1 CSCI 8980: Algorithmic Techniques for Big Data Analysis

When: 6:30 pm-9:00 pm, Thur
Where: MechE 221
Instructor: Barna Saha, AT&T Shannon Laboratory
Contact:barna@research.att.com
Office Hours: Friday, 1pm to 5pm by Appointment

Course Overview

For decades researchers across different disciplines of computer science have envisioned the need of techniques to handle data-intensive computing. With the boom of internet and the explosion of data in every socio-economical aspect, once what was a futuristic research, has now transformed itself into a dire requirement. Big Data comes with immense opportunity, but turning this seriously high volume, high velocity, structured or unstructured, heterogeneous, often noisy and high-dimensional data into something one can use is a huge challenge. This course aims at timely dissemination of foundational algorithmic developments for big data analysis and exposing students to cutting edge research in this area. The course will involve deep theoretical analysis with the goal of developing practical algorithms with variety of applications.

We will explore trade-offs among space, time and accuracy for algorithm design. The course will cover different sampling methodologies, streaming algorithms where only a small fraction of data can be stored, semi-streaming model that allows a few sequential access to disk and some parallel algorithms, specifically, the map-reduce paradigm. We will study the algorithmic and complexity aspects of these frameworks. A primary focus of this course will be on designing scalable, sub-quadratic, often near-linear or even sub-linear algorithms. We will explore property testing methodology that allows in sub-linear time to test whether a data set has certain property. We will go through the recent progress in developing fast algorithms for basic graph problems such as max-flow min-cut, matching etc., sparse transformations such as sparse fast Fourier transform, and hashing methodologies involving min-hash and locality sensitive hashing. The other topics will include dimensionality reduction techniques to handle high-dimensional data comprising of random projection method, Johnson Lindenstrauss Lemma, metric embedding and graph sparsifiers. Most of the algorithms that we will study in this course will crucially use randomization and will give an answer that is a good approximation of the optimal solution.

Prerequisites

Students should have basic knowledge of algorithms: running time analysis, graphs algorithms, and must be familiar with discrete probability. Undergraduates are welcome to attend if they satisfy the requirements.

Syllabus

There will be no required text book for this course, instead we will use assorted materials from the web. A tentative list of topics is as follows:

- Streaming: Sampling and Sketching
- Dimensionality Reduction

- External Memory and Semi-streaming Algorithms
- Map-Reduce Framework and Extensions
- Near Linear Time Algorithm Design
- Property Testing
- Metric Embedding
- Sparse Transformation
- Crowdsourcing

The topics will evolve as the course progresses.

Workload

There will be no exams. Each student will be required to write a survey paper and do a project possibly in a group of two or three. A list of topics for the survey papers will be provided. Projects can be either more applied/ implementation based or theoretical depending on interest. In addition, students will be asked to scribe one or two lecture notes and participate in a course blog.