Statistical Data Mining II-Project 4

Problem 1

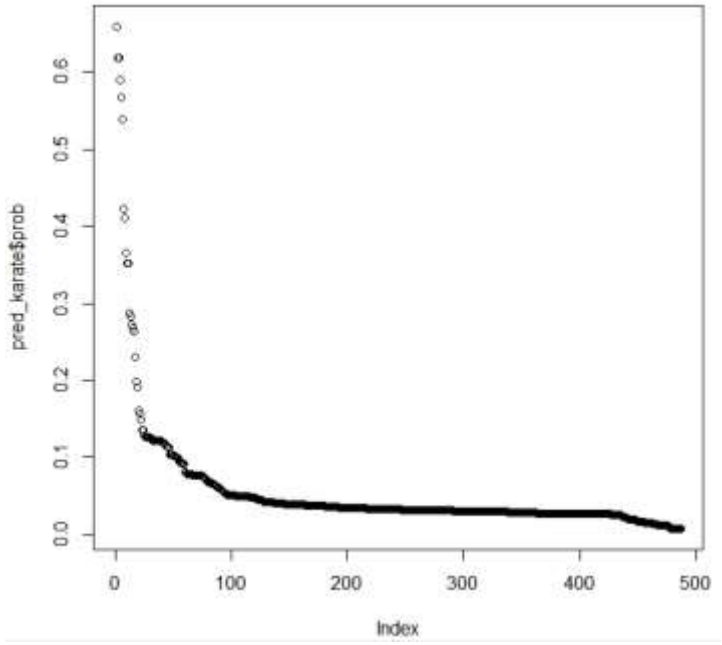
In this problem we had to consider two networks “karate” and “kite”, which are available in the package “igraphdata”.

Part A

In this part we used the hierarchical random graphs functions in “igraph” and focused on the karate network. We had to create noisy datasets by deleting 5% of the edges randomly. The total number of edges found in the karate data was 78 edges. 5% of them were deleted randomly resulting in 4/78 edges being deleted, which included the following:

Actor 31--John A
Mr Hi --Actor 9
Mr Hi --Actor 5
Actor 27--John A

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.61 and as the index/rank increase the probability for getting a missing edge decreases.

Using the link- prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

Node1	Node2	Rank	Probofedgemissing
1	5	2	0.61923256
1	9	110	0.04979663
27	34	16	0.26323107
31	34	5	0.56813381

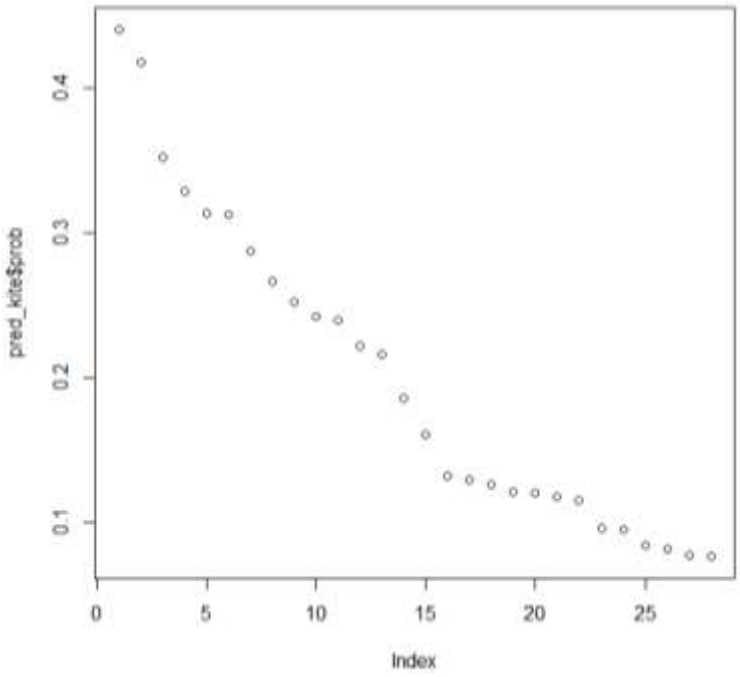
From the above table we can see that we were only able to predict some of the edges that would be deleted. These included the edge 1-—5, which had the highest predicted probability of 0.619 and the edge 31---34 which had the 5th highest probability pf 0.568. The link prediction was not able to predict the other edges 1—9 and 27—34, as they had lower probability of being deleted, 0.04 and 0.26, respectively.

