

Report Homework 4

Name: Akshay Adlakha

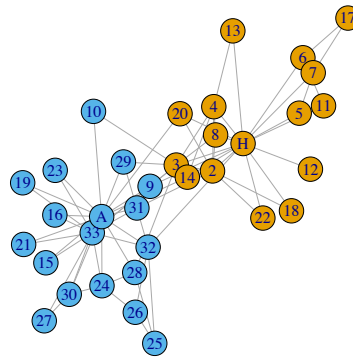
UBIT: akshayad

Person#: 50317479

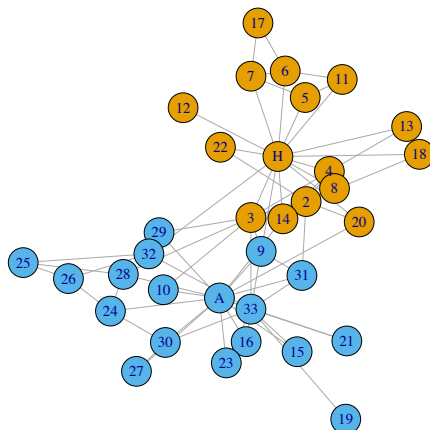
Problem 1: In this problem, I was given a karate and kite data from igraphdata package. I had to perform the MCMC for a random graphical model.

I removed the 5% of edges randomly from the karate network using the sample method and delete_edges. To find the number of edges which 5% of network, used the gsize function and multiply by 0.05 to get the number and round the number to the next integer.

Here is the karate network plot.



Here is the network after deleting edges:



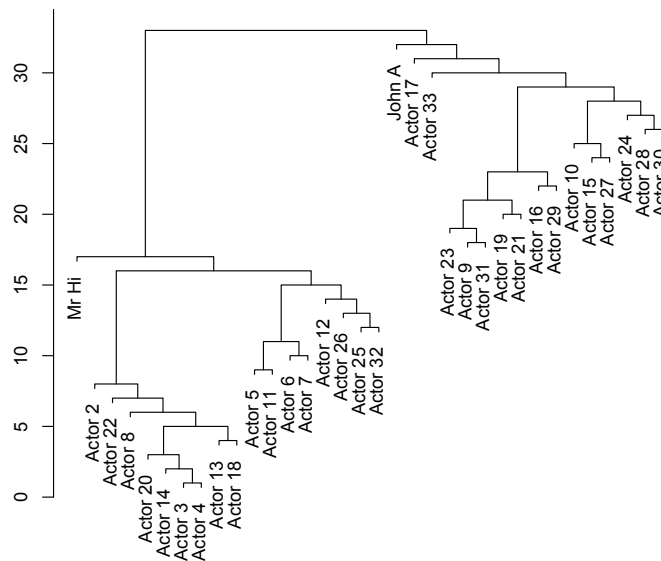
Originally I had the 78 edges and after deleting 5% edges left with 74 edges.

```
> gsize(g)
[1] 78
> gsize(g1)
[1] 74
```

The deleted edges are stored to check the whether model be able to predict them or not. These the deleted edges.

