CMSC 454
Title: Introduction to Algorithms for Massive Data Sets
Credits: 3

Description: This course will deal with fundamental methods for processing a high volume of data; these methods include stream processing, locally sensitive hashing, web search methods, page rank computation, network and link analysis, dynamic graph algorithms and methods to handle high dimensional data/dimensionality reduction.

Grading: Homeworks, Midterms, Project, Exam

Prerequisites: Minimum grade of C- in CMSC 320, 330 and 351. Students are expected to have taken STAT 400/STAT 410.

Learning Goals: Modern algorithm design is complex. Trying to solve problems at scale means that the algorithms deployed need to be extremely efficient, and often randomization plays a central role in the design of such schemes with extensive use of hashing. This course will train students to think about data at scale, and to develop a deeper appreciation for what it takes to build a search engine or any system that has to analyze vast amounts of data.

Detailed Topics:

- Importance of sampling, streaming models, external memory
- Streaming algorithms (estimating statistics, number of distinct elements, frequency moments, histogram construction, reservoir sampling)
- Near duplicate detection: Min-hash, Jaccard distances, Locally Sensitive Hashing
- PageRank, Link Analysis, Social Network mining
- Dynamic graph algorithms
- Association rules and frequent itemsets
- Clustering
- Recommendation Systems
- Dimensionality reduction (PCA)

Readings: Mining of Massive Data Sets by Leskovec, Rajaraman and Ullman. Available online for free.