Part B

In this part we used the hierarchical random graphs functions in "igraph" and focused on the kite network. We had to create noisy datasets by deleting 5% of the edges randomly. The total number of edges found in the karate data was 18 edges. 5% of them were deleted randomly resulting in 1/18 edges being deleted, which included the following:

B—G

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.41 and as the index/rank increase the probability for getting a missing edge decreases.

Using the link- prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

Node1	Node2	Rank	Probofedgemissing
2	5	13	0.215595

From the above table we can see that we were not able to predict the edges that would be deleted. The link-predicted that the B-D edge would be deleted, but it did not do it well, since this edge has a rank of 13th of being deleted with a probability of 0.215, whereas we found that the edge B-G should be the deleted edge.

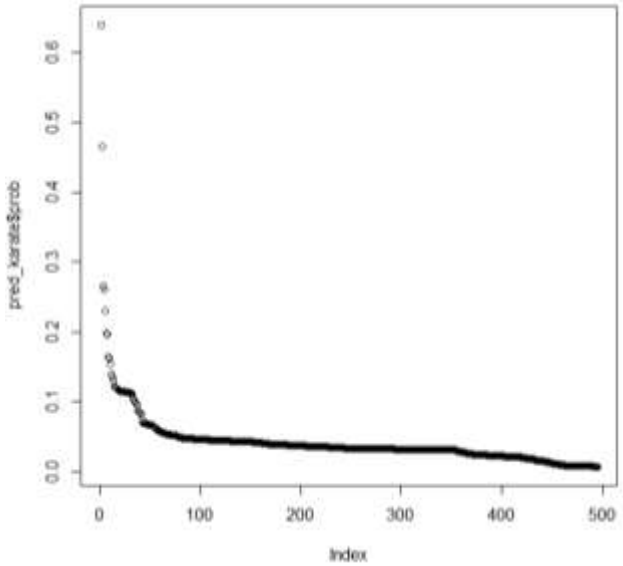
Part C

In this part we had to repeat the exercise in part (a) and (b) after deleting 15%, and 40% of the edges.

First we started with the karate network. Using the hierarchical random graphs functions in "igraph", we had to create noisy datasets by deleting 15% of the edges randomly. The total number of edges found in the karate data was 78 edges. 15% of them were deleted randomly resulting in 12/78 edges being deleted, which included the following:

Actor 2 --Actor 22
Actor 24--Actor 33
Actor 3 --Actor 33
Actor 27--Actor 30
Actor 29--Actor 32
Mr Hi --Actor 5
Actor 6 --Actor 11
Actor 25--Actor 28
Actor 30--Actor 33
Actor 32--Actor 33
Actor 26--Actor 32
Actor 3 --Actor 29

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.61 and index 2 is 0.46.

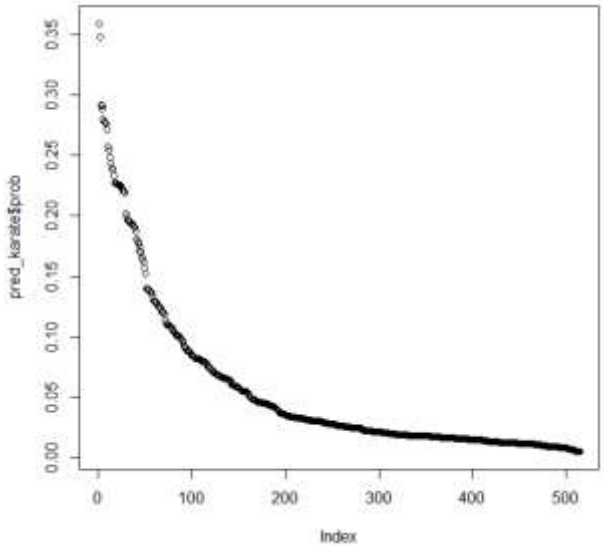
Using the link- prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