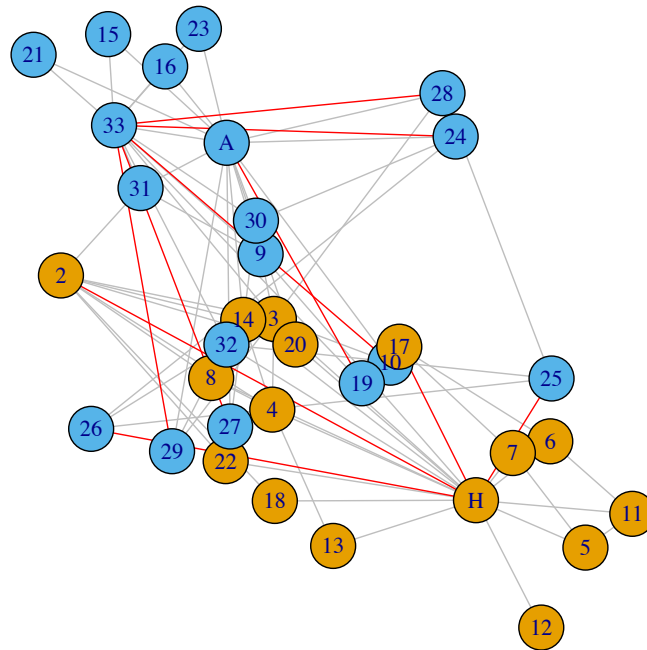
```
IGRAPH 2b38130 UNW- 34 4 -- Zachary's karate club network
+ attr: name (g/c), Citation (g/c), Author (g/c), Faction (v/n), name (v/c), label (v/c), color (v/n),
+ weight (e/n)
+ edges from 2b38130 (vertex names):
[1] Mr Hi --Actor 2 Actor 19--John A Actor 24--Actor 33 Actor 32--John A
```

Here is the dendrogram to understand the clustering.



This is clustering the similar nodes together.

Then, to predict edges performed MCMC for a random graph model. Used the predict_edges function from the igraph package. This does the link prediction for graph in hrg method and predicts probabilities of all possible edges. Here is plotted. I chose top 10 prediction with highest probabilities by model and colored them with red.

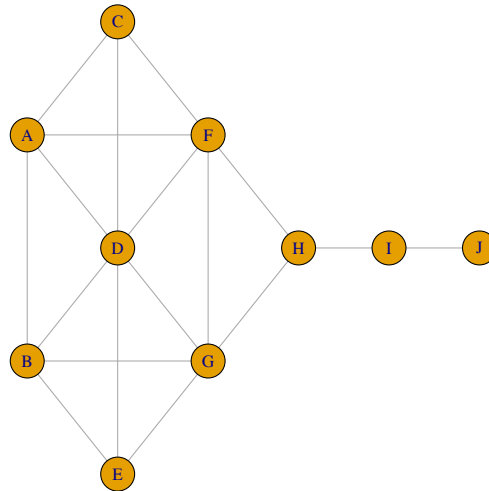


Here is the probabilities with corresponding edges.

```
> pred$edges
      [,1] [,2]
[1,]    1    2
[2,]    1   17
[3,]   19   34
[4,]    1   25
[5,]   27   33
[6,]    1   26
[7,]   24   33
[8,]   28   33
[9,]   29   33
[10,]  10   33
[11,]  11   17
[12,]    5   17
[13,]    7   11
[14,]    5    6
[15,]    1   19
[16,]    1   31
[17,]    1   10
[18,]    1   29
[19,]    1   34
[20,]    1   21
[21,]    1   24
[22,]    1   23
[23,]    1   16
[24,]    1   28
[25,]    1   27
[26,]    1   30
[27,]    1   15
[28,]    1   33
[29,]   32   34
[30,]   26   34
```

From this graph, able to predict the three edges(Mr Hi --Actor 2 Actor 19--John A Actor 24--Actor 33) out of 4 deleted edges in top 10 edges. The fourth deleted edge is at number 29.

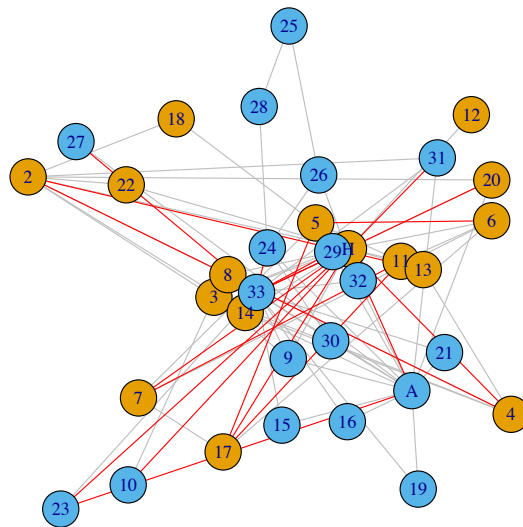
Then, performed the same by removing 5% of edges from the kite network. Here is the kite network.



