PhD Thesis Defense

"Semi-Streaming Approximation of Centrality Indices in Massive Graphs"

Benjamin Priest

Date
11am-1pm – Rm. 100 (Spanos Auditorium), Cummings Hall

Thesis Committee
George Cybenko, Ph.D. (Chair)
Eugene Santos, Ph.D.
Amit Chakrabarti, Ph.D.
Roger Pearce, Ph.D.

Abstract

The identification of important vertices or edges is a ubiquitous problem in the analysis of graphs. There are many application-dependent measures of importance, such as centrality indices (e.g. degree centrality, closeness centrality, betweenness centrality, and eigencentrality) and local triangle counts, among others. Traditional computational models assume that the entire input fits into working memory, which is impractical for very large graphs. The distributed memory model and streaming model are popular solutions to this problem of scale. In the distributed memory model a collection of processors partition the graph and must optimize communication in addition to execution time. The data stream model assumes only sequential access to the input, which is handled in small chunks. Data stream algorithms use sublinear memory and a small number of passes and seek to optimize update time, query time, and post processing time.

In this dissertation, we consider the application of distributed data stream algorithms to the sublinear approximation of several centrality indices, local triangle counts, and the simulation of random walks. We pay special attention to the recovery of *heavy hitters* - the largest elements relative to the given index.

We present new algorithms providing streaming approximations of degree centrality and a semi-streaming constant-pass approximation of closeness centrality. We achieve our results by way of counting sketches and sampling sketches. We also develop hybrid pseudo-asynchronous communication protocols tailored to managing communication on distributed graph algorithms with asymmetric computational loads. We use this protocol as a framework to develop distributed streaming algorithms utilizing cardinality sketches. We present new algorithms for estimating local neighborhood sizes, as well as vertex- and edge-local triangle counts, with special attention paid to heavy hitter recovery. We also utilize reservoir sampling and ℓ_p sampling sketches to optimize the semi-streaming simulation of many random walks in parallel in distributed memory. We use these algorithms to approximate K-path centrality as a proxy for recovery the top-k betweenness centrality elements.