2019

Homework 3

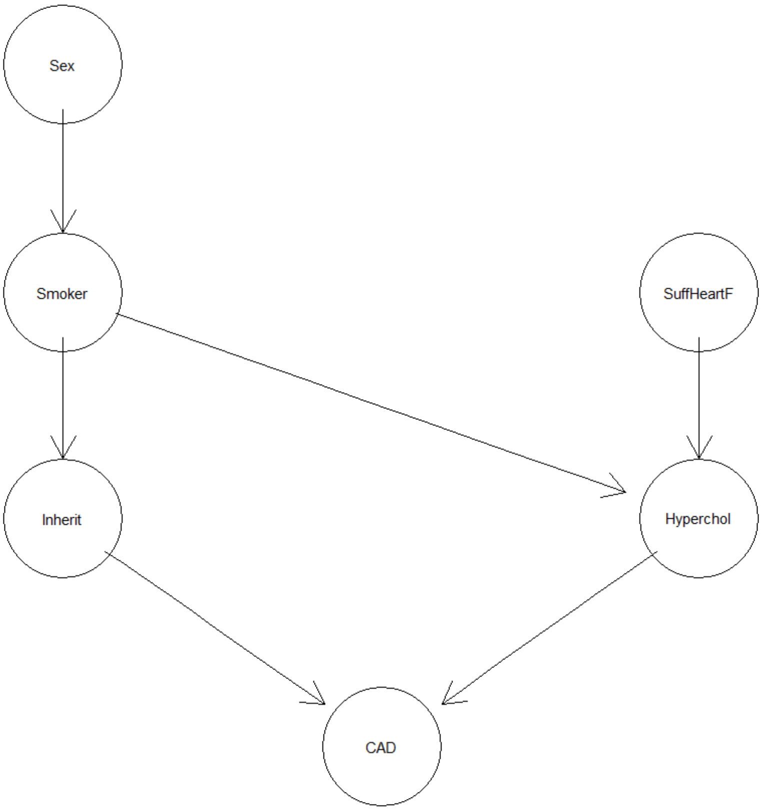
STATISTICAL DATA MINING II

RAVI TEJA SUNKARA
rsunkara
Ubit number: 50292191

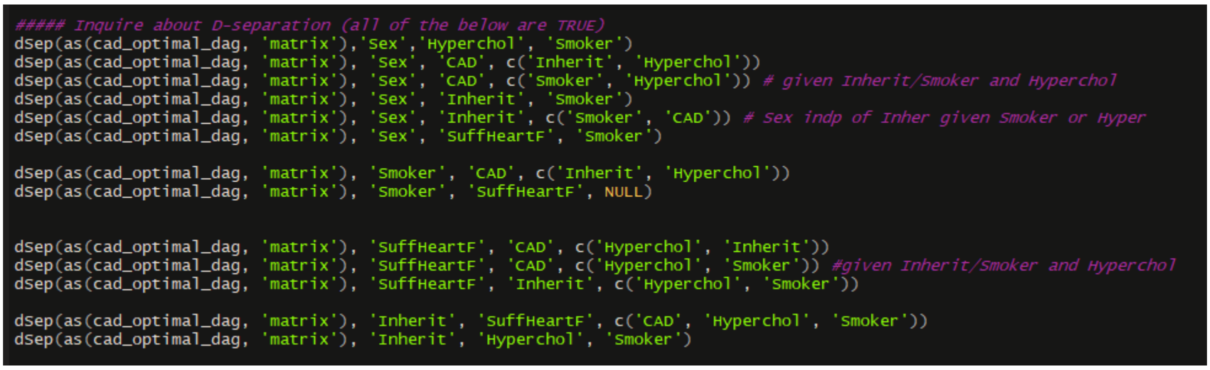
Solution of Question 1

Part - A

The plot of network:



D-Separations in Graph:



