Sublinear Algorithms for Big Data
Fall 2020

Instructor: Jasper Lee, CIT 423, jasperchlee@brown.edu

Time and Location: Tuesdays and Thursdays, 14:30–15:50 at 85 Waterman, Room 130

Collaboration Hours: Tentatively, Mondays 13:00-15:00 virtually. Details on Piazza.

Website: http://cs.brown.edu/courses/csci1951-w

Piazza: http://piazza.com/brown/fall2020/csci1951w/home

Textbook: The course will not be following any particular textbooks. However, the following texts may be helpful for understanding course material (though possibly far too advanced for our scope) and as related reading:

- *Introduction to Property Testing* by Oded Goldreich, published by the Cambridge University Press and a draft pdf available on his webpage

Links to related courses, other relevant lecture notes and further reading will be posted to the course webpage.

Special Measures During the COVID-19 Pandemic

Lectures will be in-person, with at most 19 students physically attending at the same time. We will be planning for that once the enrolment settles down.

To accommodate students situated outside Providence and possibly in different time zones, or students who otherwise cannot attend lectures, all (in-person) lectures will be recorded and broadcast live on Zoom, with whiteboard writing replaced by writing on the instructor’s iPad as part of the recording. Details to follow.

International students are welcome to document their in-person lecture attendance for immigration purposes, in any way that does not intrude on the privacy of other students, e.g. taking a photo with the instructor and the virtual whiteboard after class.

If you are feeling unwell, please do not come to class. We take the COVID-19 pandemic seriously, given its empirically-observed infectiousness and potential severity of symptoms. You will be able to catch up by watching the Zoom recordings, and we can always schedule video meetings to help you. If you are symptomatic and still come to class, you will get a C for your course grade.

We all have our parts to play in limiting the spread of the disease.
Description

A huge quantity of data is worth little unless we can extract insights from it. Yet, the large quantities mean that classic algorithms (running in linear, quadratic or even more time) can be infeasible in practice. We must instead turn to new algorithmic approaches and paradigms, which allow us to answer valuable questions about our data in runtime that is still feasible even when the data set is Facebook-sized.

Surprisingly, to answer many computational and statistical questions, sometimes there is no need to read/store every piece of data! This course focuses on this exciting “sublinear” algorithmic regime. We will study practical algorithms, making clever use of randomness with strong theoretical guarantees, on the following (tentative and non-exhaustive) list of topics:

- Testing and learning information about structures such as lists (e.g. testing approximate sortedness/monotonicity) and massive graphs (e.g. testing approximate connectedness), using very few queries into the structure
- Learning information about a probability distribution from a small number of independent samples, e.g. deciding whether the distribution of users on your online platform has changed substantially since you introduced a UI change
- Computation on high-volume streams of data, whilst only maintaining a small memory buffer, e.g. approximately estimating the number of unique visitors to a webpage over a given time period without storing them one by one

Prerequisites

This is an advanced undergraduate level (1000 level) to graduate level (2000 level) course. The key prerequisites are

1. Mathematical maturity, as we will prove most if not all of the covered results both in class and on homeworks
2. Familiarity with basic probability (you should be comfortable applying Markov’s and Chebyshev’s inequalities, and understand how a random walk can be characterised by a matrix)
3. Familiarity with basic analysis of algorithms

Concrete course prerequisites are

- CS22 or equivalent, and
- CS145 or APMA1650/1655 or equivalent, and
- CS157 or CS155

CS155 is recommended but not required: all required materials not taught in CS145 or equivalent will be covered in this class.
Learning Goals

The primary goals of the course are to 1) introduce students to the models and problem formulations for sublinear algorithms and related topics, 2) give basic tools for students to read, understand and implement existing results, and 3) enable students to communicate these results effectively.

To achieve the first goal, this course surveys 3 main areas in sublinear algorithms: 1) combinatorial property testing and learning, 2) distributional property testing and learning, and 3) streaming algorithms. There are also lectures tentatively planned to cover newer/related frameworks, such as sampling large random structures sublinearly/locally, sampling correction, and robust statistics.

In surveying these algorithmic results, relevant design and analysis tools are introduced, which are applicable in general for understanding papers in the area.

The homeworks, lecture note scribing (as described below) and final project all complement the lectures to help students master the key concepts and tools. Furthermore, they are the primary venues for developing the students’ communication skills for these sophisticated and technical results.

Lastly, for students intending to go onto graduate studies in theoretical computer science, the mathematical and paper-reading techniques from this course will hopefully form part of their research toolkit.

