chiehjul
50341634

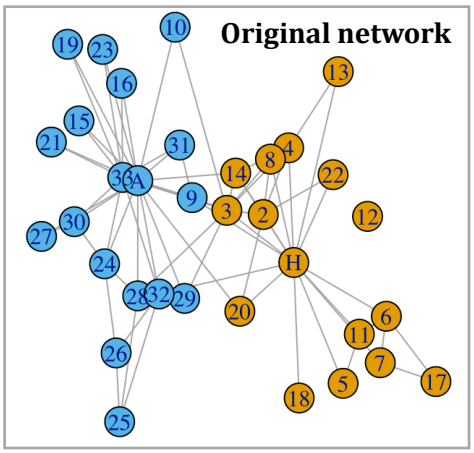
Statistical Data Mining II
Homework 4

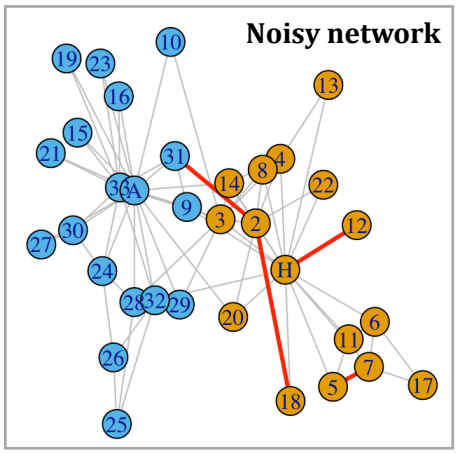
Due: Monday April 30th (11:59 pm)
30points

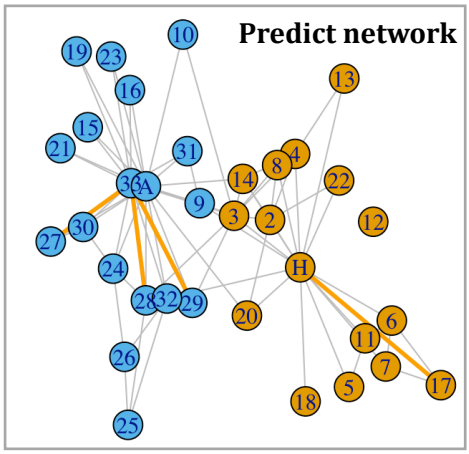
Directions: **Select only two exercises – a third may be done for extra credit.**

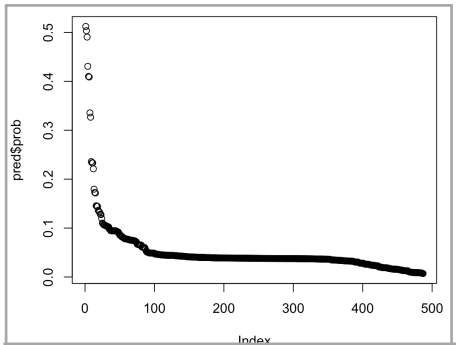
1) (15 points) Consider two networks "karate" and "yeast" (or "kite"), which are available in the package "igraphdata".
> library(igraphdata)
> data(yeast)
> ?yeast
> data(karate)
> ?karate
> data(kite)
> ?kite
Using the hierarchical random graphs functions in "igraph" perform the following tasks:

(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?









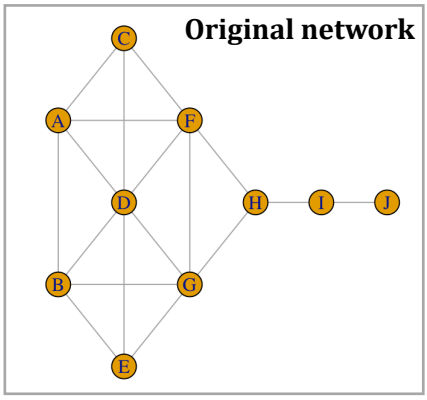
Ans: There are total 78 edges in karate network, so I deleted 4 edges (round of 5%) to create the noisy dataset. I've shown the deleted edges with red lines in the noisy network graph, and the predicted edges with orange lines in the predicted network plot.