After deleting the edge from the network, kept the deleted edge for tracking.

```
IGRAPH ff12970 UN-- 10 1 -- Krackhardt's kite
+ attr: name (g/c), layout (g/n), Citation (g/c), Author (g/c), URL (g/c), label (v/c), Firstname
  (v/c), name (v/c)
+ edge from ff12970 (vertex names):
[1] A--D
```

A to D edge is removed from the network. Then predicted to see if it is able to have the deleted edge.



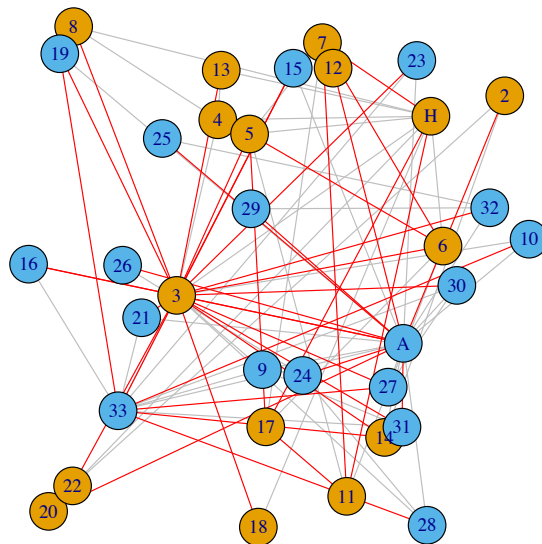
Here considered the top 20 predicted edges. It is able to predict 8 out of 12 deleted edges in top 20.

Then, deleted 40% of edges from the network.

```
> gsize(g)
[1] 78
> gsize(g140)
[1] 46
```

```
IGRAPH b380ba4 UNW- 34 32 -- Zachary's karate club network
+ attr: name (g/c), Citation (g/c), Author (g/c), Faction (v/n), name (v/c), label (v/c), color (v/n),
+ weight (e/n)
+ edges from b380ba4 (vertex names):
[1] Mr Hi --Actor 32 Mr Hi --Actor 20 Mr Hi --Actor 14 Mr Hi --Actor 12 Mr Hi --Actor 11
[6] Mr Hi --Actor 9 Mr Hi --Actor 7 Mr Hi --Actor 2 Actor 2 --Actor 31 Actor 2 --Actor 20
[11] Actor 2 --Actor 18 Actor 2 --Actor 8 Actor 2 --Actor 4 Actor 2 --Actor 3 Actor 3 --Actor 29
[16] Actor 3 --Actor 14 Actor 3 --Actor 8 Actor 4 --Actor 14 Actor 6 --Actor 7 Actor 15--Actor 33
[21] Actor 16--John A Actor 19--Actor 33 Actor 20--John A Actor 24--John A Actor 24--Actor 33
[26] Actor 24--Actor 30 Actor 25--Actor 28 Actor 25--Actor 26 Actor 26--Actor 32 Actor 29--John A
[31] Actor 31--John A Actor 32--Actor 33
```

The graph with top 40 predicted edges.



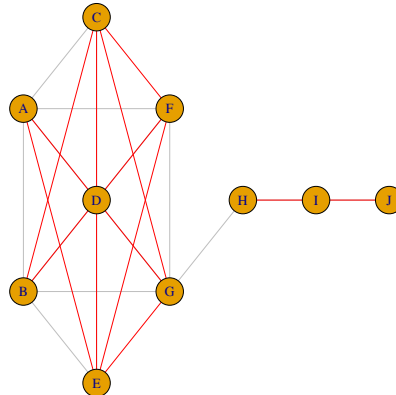
It is able to predict some of the edges from the deleted edges in the network but not all edges. As on increasing the number of edges to be removed, the prediction performance goes down. As we can see that the edges predicted in the prediction are changing with the number of edges removed in the network.

