Statistical Data Mining II
Homework 4

Due: Monday April 30th (11:59 pm)
30points

Directions: **Select only two exercises – a third may be done for extra credit.**

1) (15 points) Consider two networks "karate" and "yeast" (or "kite"), which are available in the package "igraphdata".

```
> library(igraphdata)
> data(yeast)
> ?yeast
> data(karate)
> ?karate
> data(kite)
> ?kite
```

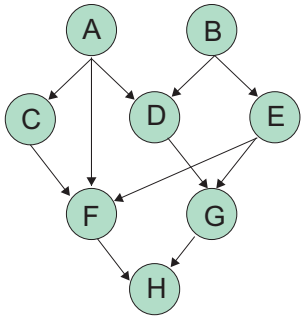
Using the hierarchical random graphs functions in "igraph" perform the following tasks:

(a) Focus on the karate network. Create noisy datasets. Do this by deleting 5% of the edges randomly (track which ones they are). Perform MCMC for a random graph model (as in Clauset et al.) on this data followed by link-prediction. Are you able to predict the edges that you deleted?

(b) Focus on the yeast network (or kite network). Create noisy datasets. Do this by deleting 5% of the edges randomly (track which ones they are). Perform MCMC on this data followed by link-prediction. Are you able to predict the edges that you deleted at random well?

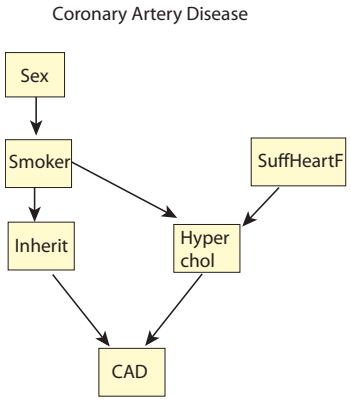
(c) Repeat the exercise in part (a) and (b) after deleting 15%, and 40% of the edges. Comment on your findings.

2) (15 points) Determine if the following statements are "TRUE OR FALSE" based on the DAG. You do not have to show work, e.g., provide your rationale. However, if you do provide an explanation, it will be considered for partial credit.



A) C and G are d-separated.
B) C and E are d-separated.
C) C and E are d-connected given evidence about G.
D) A and G are d-connected given evidence about D and E.
E) A and G are d-connected given evidence on D.

3) (15 points) Consider the "cad1" data set in the package gRbase. There are 236 observations on fourteen variables from the Danish Heart Clinic. A structural learning algorithm has identified the "optimal network" as given below. For simplicity, not all variables are represented in the network.



a) Construct this network in R, and infer the Conditional Probability Tables using the cad1 data. (Hint: the function or extractCPT or cptable may be used from gRain). Identify any d-separations in the graph.
> library(gRbase)
> data(cad1)
> ?cad1

b) Now, we are going to "absorb" evidence into the graph, and propagate this evidence using belief propagation. Once propagated, the "beliefs" (aka probabilities will be updated).

Suppose it is known that a new observation is female with Hypercholesterolemia (high-cholesterol). Absorb this evidence into the graph, and revise the probabilities. How does the probability of heart-failure and coronary artery disease (CAD) change after this information is considered?

c) Building on what you did in part B. I want you to simulate a new data set with **25 observations** conditional upon this new information from part B. Present this new data in a table and include it as a separate attachment.

Using the new data set estimate the probability of "Smoker" and "CAD" given the other variables in your model. Comment on your results. With only 25 observations in your simulated dataset, do they reflect the updated distributions in part B.

(Hint: try the function "simulate.grain" in the gRain package, "table" can also help you determine the frequencies in the simulated data, you may also use "predict").

Do the same thing, but create a larger dataset! Create a new data set, as done in part C, this time with **500 observations**. Save this data and submit it with your assignment as a separate file. Use this data to estimate the probability of

“Smoker” and “CAD” given the other variables in your model. Comment on your results when compared with Part C.

(Hint: try the function “simulate.grain” in the gRain package, “table” can also help you determine the frequencies in the simulated data, you may also use “predict”).