Node1	Node2	Rank	Probofedgemissing
1	5	2	0.46517476
2	22	141	0.04356147
24	33	54	0.06417176
25	28	73	0.05222544
26	32	50	0.06726366
27	30	419	0.02030468
29	32	381	0.02404648
3	29	81	0.04858282
3	33	65	0.05451842
30	33	52	0.06699109
32	33	42	0.07445415
6	11	36	0.09895997

From the above table we can see that we were only able to predict one edge out of the 12 deleted edges, which is the Mr Hi --Actor 5 i.e. 1—5. This is because it has a high probability of being deleted, with probability of 0.46 and it has a rank of 2, whereas the other deleted edges have lower probabilities and are not close to being in the top 20 list of being deleted edges. This it was not able to predict the random deleted edges well.

After, we repeated the same process, but by deleting 40% of the edges randomly. 40% of deleted edges results in 31/78 edges being deleted, which included the following:
Actor 24--Actor 26
Actor 15--John A
Actor 20--John A
Actor 32--John A
Actor 3 --Actor 8
Actor 2 --Actor 3
Actor 3 --Actor 9
Actor 6 --Actor 7
Actor 2 --Actor 22
Mr Hi --Actor 22
Actor 30--John A
Actor 4 --Actor 8
Actor 2 --Actor 20
Actor 9 --Actor 31
Actor 5 --Actor 11
Actor 23--Actor 33
Actor 9 --John A
Actor 9 --Actor 33
Mr Hi --Actor 11
Actor 15--Actor 33
Actor 24--John A
Actor 4 --Actor 14
Mr Hi --Actor 8
Mr Hi --Actor 12
Actor 21--Actor 33
Mr Hi --Actor 14
Actor 31--John A
Actor 7 --Actor 17
Actor 23--John A
Actor 26--Actor 32
Actor 10--John A

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.36 and index 2 is 0.35.

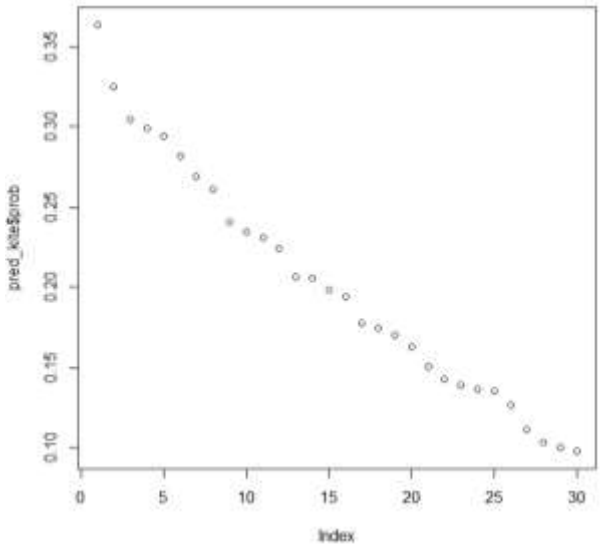
Using the link- prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

Node1	Node2	Rank	Probofedgemissing
1	11	8	0.27727198
1	12	33	0.19544173
1	14	13	0.24880639
1	22	37	0.19312118
1	8	10	0.27128845
10	34	44	0.17439071
15	33	156	0.05447250
15	34	130	0.06733734
2	20	81	0.10398606
2	22	121	0.07189492
2	3	41	0.18050642
20	34	118	0.07492496
21	33	27	0.22124993
23	33	149	0.05797076
23	34	119	0.07370377
24	26	162	0.04975715
24	34	1	0.35966351
26	32	132	0.06642054
3	8	52	0.13969223
3	9	73	0.11073668
30	34	2	0.34869777
31	34	9	0.27638585
32	34	4	0.29096569
4	14	174	0.04523430
4	8	190	0.04055977
5	11	276	0.02408630
6	7	236	0.02970409
7	17	272	0.02426374
9	31	391	0.01499601
9	33	125	0.07007566
9	34	106	0.08166048

