Parallel Algorithms and Generalized Frameworks for Learning Large-Scale Bayesian Networks

Ankit Srivastava and Srinivas Aluru (advisor)
School of Computational Science and Engineering, Georgia Institute of Technology

OAC-1828187, OAC-1854828

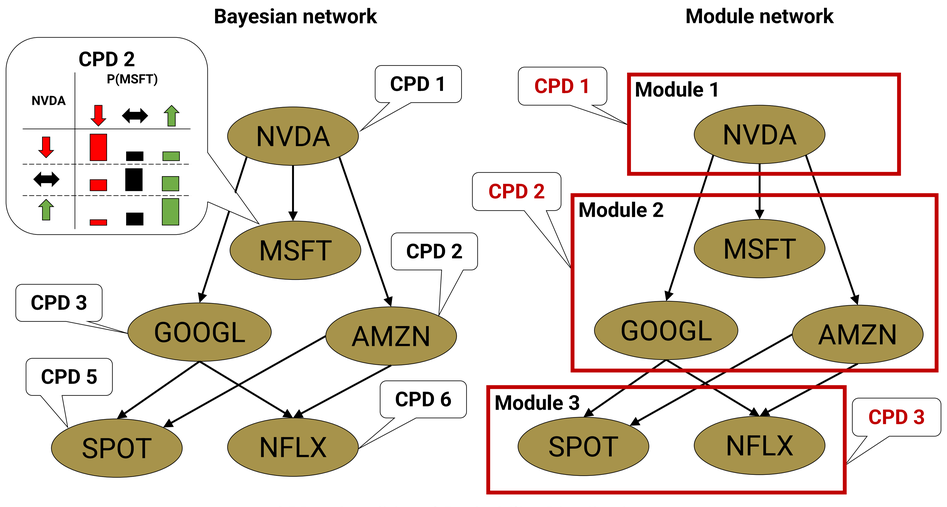
Bayesian networks (BNs) are an important subclass of graphical machine learning (ML) models that enable probabilistic reasoning about interactions between variables of interest. Their interpretability makes them an ideal model for making high-stakes decisions in fields where explainability is desirable. However, learning BNs with even few thousand variables using existing software libraries requires an infeasible amount of time. This has prevented BNs from becoming a viable alternative to other ML models. To address this, we have developed scalable high-performance libraries for learning large-scale BNs. In this poster, we present our work on parallelizing a variety of popular BN learning algorithms, including a method for constructing parameter-sharing specialization of BNs – module networks (MoNets). Our experiments show that the optimized open-source implementations of our parallel algorithms reduce the time required for learning networks with tens of thousands of variables from multiple months to a few hours by efficiently utilizing thousands of cores.

Introduction

- Current need for explainability in high-stakes decisions made by ML models – GDPR, Equal Credit Opportunity Act, etc.
- BNs enable probabilistic reasoning about links between the variables of interest – facilitates interpretable decision-making
- Learning BNs is computationally intensive; existing software libraries support limited parallelism, e.g., *bnlearn*, *Tetrad*, *pcalg*
- High-performance library for expeditious learning of large-scale BNs is required to ascertain their viability as ML models for making interpretable decisions with high human-impact

Background

- Structure of **BNs** represents dependence graph for a set of random variables – direct interactions between variables
- **MoNets** additionally identify groups of variables that organize together for emergent behaviors – indirect interactions
- e.g., BN and MoNet for stock prices of cloud-related companies



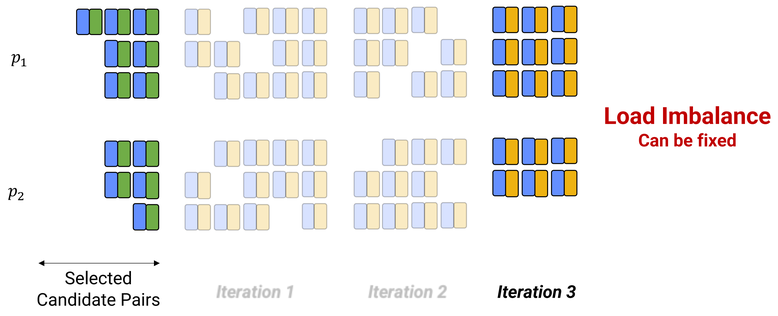
CPD = Conditional Probability Distribution

