**BDCC 19/20**
**Worksheet #4**
**May 26th, 2020**

# Questions about paper:

- A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning

## General questions

1. What is this paper about? Could you summarise its contribution in a paragraph?

2. How does this work differ from others mentioned in the paper?

3. Do the authors present experiments? What is the methodology used? Does it sound correct? Why?

4. What are the main results/findings/conclusions? Are the results useful/relevant? Why?

### Technical questions

1. What is Distributed Data Parallelism (DDP)?

2. In the sentence: “the SBNL workflow partitions the data set into data partitions of reasonable size” (Section IV), what woud be a “reasonable size”? Explore SBNL and find out what is the criterium to choose the data partition size.

3. Discuss the consequences of building a Bayesian network by partitioning the data and building local Bayes nets that later will be combined. Is this affecting the results when compared with the sequential execution? How do you compare the results produced by the parallel implementation with the sequential one? What would be the best model?

4. From a theoretical point of vew, does the probabilistic model generated in parallel approximate the optimal probabilistic function?