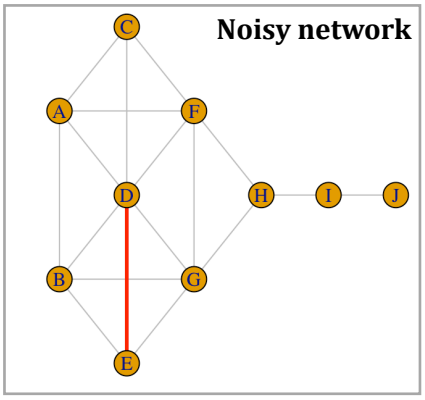
I selected four predicted edges with the highest probability, but the result showed that it was not able to predict the edges accurately in this case.

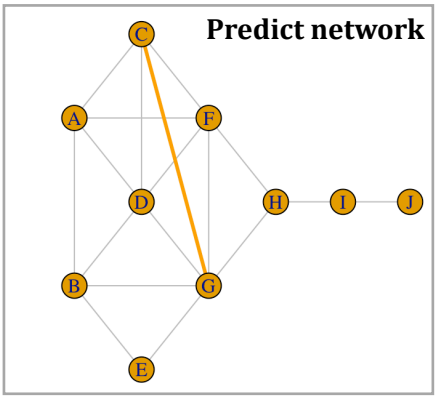
```
Hierarchical random graph, at level 3:
g1        p=0.032
'- g6     p= 0.93  Mr Hi
   '- g11 p=    0  Actor 18 Actor 2  Actor 22 Actor 4  Actor 13 Actor 3  Actor 14 Actor 20 Actor 8  Actor 11
                   Actor 7  Actor 17 Actor 6  Actor 5
'- g16    p= 0.61  Actor 33
   '- g21 p= 0.82  Actor 10 Actor 19 Actor 21 Actor 9  Actor 31 Actor 12 Actor 27 Actor 23 Actor 26 Actor 28
                   Actor 24 Actor 30 Actor 29 Actor 32 Actor 25 Actor 16 Actor 15 John A
```

(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?







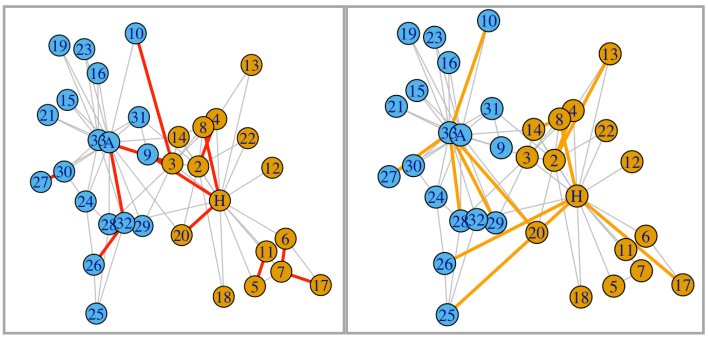

```
Hierarchical random graph, at level 4:
g1          p=0.22  I
'- g8       p=   0  J
   '- g6    p=0.71  G
      '- g5 p=0.25  C  A  F  D  H  E  B
```

```
Hierarchical random graph, at level 7
g1                p=0.22  I
'- g8             p=   0  J
   '- g6          p=0.71  G
      '- g5       p=0.25
         '- g4    p=   1
            '- g7 p=   1  A  C
            '- g3 p=   1  D  F
         '- g9    p=   0  H
            '- g2 p=   1  B  E
```

Ans: In this case, it only deleted one edge due to the less total edges, and it was still not able to predict the deleted edge.
According to the HRG model, if I set the level from 3 to 4, we can observe that node C, A, F, D, H, E, B are in a groups, and will combine with another group which only have node G. And in the level = 7, we can find that there is higher probability that G would have a link with the group {A, C}.

