SAND2011-5757 C

SAND2011-5757C

Challenges in Streaming Graph Analysis

The volume of streaming data for cyber analysis is increasing at a rate much greater than anybody's ability to hire human analysts. As a preliminary step to automating significant portions of their workload, we consider the problem of modeling cyber data. Since the latter tends to be relational in nature, graphs are a natural abstraction. This motivates future research into efficient algorithms for fundamental graph problems in a high-volume, streaming environment. Algorithms designed using current theoretical models for streaming graph algorithms are not directly suitable for operations. In this talk, we propose a new streaming model that can be implemented on a parallel system with extremely simple topology. Our model assumes an infinite stream which must be analyzed using finite resources. We illustrate the associated challenges by giving an algorithm for maintaining and querying the connected components of a graph in which edges may be expired periodically.

Demetrescu, Finocchi, and Ribichini (SODA 2006, ACM TALG 2009) gave an elegant algorithm for computing the connected components of an n-node graph in their W-stream model. In this model, edges arrive one by one in a finite stream. The algorithm has access to finite space s generally much smaller than n, the number of nodes in the graph. The algorithm can emit a new intermediate stream, and reprocess that stream after the current finite stream ends. After O(n lg n/s) such passes, the algorithm has computed the connected components of the graph.

In this work, we consider maintaining the connected components of a graph that arrives in an infinite stream. The algorithm supports queries about connected components and supports aging operations, where all edges at least t seconds old leave simultaneously. We propose a new streaming model called X-stream, essentially a parallel processing model for streaming applications.

We describe an algorithm for maintaining connected components with aging in the X-stream model based on the Demestrescu et al W-stream (finite) connected components algorithm. The Destrescu et al algorithm will stream through smoothly end-to-end on the X-Stream model. However, once the system must handle an infinite stream and must store all edges, there are significant storage management complications.

This is a work in progress, as the performance proofs are not finished. However, we believe the algorithm has the following properties: (1) It correctly answers connected component queries, except for a period of stabilization after an aging command. (2) It does not drop any edges provided there is space in the overall system, even during aging. (3) It effectively uses the total system space by storing Omega(ps) edges on p processors. We will present the main pieces of the algorithm, with arguments for this list of properties.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.