DS-563 / CD-543: Algorithmic Techniques for Taming Big Data

Boston University

Fall 2021

Course Description: Growing amounts of available data lead to significant challenges in processing them efficiently. In many cases, it is no longer possible to design feasible algorithms that can freely access the entire data set. Instead of that we often have to resort to techniques that allow for reducing the amount of data such as sampling, sketching, dimensionality reduction, and core sets. Apart from these approaches, the course will also explore scenarios in which large data sets are distributed across several machines or even geographical locations and the goal is to design efficient communication protocols or MapReduce algorithms.

The course will include a final project and programming assignments in which we will explore the performance of our techniques when applied to publicly available data sets.

Instructor: Krzysztof Onak (konak@bu.edu)
Office hours: Tue 4–6pm, MCS 138N (or adjacent common space)

Teaching Fellow: Nadya Voronova (voronova@bu.edu)
Office hours: Wed 2–4pm, MCS 141

Course website: `https://onak.pl/ds563`
Piazza: `https://piazza.com/bu/fall2021/ds563cs543/home`

HUB Units: Quantitative Reasoning II (QR2) and Creativity/Innovation (CRI)

Prerequisites

- **Programming:** Fluency with programming and basic data structures is required (commensurate with CS-111, EK-125, or equivalent). Familiarity with C++, Java, or Python is recommended.
- **Algorithms:** Familiarity with basic topics on algorithms and computation complexity (commensurate with CS-330, EC-330, or equivalent) is required. These topics are covered in textbooks such as Cormen, Leiserson, Rivest, Stein "Introduction to Algorithms."
- **Mathematics:** Familiarity with basics of linear algebra (commensurate with CS-132, MA-242, or equivalent) and probability (commensurate with MA-115, CS-237, EK-381, or equivalent) is required. These topics are covered in textbooks such as Strang "Introduction to Linear Algebra," Lay, Lay, McDonald "Linear Algebra and Its Applications," and Pishro-Nik "Introduction to Probability, Statistics, and Random Processes."

To help you assess your preparation for the course, there is a Self–Assessment Questionnaire on the course webpage.

Course Requirements

Apart from active participation, the class will require solving homework problem sets, two experimental programming assignments, and a final project. The overall grade will be based on the following factors:

- class participation: 5%
- homework: 25%
- two programming assignments: 25%
- project proposal: 5%
- final project: 40%

Programming assignments: The course will feature two programming assignments in which students will implement algorithms covered in class and apply them to data sets of their choice. Collaboration here is not allowed (except for discussing high–level ideas), i.e., students are required to implement algorithms and run experiments on their own.

Final project: Possible final projects ideas include but are not limited to

- implementing an algorithm not covered in class and testing its practical performance on real–world data,
- creating an open–source implementation of one of the algorithms with easy to follow documentation,
- developing a new algorithm with good theoretical or practical guarantees.

The outcome of a project will be a short technical report, describing obtained results and conclusions. As opposed to programming assignments, students are allowed to work in teams of 2 or 3. A list of potential projects topics will be provided, but students are encouraged to develop their own ideas. These projects have to be approved by the instructor.

Laptop and Cellphone Policy

Using laptops, cellphones, tablets, and other similar electronic devices is generally not allowed. If you want to use your laptop or tablet for taking notes, you have to email a copy of your notes to me after the class and you are not allowed to use your device for other purposes, such as replying to emails or browsing the web.

Materials

There is no textbook. A good list of resources on many of the topics covered in this class—including books, surveys, lectures notes, and presentations—can be found at

`https://sublinear.info/index.php?title=Resources`

Very Tentative Schedule

The course will consists of 27 lectures (with two lecture dedicated to final project presentations and discussions), and will cover the following topics that cover sampling, sketching, dimensionality reduction techniques, and modern distributed parallel computation.

Lecture	Date	Topics
Section 1: Data projections		
Lecture 1	Sep 2	Course overview. Frequency estimation (CountMin sketch).
Lecture 2	Sep 7	Approximate counting. Estimation of data frequency moments.
Lecture 3	Sep 9	Estimation of data frequency moments.
Lecture 4	Sep 14	Applications to distributed monitoring. Adversarially robust streaming algorithms.
Lecture 5 Lecture 6	Sep 16 Sep 21	Compressed graph representations with applications (graph sketches).
Lecture 7	Sep 23	Data dimensionality reduction (Johnson–Lindenstrauss Lemma).
Lecture 8	Sep 28	Data dimensionality reduction for clustering.
Lecture 9 Lecture 10	Sep 30 Oct 5	Finding similar data points (nearest–neighbor search).
Section 2: Selection of representative subsets		
Lecture 11	Oct 7	Simple geometric problems. Clustering via core sets.
Lecture 12	Oct 14	Clustering via core sets.
Lecture 13	Oct 19	Diversity maximization via core sets.
Section 3: Sampling from probability distributions		
Lecture 14	Oct 21	Estimation of distributions and their properties.
Lecture 15	Oct 26	Verification of a distribution's uniformity.
Lecture 16	Oct 28	Verification of other properties. Access methods beyond sampling.
Section 4: Querying and sampling subsets of data sets		
Lecture 17	Nov 2	Estimation of data parameters and approximate verification of properties.
Lecture 18	Nov 4	Efficient local sparse graph exploration techniques. Estimating graph parameters.
Section 5: Distributed computation		
Lecture 19	Nov 9	MapReduce and the Massively Parallel Computation model. Sample MPC algorithms.
Lecture 20	Nov 11	Clustering on MPC.
Lecture 21	Nov 16	Graph algorithms on MPC.
Lecture 22	Nov 18	Limitations of distributed algorithms.
Section 6: Closing lectures		
Lecture 23 Lecture 24	Nov 23 Nov 30	Efficient sparse least–square regression.
Lecture 25	Dec 2	Overview of additional topics not covered in detail.
Lecture 26 Lecture 27	Dec 7 Dec 9	Project presentations and discussions.