CS 598: Algorithms for Big Data

Fall 2014

UIUC
Computer Science Department

Chandra Chekuri

Course Summary

This course will describe some algorithmic techniques developed for handling large amounts of data that is often available in limited ways. Topics that will be covered include data stream algorithms, sampling and sketching techniques, and sparsification, with applications to signals, matrices, and graphs. Emphasis will be on the theoretical aspects of the design and analysis of such algorithms. Prerequisites: CS 573, good background in (discrete) probability

---

Administrative Information

Lectures: Tue, Thu 2 to 3.15pm in Siebel Center 1109.

Instructor: Chandra Chekuri- 3228 Siebel Center, 265-0705, chekuri at illinois dot edu

Office Hours: Wed 10-11am, and by appointment

Grading Policy: There will be 4-5 homeworks, roughly once every two weeks and final course project. Course projects could involve research on a specific problem or topic, a survey of several papers on a topic (summarized in a report and/or talk), or an experimental evaluation. I also expect students to scribe one lecture in latex.

Prerequisites: This is a graduate level class and a reasonable background in algorithms and discrete mathematics would be needed. Knowledge and exposure to probability and linear algebra is necessary.

Reference/Study material:

- Lecture notes from various places:
  - Algorithms for Big Data: Jelani Nelson (Harvard)
  - Data Stream Algorithms: Amit Chakrabarti (Dartmouth)
  - Sub-linear Algorithms: Piotr Indyk and Ronitt Rubenfeld (MIT)
  - Randomized Algorithms for Matrices and Data: Michael Mahoney (Stanford)
  - Algorithms for Modern Data Models: Ashish Goel (Stanford)

  - Algorithmic Techniques for Big Data Analysis: Barna Saha (U. Minnesota)
  - Models of Computation for Massive Data: Jeff Philips (Utah)
- A book in preparation on data stream algorithms by Andrew McGregor and Muthu Muthukrishnan
- A useful book with an emphasis on the practical aspects: Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman and Jeff Ullman.
- Foundations of Data Science, a book in preparation, by John Hopcroft and Ravi Kannan
- Survey on sketching by Graham Cormode
- Sketching as a Tool for Numerical Linear Algebra, a survey by David Woodruff. Copy for personal use from David's website
- Books on randomization: Probability and Computing (Mitzenmacher-Upfal), Randomized Algorithms (Motwani-Raghavan), The Probabilistic Method (Alon-Spencer), Concentration of Measure (Dubhashi-Panconesi)
- A survey on concentration inequalities by Fan Chung and Linyuan Li.
- Course material from Summer School on Hashing: Theory and Applications
- Open Problems in Sublinear Algorithms

Potential Topics:

- Streaming, Sketching and Sampling for Signals.
- Dimensionality Reduction
- Streaming for Graphs
- Numerical Linear Algebra
- Compressed Sensing
- Map-Reduce model and basic algorithms
- Introduction to Property Testing
- Lower Bounds via Communication Complexity

Note: The above list is suggestive/tentative and we will cover only a subset of the topics.

Piazza site for questions and discussion
Moodle site for submitting homeworks.

Homework:

Homework 1 given on Thursday 09/2/14, due in class on Thursday, 9/11/14.
Homework 2 given on Friday 09/12/14, due on Thursday, 9/23/14.
Homework 3 given on Friday 10/10/14, due on Thursday, 10/23/14.
Homework 4 given on Friday 11/7/14, due on Thursday, 11/20/14.

Lectures:

Sample LaTeX file and algo.sty

Warning: Notes may contain errors. Please bring those to the attention of the instructor.

- Lecture 1: 8/26/14, Introduction, basics of probability, probabilistic counting (Morris's algorithm),