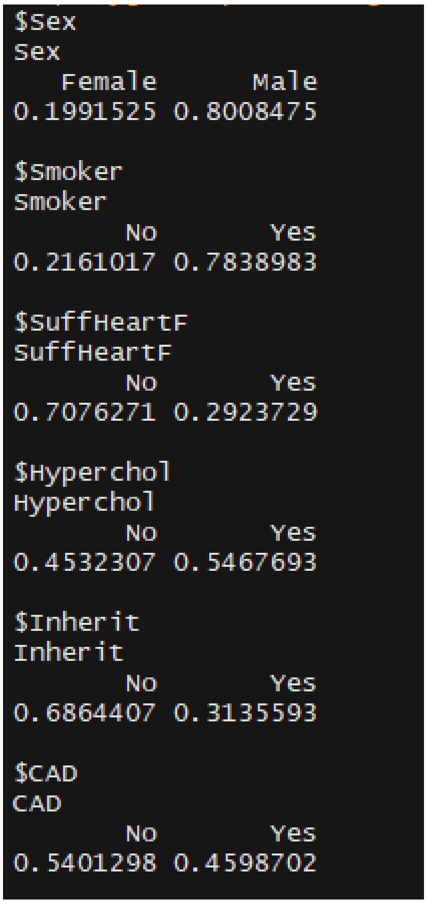
```
##### Inquire about D-separation (all of the below are TRUE)
dSep(as(cad_optimal_dag, 'matrix'),'Sex','Hyperchol', 'Smoker')
dSep(as(cad_optimal_dag, 'matrix'), 'Sex', 'CAD', c('Inherit', 'Hyperchol'))
dSep(as(cad_optimal_dag, 'matrix'), 'Sex', 'CAD', c('Smoker', 'Hyperchol')) # given Inherit/Smoker and Hyperchol
dSep(as(cad_optimal_dag, 'matrix'), 'Sex', 'Inherit', 'Smoker')
dSep(as(cad_optimal_dag, 'matrix'), 'Sex', 'Inherit', c('Smoker', 'CAD')) # Sex indp of Inher given Smoker or Hyper
dSep(as(cad_optimal_dag, 'matrix'), 'Sex', 'SuffHeartF', 'Smoker')

dSep(as(cad_optimal_dag, 'matrix'), 'Smoker', 'CAD', c('Inherit', 'Hyperchol'))
dSep(as(cad_optimal_dag, 'matrix'), 'Smoker', 'SuffHeartF', NULL)


dSep(as(cad_optimal_dag, 'matrix'), 'SuffHeartF', 'CAD', c('Hyperchol', 'Inherit'))
dSep(as(cad_optimal_dag, 'matrix'), 'SuffHeartF', 'CAD', c('Hyperchol', 'Smoker')) #given Inherit/Smoker and Hyperchol
dSep(as(cad_optimal_dag, 'matrix'), 'SuffHeartF', 'Inherit', c('Hyperchol', 'Smoker'))

dSep(as(cad_optimal_dag, 'matrix'), 'Inherit', 'SuffHeartF', c('CAD', 'Hyperchol', 'Smoker'))
dSep(as(cad_optimal_dag, 'matrix'), 'Inherit', 'Hyperchol', 'Smoker')
```

- Sex is **independent** of Hyperchol given Smoker
- Sex is **independent** of CAN given Hyperchol and Inherit
- Sex is **independent** of CAN given Hyperchol and Smoker
- Sex is **independent** of Inherit given Smoker
- Sex is **independent** of Inherit given Smoker and CAD
- Sex is **independent** of SuffHeartF given Smoker
- Smoker is **independent** of CAD given Inherit and Hyperchol

- Smoker is **independent** of SuffHeartF
- SuffHeartF is **independent** of CAD given Hyperchol and Inherit
- SuffHeartF is **independent** of CAD given Hyperchol and Smoker
- SuffHeartF is **independent** of Inherit given Hyperchol and Smoker
- Inherit is **independent** of SuffHeartF given CAD, Hyperchol and Smoker
- Inherit is **independent** of Hyperchol given smoker

Conditional Probability Tables:

Conditional Probability Tables were defined using ‘extractCPT’ function in R from ‘gRain’ package.