Performed the same with kite data removing 15% of edges.

```
IGRAPH 3da23dc UN-- 10 3 -- Krackhardt's kite
+ attr: name (g/c), layout (g/n), Citation (g/c), Author (g/c), URL (g/c), label (v/c), Firstname
  (v/c), name (v/c)
+ edges from 3da23dc (vertex names):
[1] C--F E--G F--H
```

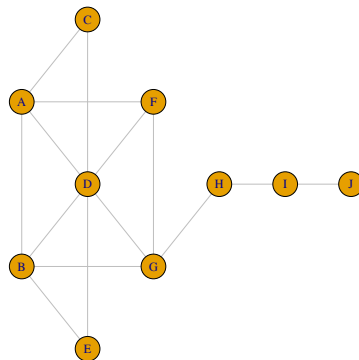
And, the predicted result:

It is able to predict only 2 edges from the three deleted edges. I.e C - - F and E - - G. The conditional dependencies are getting affected by number of edges to be removed. This is so impacting the prediction.

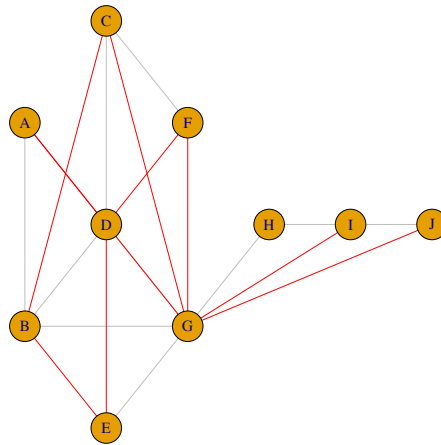


Now, 40 % are removed from the network. Here are the deleted edges.

```
IGRAPH 167d414 UN-- 10 8 -- Krackhardt's kite
+ attr: name (g/c), layout (g/n), Citation (g/c), Author (g/c), URL (g/c), label (v/c), Firstname
  (v/c), name (v/c)
+ edges from 167d414 (vertex names):
[1] A--F A--D A--C B--E D--F D--E F--H F--G
```

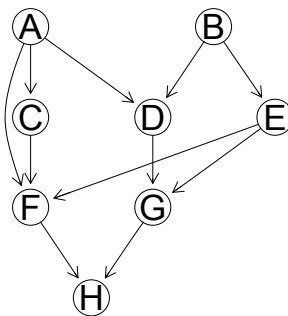


The predicted network:



The prediction are not that much good after removing 40% and probabilities are negligible. As we can see that it is predicting some of the edges not all. So, Removing the number of edges randomly impacts the prediction and changes the conditional dependencies and the prediction performance.

Problem 2: In this problem, a DAG network is given and based on this network, had to check the statements. I implemented this in R and explanation is also given below.



A) C and G are d-separated. - False

To answer the d-separation, have to check the paths from source to destination to be inactive. If any of the paths comes out to be active then, we can't say the d-separation.

To go from C to G, We have C-A-D-G path. In this, C-A-D is common cause and A-D-G is causal chain making a path active. We have path C-F-H-G. In this C-F-H makes an active triplets. So, this is an active path. In path, C-F-E-G. In this, we have F-H-G is common effect chain.

Thus we have path active to go from C to G. We can't say the d-separation. Then C and G are not d-separated.

B) C and E are d-separated. - True

To go from C to E, we have path C-F-E. This is an inactive triplets.

C-F-H-G-E - This also makes an inactive. We have two other paths C-A-D-G-E and C-A-D-B-E. They are also inactive. C-F-E, F-H-G, A-F-E and D-G-E. They are a common effect chain and blocking all the other paths to go from C to E. No active path is there.

So, we can say that C and E are d-separated.

C) C and E are d-connected given evidence about G. - True

In the we are given an evidence about G. All paths are still inactive except the path C-A-D-G-E becomes active. So, we have active path between C and E. So we can say that C and E are d-connected.

