Computer Science
Dartmouth College

Computer Science 85/185
Data Stream Algorithms
Amit Chakrabarti

Fall 2009

Where: Sudikoff 214

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

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 25)
or else: A strong mathematics background and permission of the instructor.

Lecture Plan: The following is a rough outline.

I. Basic Algorithms: Estimating Statistics.

1. Sep 24 Data stream model; Our first algorithm: the MAJORITY problem.
2. Sep 29 The DISTINCT-ELEMENTS problem; Randomized algorithms; Approximation.
3. Oct 1 Review of basic probability theory; Hash functions.
4. Oct 6 Count Sketch; Count-Min Sketch; MAJORITY revisited; The HEAVY-HITTERS problem.
5. Oct 8 More modern results on DISTINCT-ELEMENTS.
6. Oct 13 Frequency moments; Norms; The amazing second moment (F_2) algorithm.
7. Oct 15 Delving deeper into the F_2 algorithm; Johnson-Lindenstrauss Lemma.
8. Oct 20 The MEDIAN and SELECTION problems; Random order versus adversarial order.
9. Oct 22 Stable distributions; Estimating L_1 distance; The VARIATIONAL-DISTANCE problem.
10. Oct 27 Slack.

II. More Advanced Algorithms: Graphs, Geometry, Sequences

11. Oct 29 Spanners and Shortest Paths.
12. Nov 3 Matchings.
13. Nov 5 Triangle counting; Mininum Spanning Trees.
14. Nov 10 Clustering; *k*-center, *k*-median, *k*-means.
15. Nov 12 Coresets; Earthmover distance.
16. Nov 17 Slack.

III. Lower Bounds

17. Nov 19 Communication complexity; The INDEX, DISJOINTNESS and GAP-HAMMING problems.
18. Nov 24 Some reductions from communication to streaming problems.
19. Dec 1 Slack.

Textbooks: There is no set textbook for this 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.

Course Blog: Please follow this weblog for updates on all aspects of the course. This is where announcements and helpful little notes will be posted from time to time.

Lecture Notes: Here are the evolving lecture notes for the course. All registered students are required to scribe a minimum of one lecture each, and help prepare notes for that lecture, in the style of the posted notes. It is imperative to learn basic LaTeX for this purpose. We shall regularly update these notes, effectively circulating them to all other students.

Homework Sets: We will have 3 or 4 sets in total. When a new set is posted here, it will be announced on the course blog.

- Homework 1, due Thu Oct 22, at 10:00am.
- Homework 2, due Thu Oct 29, at 5:00pm.
- Homework 3, due Mon Nov 23, at 5:00pm.
- Homework 4, due Sun Dec 6, at 5:00pm.

TA and Office Hours: Our TA will be Chrisil Arackaparambil, who will hold office hours for this course every Monday, 4:00-5:00pm. Additionally, Amit will be available Wednesdays, 3:00-4:00pm (this is the X-hour for the course). Also, feel free to drop by and chat with Amit whenever his door is open.