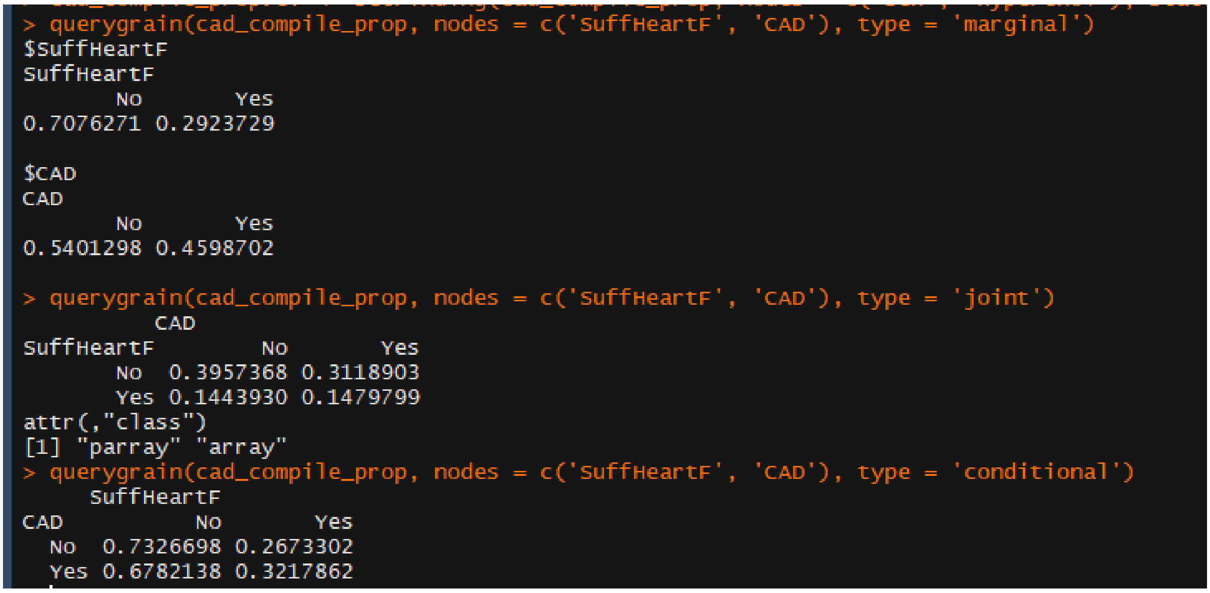


```
$Sex
Sex
   Female      Male
0.1991525 0.8008475

$Smoker
Smoker
       No       Yes
0.2161017 0.7838983

$SuffHeartF
SuffHeartF
       No       Yes
0.7076271 0.2923729

$Hyperchol
Hyperchol
       No       Yes
0.4532307 0.5467693

$Inherit
Inherit
       No       Yes
0.6864407 0.3135593

$CAD
CAD
       No       Yes
0.5401298 0.4598702
```

Part - B

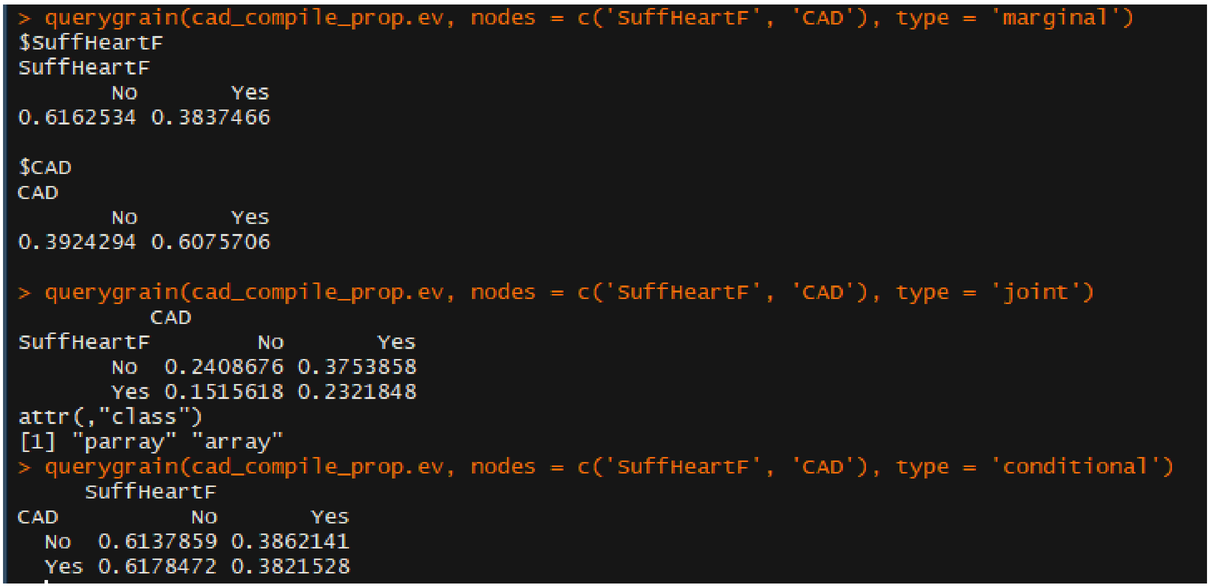
Built the network, compiled it and then propagate it. After adding the new observation, the change in probabilities was found.