From the above table we can see that we were only able to predict some of the edges that would be deleted. These include the edges in the top 10th rank, which include 20-34, 30-34, 32-34, 1—8, 31-34,1—11, which have a rank of 1,2,4,8,9,10 receptively and with the probability of being deleted were 0.36, 0.35, 0.29, 0.277, 0.276 and 0.271, respectively. Thus, the link prediction was able to predict 6/31 deleted edges, while the other edges we not predicted well as they had a much lower probability of being deleted.

Afterwards, we repeated the same process using the kite network. Using the hierarchical random graphs functions in "igraph", we created noisy datasets by deleting 15% of the edges randomly. The total number of edges found in the karate data was 78 edges. 15% of them were deleted randomly resulting in 3/18 edges being deleted, which included the following:
D--F
B--G
E—G

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.36 and index 2 is 0.33. Using the link- prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

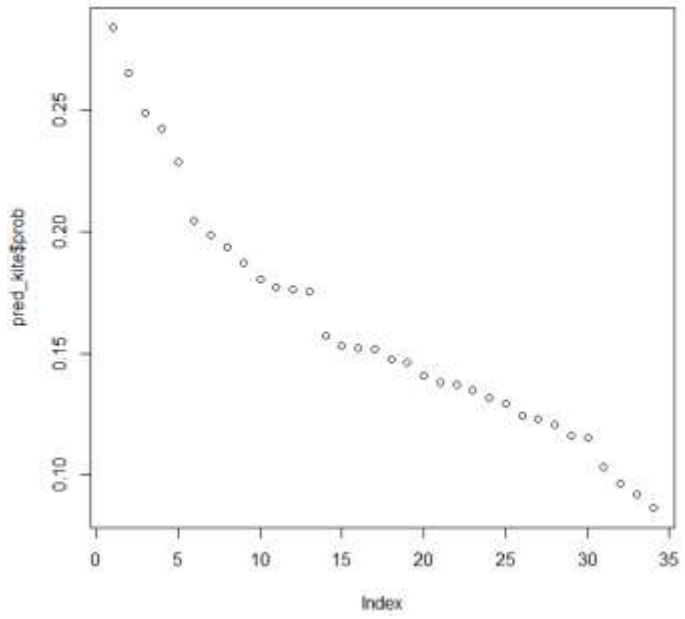
Node1	Node2	Rank	Probofedgemissing
2	7	9	0.2406719
4	6	1	0.3636387
5	7	15	0.1981642

From the above table we can see that we were only able to predict one edge out of the 3 deleted edges, which is D-F which is 4--6. This is because it has a high probability of being deleted, with probability of 0.36 and it has a rank of 1, whereas the other deleted edges have lower probabilities and are ranked lower.

The same process was repeated, but by deleting 40% of the edges randomly. 40% of deleted edges results in 7/18 edges being deleted, which included the following:

C--D
G--H
B--G
C--F
D--G
B--D
A--B

We had to perform MCMC on this data followed by link-prediction. The probability from the prediction for an edge being missing is shown below:



The graph about shows that the probability of having a missing edge for index 1 is around 0.29 and index 2 is around 0.26.

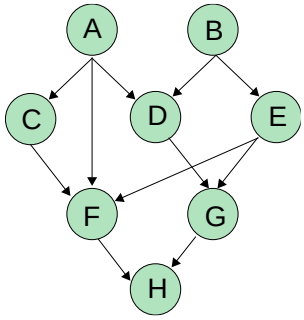
Using the link-prediction to predict which edges are likely to be missing, we generated a table which included the edges between which nodes would be deleted, their corresponding predicted probability and their rank:

Node1	Node2	Rank	Probofedgemissing
1	2	13	0.1755503
2	4	16	0.1520435
2	7	21	0.1378997
3	4	10	0.1803539
3	6	3	0.2488229
4	7	8	0.1937140
7	8	24	0.1313515

From the above table we can see that we were able to predict some of the edges that would be deleted. These include the edges in the top 10^{th} rank, which include 3—6, 4—7, 3—4 which have a rank of 1, 3 and 10 respectively and with the probability of being deleted were 0. 25, 0.19 and respectively. Thus, the link prediction was able to predict 3/7 deleted edges well, while the other edges we not predicted well as they had a much lower probability of being deleted.