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

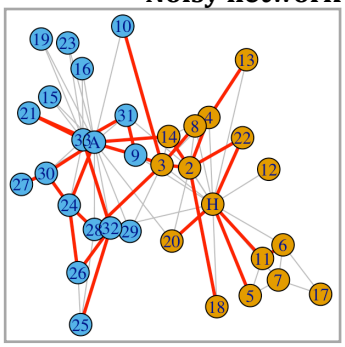
Noisy network **Predict network**

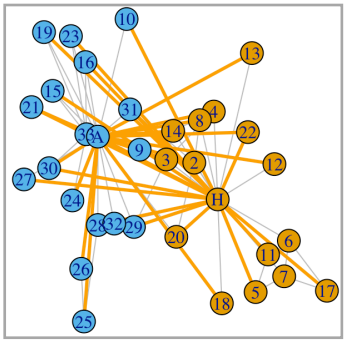


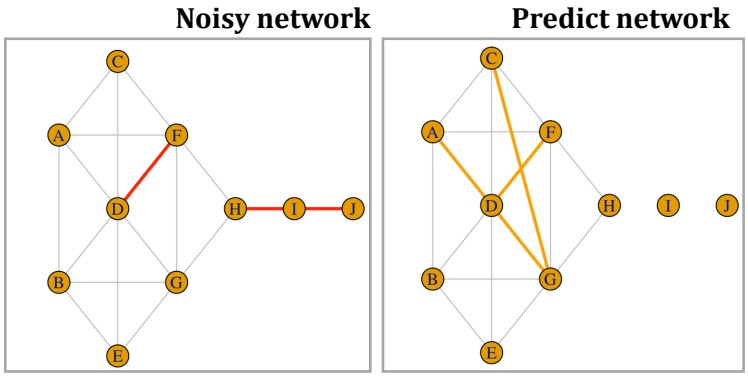
```
Hierarchical random graph, at level 3:
g1        p= 0.03  Actor 17
'- g33    p=0.033
   '- g10 p= 0.93  Mr Hi    Actor 2  Actor 22 Actor 3  Actor 14 Actor 8  Actor 4  Actor 18 Actor 12 Actor 32
                   Actor 7  Actor 13 Actor 5  Actor 11 Actor 6
   '- g5  p= 0.88  Actor 27 Actor 33 Actor 15 Actor 16 Actor 9  Actor 31 Actor 10 Actor 19 Actor 26 Actor 24
                   Actor 30 Actor 28 Actor 25 Actor 23 Actor 21 Actor 29 Actor 20 John A
```

```
Hierarchical random graph, at level 3:
g1        p=0.24  Actor 33
'- g5     p=0.38  John A
   '- g26 p=   0  Mr Hi    Actor 23 Actor 19 Actor 29 Actor 9  Actor 30 Actor 12 Actor 21 Actor 31 Actor 2
                  Actor 22 Actor 27 Actor 20 Actor 15 Actor 8  Actor 32 Actor 13 Actor 24 Actor 18 Actor 17
                  Actor 7  Actor 6  Actor 11 Actor 5  Actor 16 Actor 4  Actor 26 Actor 14 Actor 3  Actor 28
                  Actor 25 Actor 10
```