__Before Absorbing evidence:__



```
> querygrain(cad_compile_prop, nodes = c('SuffHeartF', 'CAD'), type = 'marginal')
$SuffHeartF
SuffHeartF
       No       Yes 
0.7076271 0.2923729 

$CAD
CAD
       No       Yes 
0.5401298 0.4598702 

> querygrain(cad_compile_prop, nodes = c('SuffHeartF', 'CAD'), type = 'joint')
          CAD
SuffHeartF        No       Yes
       No  0.3957368 0.3118903
       Yes 0.1443930 0.1479799
attr(,"class")
[1] "parray" "array" 
> querygrain(cad_compile_prop, nodes = c('SuffHeartF', 'CAD'), type = 'conditional')
     SuffHeartF
CAD          No       Yes
  No  0.7326698 0.2673302
  Yes 0.6782138 0.3217862
```

__After Absorbing the evidence:__

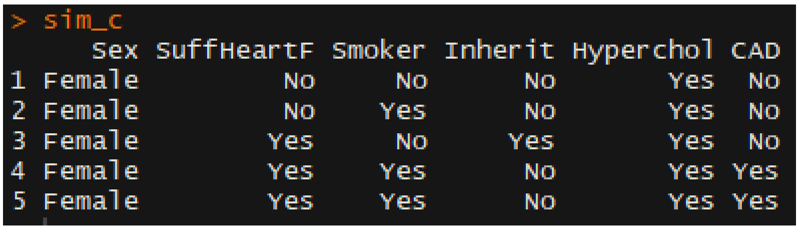


```
> querygrain(cad_compile_prop.ev, nodes = c('SuffHeartF', 'CAD'), type = 'marginal')
$SuffHeartF
SuffHeartF
       No       Yes 
0.6162534 0.3837466 

$CAD
CAD
       No       Yes 
0.3924294 0.6075706 

> querygrain(cad_compile_prop.ev, nodes = c('SuffHeartF', 'CAD'), type = 'joint')
          CAD
SuffHeartF        No       Yes
       No  0.2408676 0.3753858
       Yes 0.1515618 0.2321848
attr(,"class")
[1] "parray" "array" 
> querygrain(cad_compile_prop.ev, nodes = c('SuffHeartF', 'CAD'), type = 'conditional')
     SuffHeartF
CAD          No       Yes
  No  0.6137859 0.3862141
  Yes 0.6178472 0.3821528
```

- After absorbing the evidence, the probability of Heart Failure and Coronary Heart Disease (CAD) increases.
- The probability of Heart Failure increases by 31.5%
- The probability of Coronary Heart Disease increases by 32%