reservoir sampling
  - Lecture 1 in in Jelani Nelson's course
  - Counting large numbers of events in small registers by Morris, CACM 1978
  - Approximate counting: a detailed analysis by Flajolet.
  - Random sampling with a reservoid by Vitter.
  - Weighted random sampling with a reservoir by Efraimidis and Spirakis.
- Lecture 2: 8/28/14, Estimating Number of Distinct Elements in a Stream
  - Lecture 2 in in Jelani Nelson's course
  - Chapters 2 and 3 in Amit Chakrabarti's notes.
- Lecture 3: 9/2/14, Estimating F_k norms via AMS sampling
- Lecture 4: 9/4/14, Estimating F_2 norm, Sketching, Johnson-Lindenstrauss Lemma
- Lecture 5: 9/9/14, Estimating F_p norm for 0 < p < 2, Misra-Greis algorithm for frequent items
- Lecture 6: 9/11/14, Count and Count-Min Sketches
- Lecture 7: 9/16/14, Sparse recovery via Count-Sketch (see notes from previous lecture)
- Lecture 8: 9/18/14, \ell_2 sampling and application to near-optimal F_k estimation for k > 2
  - Slide notes by McGregor
  - Section 4 in chapter on signals in the draft book by McGregor-Muthu.
  - Paper on precision sampling by Andoni, Krauthgamer, Onak
  - Paper on \ell_p sampling by Monemizadeh and Woodruff.
  - Paper on near-optimal \ell_p sampling by Jowhari, Saglam, Tardos.
- Lecture 9: 9/23/14, \ell_0 sampling, and priority sampling
  - Paper on near-optimal \ell_p sampling by Jowhari, Saglam, Tardos.
  - Priority sampling paper by Duffield, Lund, Thorup. The arxiv version is here.
- Lecture 10: 9/25/14, Quantiles and selection in multiple passes
  - Quantiles and Equidepth Histograms over Streams, Chapter by Greenwald and Khanna in a book on data stream management.
- Lecture 11: 9/30/14, Continue previous lecture.
- No lecture on 10/2/14, discuss home work problems.
- Lecture 12: 10/07/14, Graph Streams: Connectivity, Cut/Spectral sparsifiers
  - Lecture based on excellent survey by Andrew McGregor
- Lecture 13: 10/09/14, Graph Streams: Spanners, Matchings, Sketching for graphs
  - Lecture based on excellent survey by Andrew McGregor
- Lecture 14: 10/14/14, Finish graph streams. (2+\eps) for k-center clustering in streaming setting
  - McGregor's slide notes
  - Tight results for clustering and summarizing data streams by Guha
- Lecture 15: 10/16/14, Lower bounds for streaming via communication complexity
  - McGregor's slide notes
  - Amit Chakrabarti's notes (Chapters 15, 16, 17)
  - Jelani Nelson's notes
- Lecture 16: 10/21/14, Lower bound on communication complexity of INDEX, lower bounds for graph streaming problems
  - Jelani Nelson's notes for INDEX lower bound via Fano's inequality
  - Amit Chakrabarti's notes (Chapters 17) for lower bounds on graph problems
  - See notes on basics on entropy and Fano's inequality or Cover-Thomas book on Information Theory.
- Lecture 17: 10/23/14, Similarity estimation and Locality sensitive hashing

  - Moses Charikar's paper Similarity estimation techniques from rounding algorithms
  - Min-Wise Independent Permutations by Broder, Charikar, Frieze, Mitzenmacher.
  - Piotr Indyk's paper on small min-wise hash families
  - Chapter and slides on "Finding Similar Items" from Mining Massive Data Sets
- Lecture 18: 10/28/14, Approximate Nearest Neighbor Search via Locality Sensitive Hashing
  - LSH webpage including downloadable code
  - Moses Charikar's paper Similarity estimation techniques from rounding algorithms
  - Helpful slides from a presentation by Aneesh Sharma and Michael Wand
  - Notes on LSH via p-stable distributions from course at UCSD
  - Beyond Locality-Sensitive Hashing by Andoni etal
- Lecture 19: 10/30/14, Approximate matrix multiplication.
  - Notes from Jelani Nelson's course
  - Column sampling technique from paper of Drineas and Kannan
  - Random projection technique from paper of Sarlos
  - Frequent directions technique from paper of Liberty
- Lecture 20: 11/4/14, Singular Value Decomposition
  - Chapter 3 in Foundations of Data Science by Hopcroft and Kannan
- Lecture 21: 11/6/14, Fast Deterministic Low-Rank Approximation
  - Relative Errors for Deterministic Low-Rank Matrix Approximations by Ghashami and Phillips
- Lecture 22: 11/18/14, Subspace embeddings and Fast Least Squares Regression
  - Notes from Jelani Nelsons's course. See related lectures from the same course on Fast Jonhonson-Lindenstrauss Transform etc.
  - Low Rank Approximation and Regression in Input Sparsity Time by Clarkson and Woodruff and Woodruff's talk slides.
  - Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression by Meng and Mahoney
  - OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings by Nelson and Ngyuen
- Lecture 23: 11/20/14, Compressed Sensing
  - Notes from Jelani Nelsons's course. There are several advanced topics covered in his subsequent lectures.

Course Project Information