D) A and G are d-connected given evidence about D and E. - False

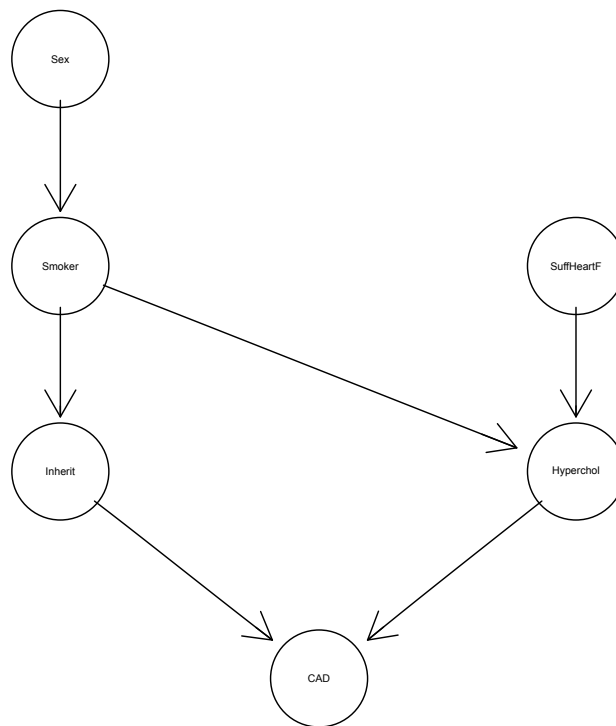
In this we are given an evidence about D and E. We have paths A-C-F-H-G and A-F-E-G which are inactive path. And, we have A-D-G which is causal chain and is an inactive triplets. In path A-D-B-E-G, B-E-G also makes an inactive path as casual chain. Paths are following these triplets. All paths are inactive. There is no connection between A and D.

E) A and G are d-connected given evidence on D. - True

Now, the evidence is given about D only, the A-D-B-E-G becomes active. So, A and G are not d-separated and there is a connection between A and G. This is true.

Problem 3: In this problem, I was given a cad1 database in the package gRbase. I had to construct the network given in the question and infer conditional probability tables.

Made the network using the dagList. Plotting the network gives the following result:



And check the d-separation for some combinations, like:

- Sex and Suffheartf are d-separated given no condition.
- Sex and Hyperchol are independent given Smoker and Suffheartf.
- Sex and inherit given smoker.
- Inherit and Suffheartf are d-separated.
- Smoker and cad given inherit and hyperchol.

Conditional probability table using extractCPT. Here is the result.

Smoker			
Hyperchol	No	Yes	
No	0.6741294	0.4646226	
Yes	0.3258706	0.5353774	

, , SuffHeartF = Yes

Smoker			
Hyperchol	No	Yes	
No	0.2767857	0.3281787	
Yes	0.7232143	0.6718213	

\$CAD

, , Hyperchol = No

Inherit			
CAD	No	Yes	
No	0.8206651	0.5000000	
Yes	0.1793349	0.5000000	

, , Hyperchol = Yes

Inherit			
CAD	No	Yes	
No	0.4488491	0.2609562	
Yes	0.5511509	0.7390438	

Here is the summary result.

```
Independence network: Compiled: TRUE Propagated: FALSE
Nodes : Named chr [1:6] "Sex" "Smoker" "SuffHeartF" "Inherit" "Hyperchol" "CAD"
- attr(*, "names")= chr [1:6] "Sex" "Smoker" "SuffHeartF" "Inherit" ...
Number of cliques: 4
Maximal clique size: 3
Maximal state space in cliques: 8
```

Now, the evidence is given to absorb into the graph. I.e, the sex is female with Hypercholesterolemia (high-cholesterol).

Following are the Marginal probabilities before and after absorbing the evidence.

Before:

\$SuffHeartF		
SuffHeartF	No	Yes
	0.7074513	0.2925487
\$CAD		
CAD	No	Yes
	0.5401277	0.4598723

After:

\$SuffHeartF		
SuffHeartF	No	Yes
	0.6167542	0.3832458
\$CAD		
CAD	No	Yes
	0.3927255	0.6072745

From the above results, we can see that having CAD and heart failure increases with evidence.

