on Sliding Windows, Braverman etal, APPROX-RANDOM 2018.

Dimensionality Reduction, Core sets

- Universality Laws for Randomized Reduction with Applications, S Oymak, J.A. Tropp, 2015.
- Near-Optimal Bounds for Binary Embeddings of Arbitrary Sets, S Oymak, B. Recht, 2015.
- Isometric sketching of any set via the Restricted Isometry Property, S. Oymak et al, 2015.
- Dimensionality Reduction for k-Means Clustering and Low Rank Approximation, M. B. Cohen et al., STOC 2015.
- Towards a Unified Theory of Sparse Dimensionality Reduction in Euclidean Space, J Bourgain et al, STOC 2015
- Spectral Approaches to Nearest Neighbor Search, Abdullah et al., FOCS 2014.
- Sparser Johnson-Lindenstrauss Transforms, Daniel M. Kane, Jelani Nelson, SODA 2012.
- Improved coresets for kernel density estimates, Phillips and Ming Tai, SODA 2018.
- Understanding Sparse JL for Feature Hashing, Meena Jagadeesan, 2019.
- Performance of Johnson-Lindenstrauss Transform for k-Means and k-Medians Clustering, Makarychev etal. STOC 2019.

Fast Numerical Linear Algebra

- Approximate Gaussian Elimination for Laplacians, R. Kyng, S. Sachdeva, 2016

- Streaming PCA: Matching Matrix Bernstein and Near-Optimal Finite Sample Guarantees for Oja's Algorithm, P. Jain et al, 2016.
- Optimal Principal Component Analysis in Distributed and Streaming Models, C. Boutsidis et al., STOC 2016.
- Weighted Low Rank Approximations with Provable Guarantees, I. Razenshteyn et al., STOC 2016.
- Nearly tight oblivious subspace embeddings by trace inequalities, M.B. Cohen, SODA 2016.
- Optimal approximate matrix product in terms of stable rank, M.B. Cohen et al., 2015.
- Optimal CUR matrix decompositions, Christos Boutsidis, David P. Woodruff, STOC 2014.
- OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings, Jelani Nelson, Huy L. Nguyen, FOCS 2013.
- Input Sparsity Time Low-Rank Approximation via Ridge Leverage Score Sampling. Michael Cohen, Cameron Musco, Christopher Musco, SODA 2017.

Parallel, Distributed, MapReduce

- Submodular Optimization in the MapReduce model. Paul Liu, Jan Vondr谩k. SOSA, 2019.
- Efficient Massively Parallel Methods for Dynamic Programming. Sungjin Im, Ben Moseley, and Xiaorui Sun
  STOC, 2017.
- Massively Parallel Approximation Algorithms for Edit Distance and Longest Common Subsequence MohammadTaghi Hajiaghayi, Saeed Seddighin, Xiaorui Sun. SODA, 2019.
- Massively Parallel Dynamic Programming on Trees. MohammadHossein Bateni, Soheil Behnezhad, Mahsa Derakhshan, MohammadTaghi Hajiaghayi, Vahab Mirrokni. ICALP, 2018.
- Parallel Graph Connectivity in Log Diameter Rounds. Alexandr Andoni, Clifford Stein, Zhao Song, Zhengyu Wang, Peilin Zhong. FOCS 2018.
- Approximating Edit Distance in Truly Subquadratic Time: Quantum and MapReduce Mahdi Boroujeni, Soheil Ehsani, Mohammad Ghodsi, MohammadTaghi HajiAghayi, Saeed Seddighin. SODA, 2018.
- Randomized Composable Coreset for Matching and Vertex Cover, Sephr Assadi and Sanjeev Khanna, SPAA 2017.
- Coresets Meet EDCS: Algorithms for Matching and Vertex Cover on Massive Graphs, Assadi et al, SODA 2019.
- Sparsifying Distributed Algorithms with Ramifications in Massively Parallel Computation and Centralized Local Computation Mohsen Ghaffari, Jara Uitto. SODA, 2019.
- Shuffles and Circuits (On Lower Bounds for Modern Parallel Computation). Tim Roughgarden, Sergei Vassilvitskii, Joshua R. Wang. JACM, 2018.

Learning Structured Distributions

- Learning k-Modal Distributions via Testing, C. Daskalakis, I. Diakonakolis and R. A. Servedio, 2010.
- Learning Poisson Binomial Distributions, C. Daskalakis, I. Diakonakolis and R. A. Servedio, 2015.

- Learning Mixtures of Product Distributions over Discrete Domains, J. Freedman, R. O'Donnell, and R. A. Servedio, 2008.
- Efficiently Learning Mixtures of Two Gaussians, A. T. Kalai, A. Moitra, and G. Valiant, STOC 2010.
- Robustly Learning a Gaussian: Getting Optimal Error, Efficiently, I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, SODA 2018
- Anaconda: A Non-Adaptive Conditional Sampling Algorithm for Distribution Testing, G. Kamath and C. Tzamos, SODA 2018
- Sample-Optimal Density Estimation in Nearly-Linear Time, J. Acharya, I. Diakonikolas, J. Li, L. Schmidt, SODA 2017

Miscellaneous

- Efficient Density Evaluation for Smooth Kernels, Backurs etal, FOCS 2018.
- Do Less, Get More: Streaming Submodular Maximization with Subsampling, Feldman, Karbasi, Kazemi, NIPS 2018.