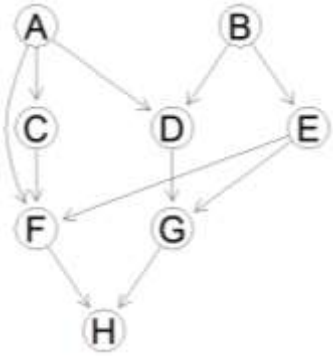
Problem 2

In this problem we had to determine if the following statements are "TRUE OR FALSE" based on the DAG below.



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.

First, the conditional probability distribution of Bayesian Network for the DAG was determined which is P(A,B,C,D,E,F,G.H)= P(A)P(B)P(C|A)P(D|A,B)P(E|B)P(F|A,C,E)P(G|D,E)P(H|F,G). The DAG graph was plotted afterward using R as shown below.



The d-separation for the above statements where determined for the DAG using R function dSep in the ggm package and answers were the reversed for d-connected statements.

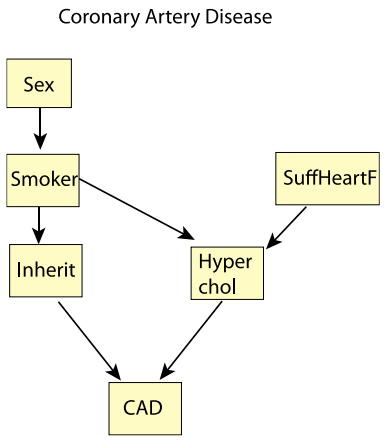
```
> dSep(as(CPD_graph,"matrix"),"C","G",cond = NULL)
[1] FALSE
> dSep(as(CPD_graph,"matrix"),"C","E",cond = NULL)
[1] TRUE
> dSep(as(CPD_graph,"matrix"),"C","E",c("G"))
[1] FALSE
> dSep(as(CPD_graph,"matrix"),"A","G",c("D","E"))
[1] TRUE
> dSep(as(CPD_graph,"matrix"),"A","G",c("D"))
[1] FALSE
```

Thus, the result will be

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

Problem 3

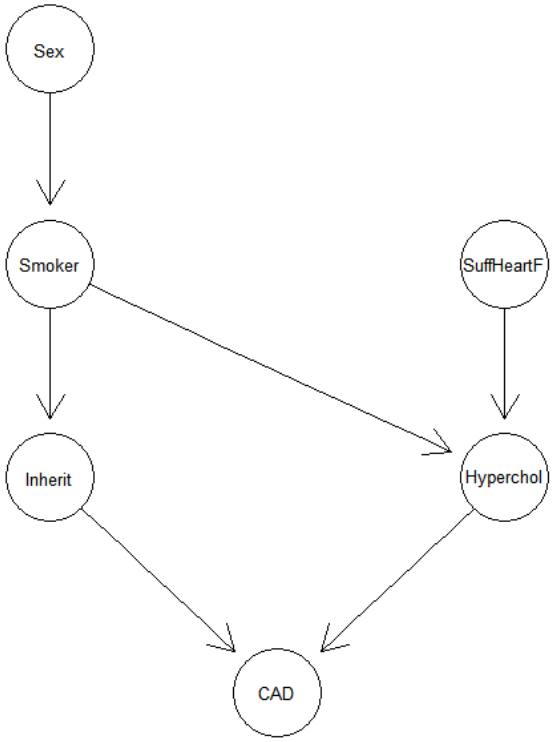
In this problem we had to consider the "cad1" data set in the package gRbase. There are 236 observations on fourteen variables from the Danish Heart Clinic. A structural learning algorithm has identified the "optimal network" as given below. For simplicity, not all variables are represented in the network.



