Final Homework
50 Points
Due Wednesday May 15th

1) Consider the Parkinsons Telemonitoring Dataset on the UCI Machine learning repository (https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring). This data set was developed with 10 medical centers. Together with a corporation, they developed a telemonitoring device to record speech signals of patients for the prediction of clinical Parkinson's disease symptom scores on a UPDRS scale. This data is designed for supervised learning, however, we are going to "pretend" that there are no labels/ no response.

a) Cluster this data over using a sensible subset of variables using two methods described in class. For example, you would not want to use variables like "subject id". Also leave out "motor_UPDRS" and "total_UPDRS". How well do the clusters capture "motor_UPDRS" and "total_UPDRS".

b) Fit a Bayesian Network using this data. Include "motor_UPDRS" and "total_UPDRS", but not both, and force this variable to be the bottom node of the network.

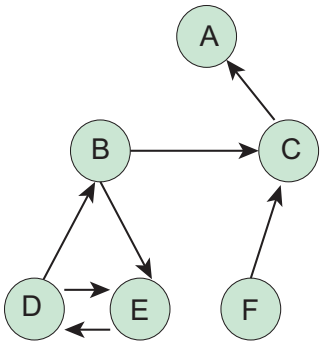
   A collaborator asks you to characterize "Jitter" related variables for a new patient that has a relatively high UPDRS score (two standard deviations above the mean). Use your Bayesian Network to answer this question.

(2) (10 points) The sinking of the Titanic is a famous event in history. The titanic data (https://www.kaggle.com/c/titanic/data) was collected by the British Board of Trade to investigate the sinking. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.

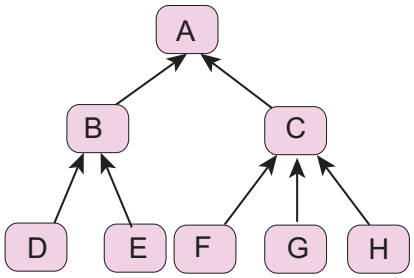
You have been petitioned to investigate this data. Analyze this data with tool(s) that we learned in STA546. Summarize your findings for British Board of Trade. Is their evidence that "women and children" were the first evacuated? What characteristics/demographics are more likely in surviving passengers? What characteristics/demographics are more likely in passengers that perished? How do your results support the popular movie "Titanic" (https://www.imdb.com/title/tt0120338/)? For example, what is the probability that Rose (1st class adult and female) and (3rd class adult and male) would not survive?

3) (10 points) Consider the following webgraphs.

Webgraph A



Webgraph B



(a) Compute the PageRank vector of Webgraph A for damping constants p = 0.05, 0.25, 0.50, 0.75, and 0.95. How sensitive is the PageRank vector, and overall ranking of importance, to the damping constant? Does the relative ranking of importance according to PageRank support your intuition?

(b) Compute the PageRank vector of Webgraph B for damping constant p = 0.15. Interpret your results in terms of the relationship between the number of incoming links that each node has. Does the relative ranking of importance according to PageRank support your intuition?

4) (10 points) Data released from the US department of Commerce, Bureau of the Census is available in R

- data(state)
- ?state

Build a Gaussian Graphical Model using the Graphical Lasso for the 8 predictors (Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. What do you find for different penalties, and how does it compliment (and/or contradict) a model fit with SOM?

5) (10 points) Write a function “from scratch” in the R programming language to implement single-linkage, average linkage and complete linkage agglomerative hierarchical clustering. Try it out on a dataset of your choice. (**Note you may not use built in functions for linkage, but you may use internal/built in functions for dissimilarity).