Course Outline

The course material is evolving. Depending on the technical background of the enrolled students, we may explore material based on knowledge outside of the prerequisites stated above.

**Introduction to Probability Tools** (Week 1)

**Sublinear algorithms on combinatorial structures** (Weeks 2–5)

**Sublinear algorithms on probability distributions** (Weeks 6–9)

**Streaming Algorithms** (Weeks 10–13)

**Miscellaneous Topics** (Week 14)

Grading

Grade components may be curved by the instructor before a final grade is calculated.

**For 1000-level credit:**

Homeworks 50%  
Participation+Scribing 20%  
Final Project 30%  
Extra Credit from 2000-level work Maximum of 10%

**For 2000-level credit:**

Homeworks 50%  
Participation+Scribing 20%  
Final Project 30%
Time Requirements

There will be 2 lectures a week, 1.5 hours each, for 14 weeks, totalling 42 hours of in-class time. Depending on the class size, each student is expected to spend 1-3 hours (over the entire semester) outside of class polishing their scribe notes. Each homework should take roughly between 8 to 20 hours including any extra reading needed, with 10 hours a week being typical (totalling 140-150 hours). The final project is expected to take 10-40 hours depending on the scope of the project. The total time requirement is expected to be at least $42 + 1 + 140 + 10 = 193$ hours.

Homework Assignments

Homeworks will be released weekly, with instructions on submission and deadlines. With the exception of Homework 0 which should be submitted but will not be formally graded, all homework submissions contribute to the homework component of the final grade as detailed in the previous section. Students will pair up in writing their homework submissions, again except only for Homework 0.

For 2000-level credit, there will be extra problems on the homeworks. Please refer to the instructions on each homework. Students registered for the 1000-level credit are welcome to complete the 2000-level problems for extra credit, as described in the Grading section.

Assignments should be typeset neatly in \LaTeX or written in clearly legible handwriting. Unreadable submissions may not be graded.

Collaboration Policy and Academic Honesty

Collaboration on homework sets (with others in addition to your partner) is not only permitted, but also encouraged. To maximise learning, we suggest you first try solving the problems on your own, before exchanging and brainstorming ideas with your classmates. You must however write out all solutions on your own. On each homework, please state who you discussed the problems with.

You are also allowed to consult other sources, for example resources on the Internet, for alternative explanations of concepts and results covered in the course, as well as related reading materials. Inevitably, some of you will (intentionally or unintentionally) stumble upon solutions to homework problems. In order to have an enforceable collaboration policy, the bottom line is that you must write out all the solutions in your own words, demonstrating that you at least understand the solution you have written. Under this policy, whilst you are allowed to just search for problem solutions online (if the problem is standard enough), and rewrite the solutions in your own words, it is of course heavily discouraged for the sake of your education.

If you do happen to base your answer (even if only partially) on outside sources, please also cite them. This is for the instructor’s (and then the class’) benefit, to see what online sources may be useful for the course.

Failure to comply with this lax collaboration policy, that is if you submit something that you plagiarised without demonstrating any understanding, results in a C in your course grade.
Late Policy

Except with the prior approval of the instructor, any submission that is late for no more than 48 hours will get a 20% reduction on that homework’s grades, and any submission more than 48 hours late will not be graded.

Permission for late submissions must be requested at least 24 hours in advance, and will only be granted in exceptional/extenuating circumstances or for religious observances. Contact the instructor directly for such requests.

Participation and Scribing

Participation and scribing constitutes 20% of the final grade. In addition to participating in class by joining in discussions and asking/answering questions, each student is expected to act as the lecture note scribe for at least 1 lecture. As with the homework assignments, scribing will also be done in pairs. Students may be asked to scribe more than once, depending on the enrolment numbers, with priority going to those registered for 2000-level credit.

Lecture notes should be typeset using a supplied \LaTeX template, and submitted to the instructor in a timely manner. Late submissions may have a 5% deduction (amongst the 20%), depending on the pacing of the course at the time and at the instructor’s discretion, taking into account the reasons for lateness. Each submission will be awarded 10% for any non-trivial attempt that includes all components of the lecture. The other 10% is awarded for clear and readable notes. Extra credit may be given for notes that are particularly well-written and/or include materials supplementing the lecture, such as completing proofs of assumed results.

The staff will try to schedule such that students residing in different time zones are responsible for scribing on Thursdays, allowing them more time to view the lecture recording and write up their notes.

The instructor will give timely feedback on the submission, such that the students can revise and improve their notes for a higher grade, with no penalty.