Part A

In this part we had to construct this network in R, which required the conditional probability distribution for Bayesian network:

P(sex,smoker,inherit,suffheartf,hyperchol,cad)=P(sex,smoker|sex)P(suffheartf)
*P(inherit|smoker)*P(hyperchol|suffheartf,smoker)*P(cad|inherit,hyperchol)



We also had to infer the Conditional Probability Tables for the above network using the cad1 data, as shown below.

CPD for sex

```
Sex
   Female      Male
0.1991525 0.8008475
```

CPD for suffering from heart failure

```
SuffHeartF
       No       Yes
0.7076271 0.2923729
```

CPD for smoker

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

CPD for inherit

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

CPD for Hyper cholesterol

```
$Hyperchol
, , SuffHeartF = No

          Smoker
Hyperchol        No       Yes
      No  0.6750000 0.4645669
      Yes 0.3250000 0.5354331

, , SuffHeartF = Yes

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

CPD for Coronary artery disease

```
$CAD
, , Hypercho1 = No

      Inherit
CAD          No       Yes
  No  0.8214286 0.5000000
  Yes 0.1785714 0.5000000

, , Hypercho1 = Yes

      Inherit
CAD          No       Yes
  No  0.4487179 0.2600000
  Yes 0.5512821 0.7400000
```

We had to identify any d-separations in the graph network as shown below:

```
> dSep(as(cad_dag, "matrix"),"Sex", "SuffHeartF",cond = NULL)
[1] TRUE
> dSep(as(cad_dag, "matrix"), "SuffHeartF","Smoker",cond = NULL)
[1] TRUE
> dSep(as(cad_dag, "matrix"), "SuffHeartF","Inherit",cond = NULL)
[1] TRUE
> dSep(as(cad_dag, "matrix"), "Sex", "CAD", c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"), "Sex", "Hyperchol", c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"), "Sex", "SuffHeartF", c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"Inherit", "Hyperchol",c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"Inherit", "Sex",c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"Inherit", "SuffHeartF",c("Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"), "Hyperchol","Sex", c("Smoker", "SuffHeartF"))
[1] TRUE
> dSep(as(cad_dag, "matrix"), "Hyperchol","Inherit", c("Smoker", "SuffHeartF"))
[1] TRUE
> dSep(as(cad_dag, "matrix"), "SuffHeartF", "CAD", c("Hyperchol","Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"SuffHeartF", "Inherit",c("Hyperchol","Smoker"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"Smoker", "CAD",c("Inherit", "Hyperchol"))
[1] TRUE
> dSep(as(cad_dag, "matrix"),"SuffHeartF", "CAD",c("Inherit", "Hyperchol"))
[1] TRUE
```

Part B

In this part we supposed that it is known that a new observation is female with Hypercholesterolemia (high-cholesterol).We had to absorb this evidence into the graph, and revise the probabilities.

After absorbing the evidence about the female with high-cholesterol, the probability of heart-failure and coronary artery disease (CAD) changed very minimally after this information is considered. We can see this below from the tables of the joint, marginal and conditional probabilities before and after the information was considered.

Marginal probability table before absorbing the information for CAD and SuffHeartF:

```
$SuffHeartF
SuffHeartF
       No       Yes
0.7076271 0.2923729
```

```
$CAD
CAD
       No       Yes
0.5531269 0.4468731
```

Marginal probability table after absorbing the information for CAD and SuffHeartF:

```
$SuffHeartF
SuffHeartF
       No       Yes
0.7076271 0.2923729
```

```
$CAD
CAD
       No       Yes
0.5401298 0.4598702
```

Joint probability table before absorbing the information:

```
          CAD
SuffHeartF        No       Yes
       No  0.3957368 0.3118903
       Yes 0.1443930 0.1479799
```

Joint probability table after absorbing the information:

```
          CAD
SuffHeartF        No       Yes
       No  0.4078210 0.2998061
       Yes 0.1453059 0.1470670
```

Conditional probability table before absorbing the information:

```
      SuffHeartF
CAD          No       Yes
  No  0.7326698 0.2673302
  Yes 0.6782138 0.3217862
```

Conditional probability table after absorbing the information:

```
      SuffHeartF
CAD          No       Yes
  No  0.7373010 0.2626990
  Yes 0.6708976 0.3291024
```

Part C

i) In this part we had to simulate a new data set with 25 observations conditional upon this new information, as shown below.

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

We had to save this data and we later used this data and the “predict” function to estimate the probability of “Smoker” and “CAD” given the other variables in the model. We used the function “simulate.grain” in the gRain package.

Marginal probability of “Smoker” using 25 observations

```
  No  Yes
0.12 0.88
```

Marginal probability of “Smoker” from part b

```
Smoker
       No       Yes
0.3617021 0.6382979
```

Marginal Probability of “CAD” using 25 observations

```
  No  Yes
0.48 0.52
```

Marginal probability for CAD from part b

```
CAD
       No       Yes
0.5401298 0.4598702
```

From the tables above we can see that the probability of being a smoker or not and having CAD or not are different for the 25 observation dataset and the update distribution in part b. This is because we have a small dataset of observations and so it is not representative of the prediction given the other factors in the model.

Part C

ii) In this part we had to create a new data set, as done in part C, this time with 500 data points. A portion of the dataset is shown below:

	Sex	Smoker	SuffHeartF	Inherit	Hyperchol	CAD
1	Female	Yes	No	No	No	No
2	Female	Yes	No	No	Yes	Yes
3	Female	No	No	No	No	No
4	Female	Yes	Yes	Yes	Yes	Yes
5	Female	No	Yes	No	Yes	Yes
6	Female	No	Yes	No	Yes	No
7	Female	No	No	No	Yes	Yes
8	Female	Yes	No	Yes	Yes	Yes
9	Female	No	No	No	Yes	Yes
10	Female	No	No	No	No	No
11	Female	Yes	Yes	No	Yes	No
12	Female	No	No	No	No	No
13	Female	Yes	Yes	No	Yes	Yes
14	Female	No	No	No	Yes	Yes
15	Female	Yes	Yes	No	Yes	No
16	Female	Yes	No	No	Yes	No
17	Female	Yes	Yes	No	Yes	Yes
18	Female	Yes	No	Yes	No	Yes
19	Female	Yes	No	No	Yes	Yes
20	Female	Yes	Yes	Yes	Yes	No
21	Female	Yes	No	Yes	No	Yes
22	Female	Yes	No	No	No	No
23	Female	Yes	No	Yes	Yes	Yes
24	Female	No	No	No	Yes	No
25	Female	Yes	No	No	No	Yes
26	Female	No	Yes	No	No	No
27	Female	No	No	No	Yes	No
28	Female	No	No	No	No	No
29	Female	Yes	No	No	Yes	No
30	Female	Yes	No	No	Yes	Yes
31	Female	Yes	No	Yes	Yes	Yes
32	Female	No	No	No	No	No
33	Female	Yes	No	No	No	Yes
34	Female	Yes	No	No	No	No
35	Female	No	Yes	No	Yes	Yes

We had to save this data and we later used this data and the “predict” function to estimate the probability of “Smoker” and “CAD” given the other variables in the model.

Marginal probability of “Smoker” using 500 observations

```
   No   Yes
0.262 0.738
```

Marginal probability of “Smoker” from part b

```
Smoker
       No       Yes
0.3617021 0.6382979
```

Marginal Probability of “CAD” using 500 observations

```
    No    Yes
0.494 0.506
```

Marginal probability for CAD from part b

```
CAD
       No       Yes
0.5401298 0.4598702
```

From the tables above we can see that the probability of being a smoker or not and having CAD or not are similar and closer for the 500 observation dataset and the update distribution in part b, compared to the 25 observations in part c. This is because of sample size, as using 500 observations is more representative of the data distribution and gives a closer prediction to the original distribution we are inquiring about in part b.