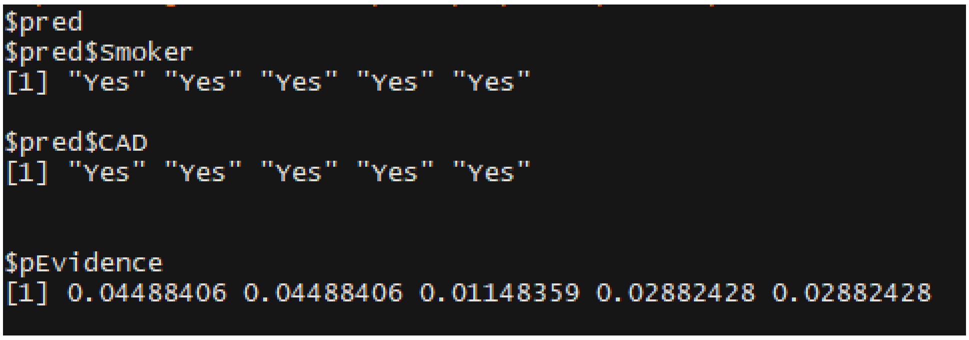
Part – C

The new dataset with 5 observations generated using simulate function is:



```
> sim_c
     Sex SuffHeartF Smoker Inherit Hyperchol CAD
1 Female         No     No      No       Yes  No
2 Female         No    Yes      No       Yes  No
3 Female        Yes     No     Yes       Yes  No
4 Female        Yes    Yes      No       Yes Yes
5 Female        Yes    Yes      No       Yes Yes
```

The predictions are:

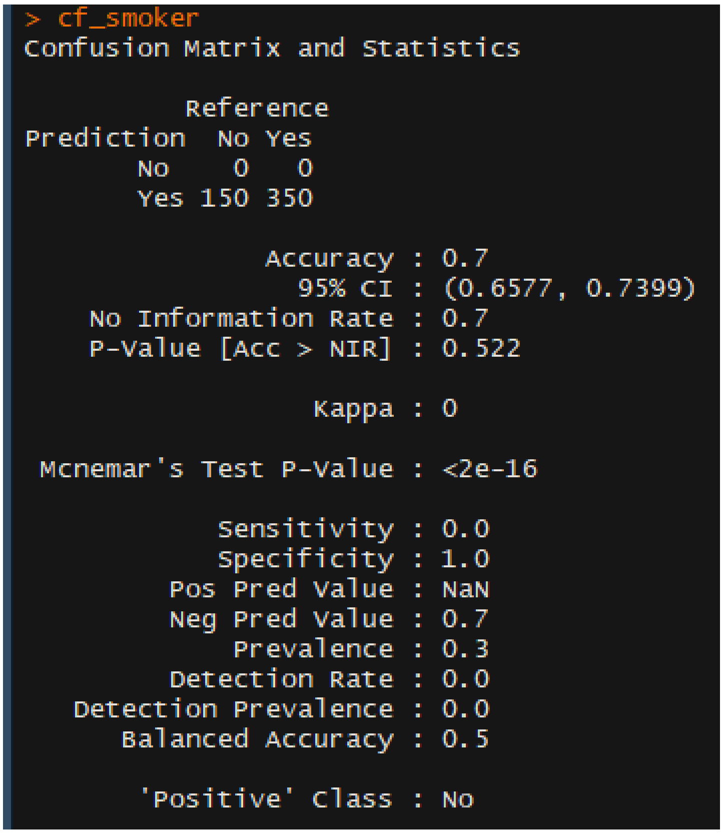


```
$pred
$pred$Smoker
[1] "Yes" "Yes" "Yes" "Yes" "Yes"

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


$pEvidence
[1] 0.04488406 0.04488406 0.01148359 0.02882428 0.02882428
```

Part - D

Simulated a new dataset with 500 observations. Saved this as ‘q1_simulate_d.txt’. Estimated the probabilities of ‘Smoker’ and ‘CAD’ given other variables in the data.

Confusion Matrix of Smoker:

