Identifying close-knit communities (or “clusters”) in graphs is an advanced task with a broad range of scientific applications. While theoretical formulations of this operation are either intractable or computationally prohibitive, practical algorithmic heuristics exist to efficiently tackle the problem. However, implementing these heuristics to work for large real-world graphs still remains a significant challenge, owing to a combination of factors that include magnitude of the data, irregular data access patterns, and compute-intensive operations to better the approximation. In this paper, we propose a novel MapReduce-based algorithm for a well known serial graph clustering heuristic called Shingling [3]; and ii) a novel application of the method to cluster biological data, proteins and domains. Operating on an input graph that is simply represented as a list of edges, our algorithm uses a combination of shuffling and sorting operations, and pipelined MapReduce stages to implement the various phases of the algorithm. Preliminary results show linear scaling of the time-dominant phase up to 64 cores on a Hadoop cluster at National Energy Research Scientific Computing Center (NERSC) as our experimental platform. The cluster has 78 nodes with a total of 624 processors and 24 GB DDR3 1333 MHz RAM. These nodes run Cloudera’s distribution for processing on large clusters.

EXPERIMENTAL RESULTS

We presented a novel MapReduce algorithm called pClust-mr that can be used to identify dense subgraphs from bipartite graphs. We compare it with a traditional application of this method to a problem pertaining to clustering protein sequences based on domain knowledge. Although the research is at an early stage, preliminary results indicate linear scaling behavior of the most dominant phase (Shingling Phase I). More importantly, the results demonstrate the effectiveness of the MapReduce framework for solving this irregular graph problem, and the potential of the proposed method to scale to much larger input sizes. Several studies and extensions have been planned as part of future work. Importantly, we plan to test out our implementation for larger real-world graphs, optimize the performance of the remaining phases, and analyze the qualitative and scientific merits of protein-domain clustering.

REFERENCES


ACKNOWLEDGMENT AND CONTACTS

This research was supported in parts by DOE award DE-SC-0006516 and NSF grant IIS 0916463.

Ananth Kalyanaraman
School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA, USA
Email: ananth@eecs.wsu.edu

Inna Rytsareva
School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA, USA
Email: inna.rytsareva@email.wsu.edu