Noisy network **Predict network**

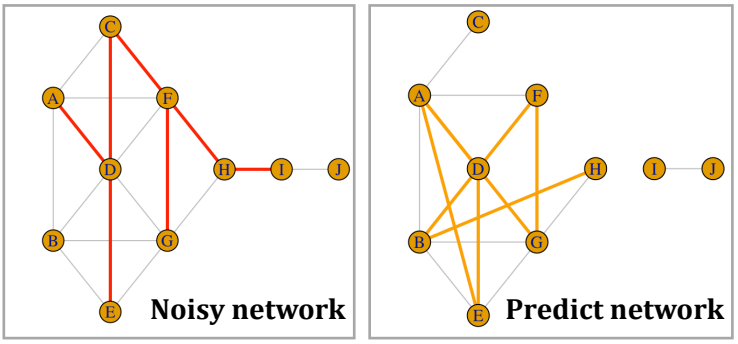






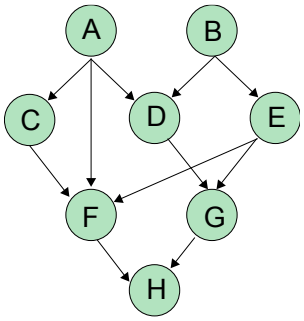

```
Hierarchical random graph, at level 3:
g1          p=   0  I
'- g2       p=   0  J
   '- g6    p=0.53  H  D  A  C  B  G  F  E
```

```
Hierarchical random graph, at level 3:
g1          p=0.11  H
'- g8       p=0.38  A
   '- g2    p=0.14  C  J  B  F  D  G  E  I
```



Ans: As the deleted edges increase, we can find that the difference between the predicted network and the original network is getting larger and larger.
In the left plot (40% of the edges are deleted), the HRG model has only one node in the first and second clusters, but the third cluster contains all nodes. The tree will become more unbalanced if the more edges deleted.

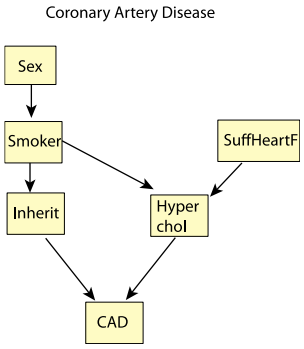
2) (15 points) Determine if the following statements are "TRUE OR FALSE" based on the DAG. You do not have to show work, e.g., provide your rationale. However, if you do provide an explanation, it will be considered for partial credit.



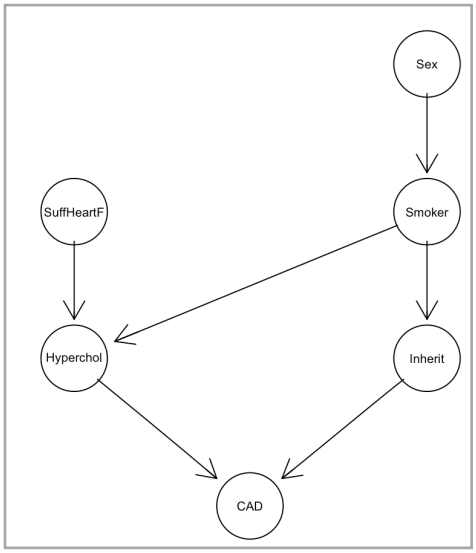
A) C and G are d-separated.
B) C and E are d-separated.
C) C and E are d-connected given evidence about G.
D) A and G are d-connected given evidence about D and E.
E) A and G are d-connected given evidence on D.

Ans:
A) False, C <- A-> D -> G is active.
B) True, all path have v-structure.
C) True, C<- A -> D -> G <-E will be active given G.
D) False
E) True, A -> D <- B -> E -> G will be active given 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




```
Sex
   Female      Male
0.1991525 0.8008475
```

```
Smoker     Female      Male
   No  0.3617021 0.1798942
   Yes 0.6382979 0.8201058
```

```
SuffHeartF
        No       Yes
0.7076271 0.2923729
```

```
Inherit         No       Yes
    No  0.8235294 0.6486486
    Yes 0.1764706 0.3513514
```

```
, , SuffHeartF = No

        Smoker
Hyperchol    No       Yes
      No  0.675 0.4645669
      Yes 0.325 0.5354331

, , SuffHeartF = Yes

        Smoker
Hyperchol        No       Yes
      No  0.2727273 0.3275862
      Yes 0.7272727 0.6724138
```

```
, , Inherit = No

     Hyperchol
CAD           No       Yes
  No  0.8214286 0.4487179
  Yes 0.1785714 0.5512821
```

```
, , Inherit = Yes

     Hyperchol
CAD    No  Yes
  No  0.5 0.26
  Yes 0.5 0.74
```

Ans:
- Smoker and SuffHeartF are d-separated.
- Smoker and SuffHeartF are not d-separated given CAD.
- Sex and Hyperchol are d-separated given smoker.

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").