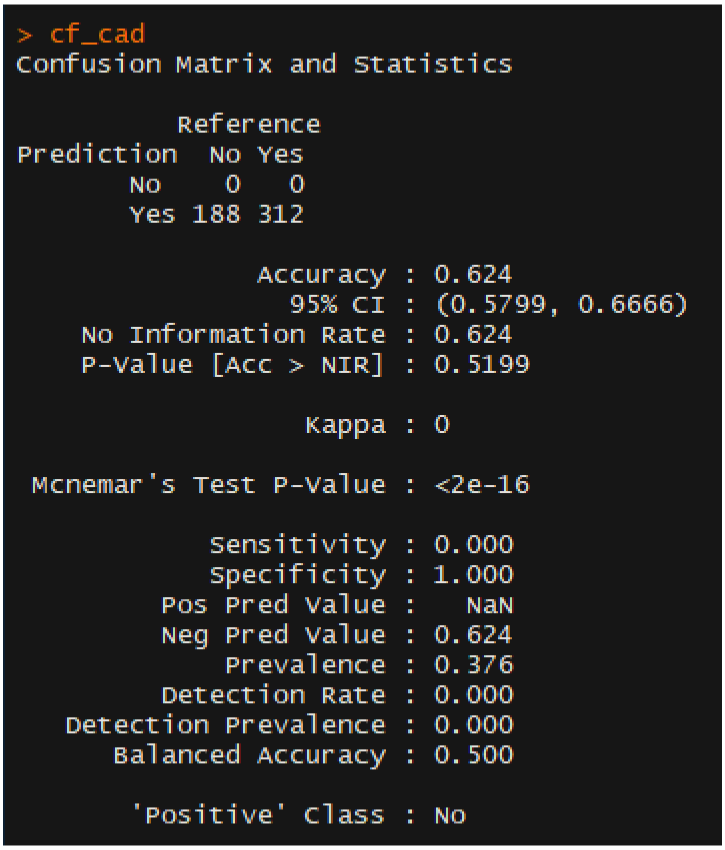


```
> cf_smoker
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   0
       Yes 150 350

               Accuracy : 0.7
                 95% CI : (0.6577, 0.7399)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.522

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0
            Specificity : 1.0
         Pos Pred Value : NaN
         Neg Pred Value : 0.7
             Prevalence : 0.3
         Detection Rate : 0.0
   Detection Prevalence : 0.0
      Balanced Accuracy : 0.5

       'Positive' Class : No
```

- Misclassification rate for Smoker – 30%

Confusion Matrix of CAD:

- Misclassification rate for CAD – 37.6%



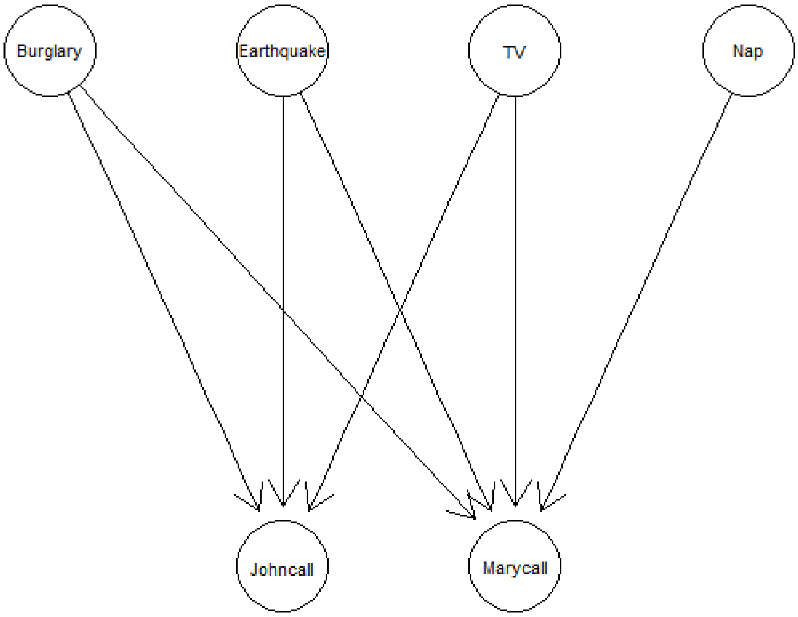
```
> cf_cad
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   0
       Yes 188 312
                                          
               Accuracy : 0.624           
                 95% CI : (0.5799, 0.6666)
    No Information Rate : 0.624           
    P-Value [Acc > NIR] : 0.5199          
                                          
                  Kappa : 0               
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.000           
            Specificity : 1.000           
         Pos Pred Value :   NaN           
         Neg Pred Value : 0.624           
             Prevalence : 0.376           
         Detection Rate : 0.000           
   Detection Prevalence : 0.000           
      Balanced Accuracy : 0.500           
                                          
       'Positive' Class : No              
```