Students should not see scribing as a burden, but an educational opportunity in line with the learning goals of the course. Scribing lecture notes is good practice for explaining technical ideas, with a mechanism to get feedback on writing, as well as a way to consolidate understanding before putting such understanding into words.

Final Project

As stated in an earlier section, the primary learning goals of the course include enabling students to read, understand, and communicate results in sublinear algorithms and related areas. To that end, the final project is to produce a written report/survey on one or more results in the area, allowing students both to 1) learn about new models/results and to 2) demonstrate their ability to understand and communicate these ideas.

For 1000-level credit, students are expected to read at least one sublinear algorithmic result and 1) implement it (if there is sufficient scope and complexity in the algorithm), reporting on the empirical findings, and/or 2) write up the analysis of the algorithm in their own words.
For 2000-level credit, students are expected to read multiple research papers or a single sufficiently complex result, and produce a coherent survey on the analyses and relationship between the different results. In general, a 2000-level credit final project will have a format similar to that of a 1000-level credit one, but based on more difficult material.

Projects are graded both on their scope and on the quality of the write-up. A high quality report/survey will have 1) clear writing style, 2) well-organised logical flow to explain technical results and 3) a distillation of the intuition behind the algorithm design and analysis.

The topic choices are subject to the instructor’s approval. Students are expected to complete their projects individually, unless the students gain prior approval from the instructor for a project with significantly larger scope (e.g. a writeup of a proof of the PCP theorem). Further details will be communicated to the class during the second half of the course.

Tentatively, the final project is due on 11 Dec 2020 at 23:59 (Providence time).

**Accommodations (SEAS and Religious Observance)**

If you have any disabilities (of any form and type), or any existing or new medical conditions that could affect your learning and ability to complete the coursework, please contact SEAS or a dean to discuss. Let us know as soon as possible if you require any accommodations, attaching a relevant Dean’s note or SEAS email. The staff will support you as best as we can.

Students with religious observance conflicting with the course schedule should endeavour to inform the instructor within the first 4 weeks of the semester, and no later than 1 week in advance, in order for us to make suitable arrangements.

**Mental Health**

Being a student can be very stressful. If you feel you are under too much pressure or there are psychological issues that are keeping you from performing well at Brown, we encourage you to contact Brown’s Counseling and Psychological Services. They provide confidential counselling.

CAPS: [https://www.brown.edu/campus-life/support/counseling-and-psychological-services/](https://www.brown.edu/campus-life/support/counseling-and-psychological-services/)

**Coping with Unforeseen or Difficult Circumstances**

If there are events that are upsetting to you, whether political, family-related, weather-related, etc., that affect your ability to do well in class, we are happy to take them into account with respect to our late policy. Please feel free to talk to the instructor about this. Additionally, Student Support Services Deans ([https://www.brown.edu/offices/student-support/student-support-services](https://www.brown.edu/offices/student-support/student-support-services)) can be a helpful resource for discussing current concerns and academic and personal plans. They are available for both same-day consults and scheduled appointments.
Diversity Statement

Our intent is that this course provide a welcoming environment for all students who satisfy the prerequisites. Our staff have undergone training in diversity and inclusion; all members of the CS community, including faculty and staff, are expected to treat one another in a professional manner. If you feel you have not been treated in a professional manner by any of the course staff, you are invited to talk with Jasper Lee (the instructor), Prof. Ugur Cetintemel (the department chair), Prof. Tom Doeppner (Director of Undergraduate Studies), Prof. Kathi Fisler (Associate Director of Undergraduate Studies) or Ms Laura Dobler (the department’s coordinator for diversity and inclusion initiatives). We take seriously all complaints about unprofessional behaviour.

Brown welcomes students from around the country and the world, and their unique perspectives enrich our learning community. To support students whose primary language is not English, an array of English support services are available on campus including language and culture workshops and individual appointments. For more information, contact english-support@brown.edu or (401) 863-5672.

Contacting the Staff

For any sensitive matters, you should feel free to contact the instructor or departmental staff directly, ignoring the instructions below.

- For technical questions about course content: post on Piazza, or better yet, go to hours if you are in Providence (health and safety allowing)
- For non-sensitive SEAS accommodations: email the instructor
- For requests for deadline extensions: email the instructor
- For grading concerns: email the instructor
- For concerns about the final grade: email the instructor
- For diversity and related concerns: email the instructor, or Professors Centintemel, Doeppner or Fisler, or Laura Dobler