University of Pittsburgh | School of Computing and Information

Dissertation Defense

Doctor of Philosophy in Computer Science

"**Distributed Sparse Computing and Communication for Big Graph Analytics and Deep Learning**" by **Mohammad Hasanzadeh Mofrad**

Date: October 30, 2020
Time: 09:30 – 12:00
Place: https://pitt.zoom.us/j/99334483280

Committee:

- Rami Melhem, Professor, Department of Computer Science, School of Computing and Information
- Alexandros Labrinidis, Professor, Department of Computer Science, School of Computing and Information
- John Lange, Associate Professor, Department of Computer Science, School of Computing and Information
- Dr. Balaji Palanisamy, Associate Professor, Informatics and Networked Systems Department, School of Computing and Information
- Dr. Mohammad Hammoud, Teaching Professor, Department of Computer Science, Carnegie Mellon University in Qatar

Abstract:
Sparsity can be found in the underlying structure of many real-world computationally expensive problems including big graph analytics and large scale sparse deep neural networks. In addition, if gracefully investigated, many of these problems contain a broad substratum of parallelism suitable for parallel and distributed executions of sparse computation. However, usually, dense computation is preferred to its sparse alternative as sparse computation is not only hard to parallelize due to the irregular nature of the sparse data, but also complicated to implement in terms of rewriting a dense algorithm into a sparse one. Hence, foolproof sparse computation requires customized data structures to encode the sparsity of the sparse data and new algorithms to mask the complexity of the sparse computation. However, by carefully exploiting the sparse data structures and algorithms, sparse computation can reduce memory consumption, communication volume, and processing power and thus undoubtedly move the scalability boundaries compared to its dense equivalent.

In this dissertation, I explain how to use parallel and distributed computing techniques in the presence of sparsity to solve large scientific problems including graph analytics and deep learning. To meet this end goal, I leverage the duality between graph theory and sparse linear algebra primitives, and thus solve graph analytics and deep learning problems with the sparse matrix operations. My contributions are fourfold: (1) design and implementation of a new distributed compressed sparse matrix data structure that reduces both computation and communication volumes and is suitable for sparse matrix-vector and sparse matrix-matrix operations, (2) introducing the new MPI*X parallelism model that deems threads as basic units of computing and communication, (3) optimizing sparse matrix-matrix multiplication by employing different hashing techniques, and (4) proposing the new data-then-model parallelism that mitigates the effect of stragglers in sparse deep learning by combining data and model parallelisms. Altogether, these contributions provide a set of data structures and algorithms to accelerate and scale the sparse computing and communication.