Comment:

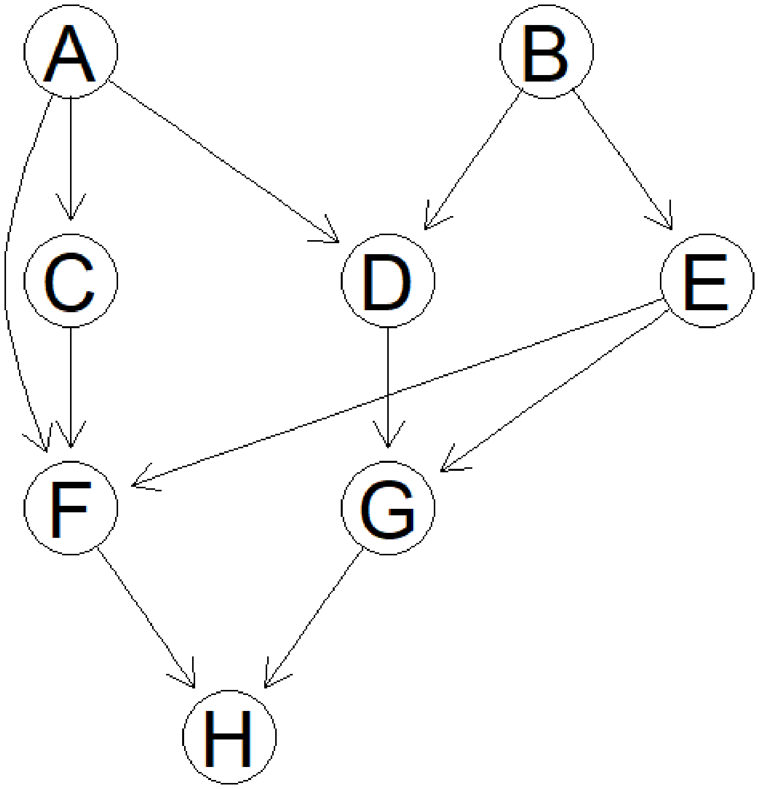
- By observing misclassification rate we can infer that the network is performing well. The accuracy is nearly 25% better than a random guess accuracy of 50%.
- It can be further improved by conditioning the probability of nodes properly, by forming a strong Bayesian Network that is, hierarchically ordering the nodes.
- Different versions could be tried to arrive at the best performing network.

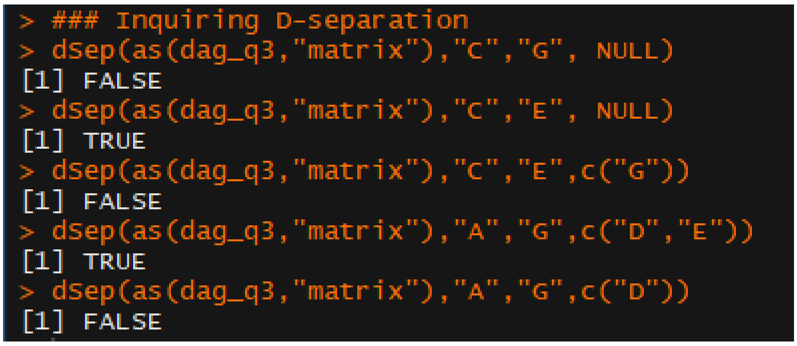
Solution of Question 2

Bayesian network obtained is:


Burglary
Earthquake
TV
Nap
Johncall
Marycall

Solution of Question 3





```
> ### Inquiring D-separation
> dSep(as(dag_q3,"matrix"),"C","G", NULL)
[1] FALSE
> dSep(as(dag_q3,"matrix"),"C","E", NULL)
[1] TRUE
> dSep(as(dag_q3,"matrix"),"C","E",c("G"))
[1] FALSE
> dSep(as(dag_q3,"matrix"),"A","G",c("D","E"))
[1] TRUE
> dSep(as(dag_q3,"matrix"),"A","G",c("D"))
[1] FALSE
```

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