Intro to Engineering Stochastic Processes II
"Stochastics at Scale"

ORIE 4520: Syllabus

Fall 2015

Essential Course information:

Lectures and Recitations
Class time: MWF 1:25-2:15pm
Class location: Phillips 403
Recitation time/location: TBD

Instructor
Prof. Siddhartha Banerjee
Office: 229 Rhodes Hall
E-mail: sbanerjee@cornell.edu
Office hours: MW 2:30pm-3:30pm (immediately after class)

Teaching Assistant
Anna Srapionyan
E-mail: as3348@cornell.edu

Course description:

The course will be based on studying a collection of stochastic algorithms and models, that together illustrate the following idea:

> *Size is a critical consideration in the design of useful stochastic models and algorithms.*

A more formal way to express this is through the notion of *scaling* – given a model/algorithm for some system, how does it behave when we grow some aspect of the system. This course will try to build intuition behind the importance of scaling by presenting examples where understanding scaling is crucial for good system design.

The course will comprise of three units:

1. **Probability tools and randomized algorithms:** I will introduce some tools (concentration inequalities, martingales, balls and bins) which we will use throughout the course, and illustrate their use in the design of randomized algorithms for various applications. These algorithms are often much simpler and more robust than their deterministic counterparts.

2. **'Big-data' scaling:** Applications involving large amounts of data often impose additional restrictions on algorithms – in particular, on the storage availability. I'll present some modern techniques (various hashing methods, sketching, sampling techniques, etc.) which make clever use of randomness to deal with these issues.

3. **Threshold phenomena:** I'll discuss some intriguing stochastic models where small changes lead to dramatically different behavior as the system grows larger. This can be exploited to design simple yet efficient algorithms for various applications – load balancing, ensuring network connectivity, controlling epidemics, etc.

Learning outcomes

After taking this course, you should be equipped with the following:

1. The ability to think about any problem you work on through the lens of scaling behavior.
2. A set of technical tools to deal with such problems – these include concentration inequalities, martingales, random walks, and fluid approximations.
3. A collection of useful probabilistic models and algorithms – balls and bins, various hashing techniques, graph sketches, branching processes, random graphs, epidemic models, etc. These are excellent modeling tools for many diverse settings.

Course Prerequisites:

I will assume knowledge of basic probability (at the level of ORIE 3500): in particular, you should be comfortable with random variables, conditional probability and expectation, common probability distributions and their properties (binomial, geometric, exponential, Poisson, Gaussian).

The latter part of the course (after the prelim) will require knowledge of stochastic processes, in particular, Markov chains (at the level of ORIE 3510). There will be a recitation session covering the essentials, and students may be able to manage without the required background. Prior exposure to algorithms and graph theory will also be useful, but not essential. Send me a mail if you are concerned about having the appropriate prerequisites.

Class websites:

I will be using Blackboard for announcements and lecture materials. You should be enrolled automatically, but if not, visit https://blackboard.cornell.edu/ and search for ORIE 4520.

Textbooks:

There is no required textbook for the course. I will cover different topics from different sources, and will periodically post notes and links to the relevant material. Some recommended textbooks:

- There are two good textbooks for the first unit of the course:
  - *Randomized Algorithms* by Rajeev Motwani and Prabhakar Raghavan
  - *Probability and Computing* by Michael Mitzenmacher and Eli Upfal

  The former is a classic in the field, but is older and more technical; the latter covers fewer topics, but is easier to read. The Cornell library has access to online versions of both books.
- A good introduction for the material in the second unit is *Mining of Massive Datasets* by Jure Leskovec, Anand Rajaraman and Jeff Ullman. An online version is freely available on the book's webpage (`http://www.mmds.org/`).
- A good introduction for some of the material in the third unit is *Networks, Crowds and Markets* (Sections V, VI) by David Easley and Jon Kleinberg. For a (much more) technical treatment, see *Epidemics and Rumours in Complex Networks* by Moez Draief and Laurent Massoulié.

Homework:

The course will have 8 homeworks – these will be weekly until the prelim, and biweekly after that. Homeworks will be due on Friday at noon in the homework mailbox (Rhodes 2nd floor lobby).

You are encouraged to discuss the problems with others, as well as consult online sources. However, *you must write your solutions independently and individually.* Submitting copied solutions will be considered a violation of the Cornell code of academic integrity.

Late homework will not be graded. If you can not submit in person by 12pm, then mail me a scan/clear picture of *all pages* of the homework by 12 pm, and then submit the physical copy as soon as possible for grading. Any work not visible in the scan will not be graded.

Exams:

The prelim will be a 90 min in-class exam, held during recitation hours – tentatively, this will be during the week of 19th to 23rd October. The exact dates and location will be confirmed soon. The exam will be closed book and closed notes. The course has no final exam.

Project:

In place of the final exam, there will be a project toward the end of the semester. A typical project will involve you choosing a topic related to things we discuss in the course, reading up on research on the topic, and then simulating the system to try and understand it better. Alternately (or additionally) you could do original research on the topic. The topic can be proposed by you or chosen from a list of suggestions that I'll put up sometime later in the semester; it can be either theoretical or more practically oriented.

You need to submit a one-page proposal to me on **Friday, October 23, 2015**. For the final report of the project, I would like an interactive document (ideally an iPython notebook, or using similar technology such as R markdown) which summarizes the problem and results, and has interactive demonstrations of the results. We will discuss this in more detail in class. The final report is due on **Friday, December 4, 2015** (the last class day).

Grading:

Your grade will be based on homeworks (45%), the prelim (25%) and the project (25%+5%). The first six homeworks will be worth 5% each, and the last two will be worth 10% each – the total will be rounded down to 45% (effectively, this means you can drop one of the first six homeworks). The project report is worth 25% of the grade. Finally, the remaining 5% is for submitting the proposal on time **and** filling the course evaluation form during the evaluation period.

Class schedule:

A class schedule will be posted on Blackboard, with details of homework and project deadlines and the prelim, by 1st September. I will need to swap recitation and class meeting times in early November due to some travel – this will be on the schedule, and you will be notified in advance.

Academic integrity:

You are expected to abide by the Cornell University Code of Academic Integrity. Any work submitted by you in this course for academic credit should be your own. The complete code is available at `http://cuinfo.cornell.edu/Academic/AIC.html`.