Joint probabilities before and after absorbing the evidence.

Before:

		CAD	
SuffHeartF		No	Yes
	No	0.3955084	0.3119429
	Yes	0.1446194	0.1479293

After:

		CAD	
SuffHeartF		No	Yes
	No	0.2412694	0.3754848
	Yes	0.1514562	0.2317897

In the joint probability, not having both CAD and SuffHeartF decreases with absorbing evidence but we can see the probability increases in all other scenarios after absorbing the given evidence.

Conditional probabilities before and after the evidence is given below.

Before:

		SuffHeartF	
CAD		No	Yes
	No	0.7322497	0.2677503
	Yes	0.6783252	0.3216748

After:

		SuffHeartF	
CAD		No	Yes
	No	0.6143460	0.3856540
	Yes	0.6183115	0.3816885

In conditional probability, the probability of SuffHeartF increases with evidence but the for CAD it decreases. But the probability of not having hear failure decreases with evidence.

Then, used simulate.grain function to simulate the data with 25 observations and stored them in data25.RData.

Here is the simulated data with 25 observations conditional upon the information given.

	Sex	Smoker	SuffHeartF	Inherit	Hyperchol	CAD
1	Female	Yes	No	No	Yes	Yes
2	Female	Yes	No	No	Yes	Yes
3	Female	Yes	Yes	Yes	Yes	No
4	Female	No	No	Yes	Yes	Yes
5	Female	Yes	No	No	Yes	Yes
6	Female	Yes	No	Yes	Yes	Yes
7	Female	No	No	Yes	Yes	Yes
8	Female	Yes	No	No	Yes	Yes
9	Female	Yes	Yes	No	Yes	Yes
10	Female	No	No	No	Yes	Yes
11	Female	No	No	Yes	Yes	Yes
12	Female	No	No	Yes	Yes	Yes
13	Female	Yes	Yes	No	Yes	Yes
14	Female	Yes	No	No	Yes	Yes
15	Female	Yes	No	No	Yes	Yes
16	Female	No	Yes	No	Yes	Yes
17	Female	Yes	No	No	Yes	No
18	Female	Yes	No	No	Yes	Yes
19	Female	Yes	Yes	No	Yes	Yes
20	Female	Yes	No	No	Yes	Yes
21	Female	Yes	Yes	Yes	Yes	Yes
22	Female	Yes	No	No	Yes	No
23	Female	Yes	Yes	No	Yes	Yes
24	Female	No	No	No	Yes	No
25	Female	Yes	No	No	Yes	Yes

The probability result using the simulated data with 25 observations:

```
$pred
$pred$Smoker
[1] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"

$pred$CAD
[1] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"

$Evidence
[1] 0.04492834 0.04492834 0.01150195 0.01989031 0.04492834 0.01989031 0.01989031 0.04492834 0.02877581 0.04492834 0.01989031 0.01989031 0.02877581 0.04492834
[15] 0.04492834 0.02877581 0.04492834 0.04492834 0.02877581 0.04492834 0.01150195 0.04492834 0.02877581 0.04492834 0.04492834
```

The predict function is used to predict. The "class" type is used to predict which assigns the value to the class with the highest probability.

As we can see that the probability of having a CAD and a smoker is 100% as all the prediction of Smoker and CAD are Yes.

And, it is reflecting the distribution from that of Part-B. In part B we didn't had 100% accuracy. But with simulated data, we got 100% result predicting all the smokers having CAD.

And then, generated 500 observations and stored the data in data500.RData and then estimated probability of Smoker and CAD given other variables.

Here is the prediction result of Smoker and CAD. The evidence gives the probability of the each configuration of the variables.

[illegible]

From the above results, we can see that all the predictions are Yes. So, the probability of a smoker having CAD comes out to 100% for 500 observations. And, we can say that it is predicting all the smokers having CAD not depending on the observations into consideration.