Methodology

- Parallelized a variety of popular sequential BN learning algorithms – focused on algorithms that lack parallel solutions
- ***Constraint-based***
  - Developed a parallel framework and used it to propose parallel versions of five popular algorithms in the space – *GS*, *IAMB*, *Inter-IAMB*, *MMPC*, and *SI-HITON*
- ***Score-based***
  - Proposed the first parallel algorithm for learning MoNets using the widely used *Lemon-Tree* method

Key Parallelization Ideas

1. Use of tuples for fine-grained definitions of the total available work for learning networks
2. Distribution of the tuples for balancing the load

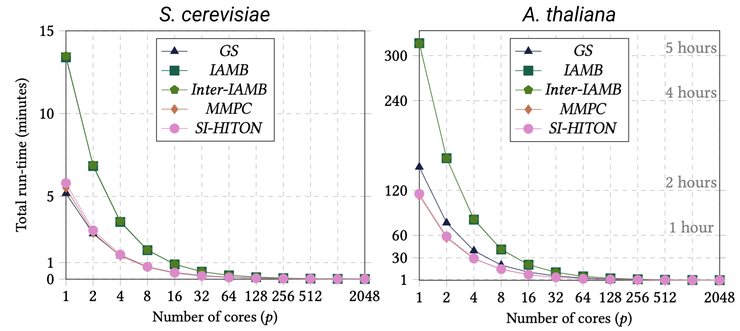


Relevant Publications at SC

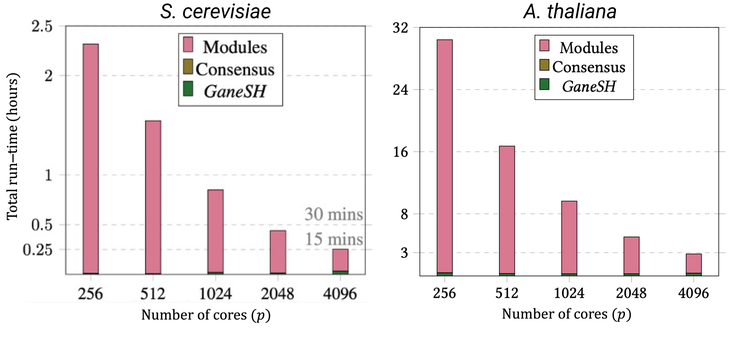
- Srivastava, A., Chockalingam, S., Aluru, M. & Aluru, S. (2021, November). Parallel Construction of Module Networks. *Accepted* in *2021 SC21: International Conference for High Performance Computing, Networking, Storage and Analysis* (***SC***). ACM.
- Srivastava, A., Chockalingam, S., & Aluru, S. (2020, November). A Parallel Framework for Constraint-based Bayesian Network Learning via Markov Blanket Discovery. In *2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (***SC***) (pp. 74-88). IEEE Computer Society.
  - **Selected as benchmark for Reproducibility Challenge, SCC 21**

Results

- Used real gene expression data sets to learn gene-regulatory networks – important application of BNs
  - *Saccharomyces cerevisiae* and *Arabidopsis thaliana*
- ***Constraint-based*** algorithms learn BNs with 18,373 variables in **< 1 minute** using 2048 cores
  - Prior state-of-the-art requires **more than a week**



- ***Score-based*** construction of MoNets with 18,373 variables can be done in **< 3 hours** using 4096 cores
  - Estimated run-time of previous state-of-the art is **> 4 years**



Open-Source Software Packages

- **ParsiMoNe** – Parallel Construction of Module Networks
  https://github.com/asrivast28/ParsiMoNe/releases/tag/v1.0.1
- **ramBLe** – A Parallel Framework for Bayesian Learning
  https://github.com/asrivast28/ramBLe/releases/tag/v2.0.0