Computer Science
Dartmouth College

Computer Science 49/149
Data Stream Algorithms
Amit Chakrabarti

Fall 2011

Where: Sudikoff 115

When: 10A hour: TTh 10:00-12:00, X-hr W 15:00-16:05

Who: Instructor: Amit Chakrabarti, Sudikoff 107, Office hours: MF 9:00-10:30
Filler TA: Ranganath Kondapally, Sudikoff 112, Office hours: M 18:00-19:00, Tu 17:00-18:00

What: This course studies algorithms that process massive amounts of data; so massive that they will not fit in a computer's storage. As we shall see, this forces one to rethink even very basic algorithmic problems, such as counting the number of distinct elements, selection, and sorting. We shall study techniques to summarize such large amounts of data into succinct "sketches" that nevertheless retain important and useful information. We shall introduce the necessary mathematical tools along the way. Most techniques we shall study come from research in the last decade or so, although a few date as far back as the 1970s. The techniques we shall study have been applied successfully to web data compression, approximate query processing in databases, network measurement and signal processing. The course itself will focus on the underlying techniques rather than the specific applications.

Prerequisites: An undergraduate-level course in Algorithms (such as our CS 31)
or else: A strong mathematics background and permission of the instructor.

Lecture Plan: The following is a rough outline, and will very likely be changed once the course gets going.

I. Basic Algorithms: Estimating Statistics.

1. Sep 27 Data stream model; Our first algorithm: the MAJORITY problem.
2. Sep 29 The DISTINCT-ELEMENTS problem; Randomized algorithms; Approximation.
3. Oct 4 More modern results on DISTINCT-ELEMENTS.
4. Oct 6 Count Sketch; Count-Min Sketch; MAJORITY revisited; The HEAVY-HITTERS problem.
5. Oct 11 Frequency moments; The AMS algorithm; Improving a Basic Estimator.
6. Oct 13 The amazing AMS second moment (F_2) algorithm; Dimension reduction; Johnson-Lindenstrauss Lemma.
7. Oct 18 Stable distributions; Estimating L_1 distance; Indyk's algorithm and L_p norms, $p \in (0,2)$.
8. Oct 20 The rest of the L_p norms; Sketch using precision sampling.
9. Oct 25 Proof of the Precision Sampling Lemma.
10. Oct 27 The MEDIAN and SELECTION problems; Random order versus adversarial order.

II. More Advanced Algorithms: Graphs, Geometry, Sequences

11. Nov 1 Geometric problems; Coresets; The min-enclosing-ball problem.
12. Nov 3 Metric streams; Clustering; *k*-center, *k*-median, *k*-means.
13. Nov 8 Finding a good *k*-center clustering; Doubling algorithm; Guha's algorithm.
14. Nov 10 Triangle counting; Mininum Spanning Trees.
15. Nov 15 Maximum Matchings.

III. Lower Bounds

16. Nov 17 Communication complexity; The INDEX, DISJOINTNESS and GAP-HAMMING problems.
17. Nov 22 Some reductions from communication to streaming problems.
18. Nov 29 Student presentations #1, #2, #3, #4.
19. Dec 5 Student presentations #5, #6, #7 (from 15:00 to 17:00).

Textbooks: There is no set textbook for this course, but we have some evolving in-house course notes, thanks to the scribing efforts of students from the previous offering of the course. A good fraction of the material we shall study only resides in research papers. There is a good survey by Muthukrishnan, somewhat dated at this point, but still very useful for the basics.

Homework Sets: We will have 4 homework sets in total.

- Homework 1, due Oct 12.
- Homework 2, due Oct 26.
- Homework 3, due Nov 18.
- Homework 4, due Dec 1.

Class Presentation: In lieu of a final exam, students are required to read up on an advanced topic in data stream algorithms

(usually represented by a research paper), and give a *short, 15-minute* presentation on its main ideas. Each presentation should be given by a team of 2 students. Below is a list of suggested papers for such a presentation.

- *The Shifting Sands Algorithm.* McGrgegor, Valiant. (SODA, 2012) [Thomas + Yu-Han]
- *Stable Distributions, Pseudorandom Generators, Embeddings, and Data Stream Computation.* Indyk. (JACM, 2006)
- *A Near-Optimal Algorithm for Estimating the Entropy of a Stream.* Chakrabarti, Cormode, McGregor. (TALG, 2010)
- *Recognizing Well-Parenthesized Expressions in the Streaming Model.* Magniez, Mathieu, Nayak. (STOC, 2010) [Andrew + Huiting]
- *Estimating the Sortedness of a Data Stream.* Gopalan, Jayram, Krauthgamer, Kumar. (SODA, 2007) [Zhiyu + Zhong]
- *CR-Precis: A Deterministic Summary Structure for Update Data Streams.* Ganguly, Majumder. (ESCAPE, 2007) [Yilong + Zhipeng]
- *Estimating PageRank on Graph Streams.* Das Sarma, Gollapudi, Panigrahy. (JACM, 2011) [Cole + Fabio]
- *Space-Efficient Online Computation of Quantile Summaries.* Greenwald, Khanna. (SIGMOD, 2001) [Shahrzad + Shrirang]
- *Sketching Information Divergences.* Guha, Indyk, McGregor. (JML, 2008)
- *Estimating Rarity and Similarity over Data Stream Windows.* Datar, Muthukrishnan. (ESA, 2002) [Chen + Vijay]

We now have a schedule for the class presentations. Please note that all students should be present for all presentations. The presentation is in lieu of a final exam and does count towards your grade, with 20% weight.

Class Presentation: In lieu of a final exam,

TA and Office Hours: Our TA will be Ranganath Kondapally. Office hours for the TA and instructor are announced at the top of this page. Additionally, feel free to drop by and chat with us about the course whenever our doors are open.