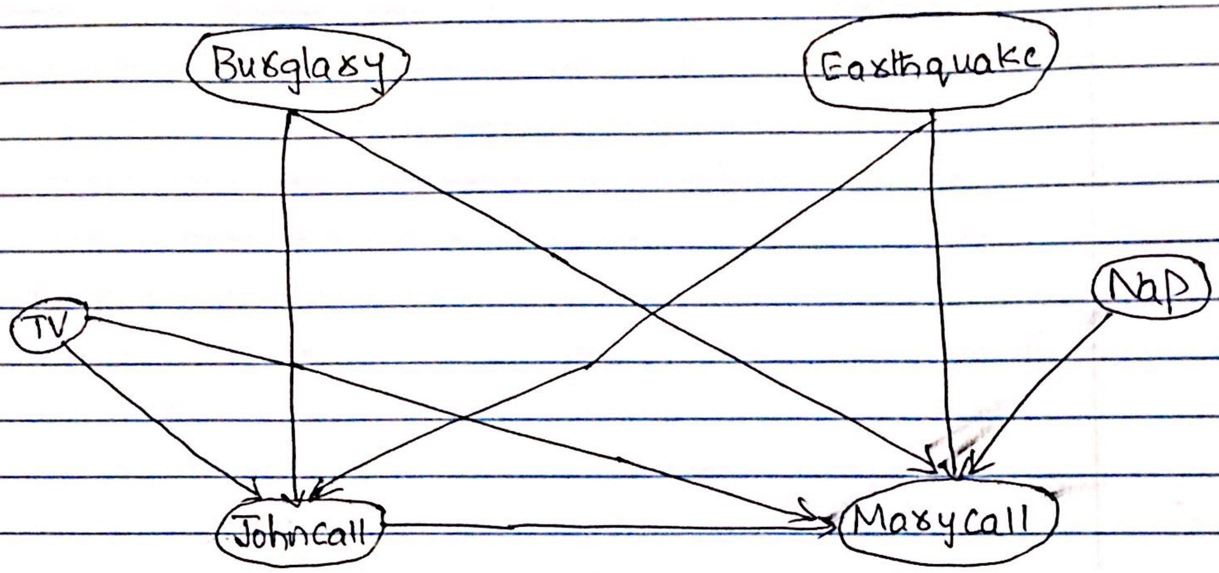
UBIT - likhithk

STA Home Work-4

2) The Bayesian network over all the nodes except for Alarm that is a minimal I-map for the marginal distribution is :



→ Here, it is drawn without considering alarm as evidence.

- Burglary and Earthquake are connected to Johncall and Mary call instead of Alarm.

- Johncall is connected to Marycall.

- Marycall is also connected to Nap as per the original graph.

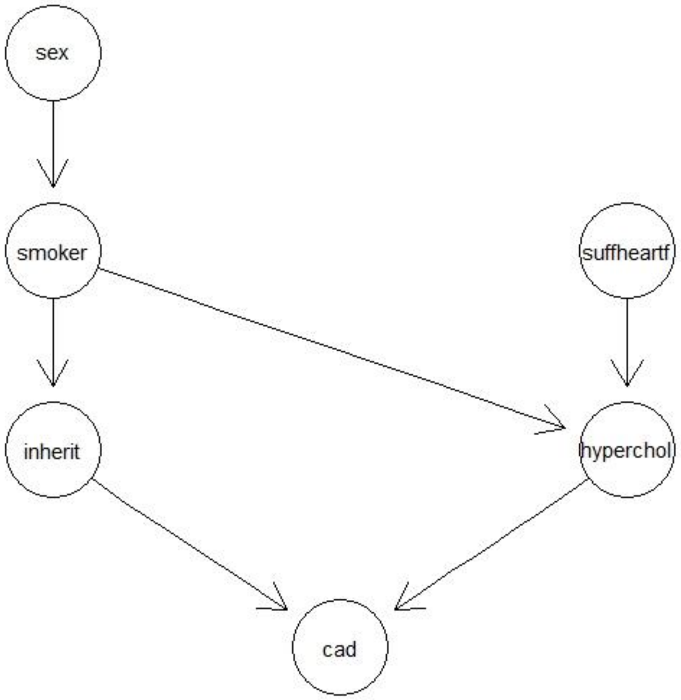
- Mary call is connected to TV in making the assumption that she did not hear the alarm due to perturbance of TV.

- Johncall (John) is not connected to Nap cause he has to be awake to make the call.

Scanned by CamScanner

STA Home Work – 4

1) Install the required packages and load the cad1 dataset from the ‘gRbase’ package. These dataset consists of 236 observations of the 14 following variables. Create a dag using daglist function for the optimal network as given in the question, which is:



Inference: This is the directed acyclic graph for the optimal network that has only 6 nodes which are taken from the cad1 dataset.

a) After constructing the network in R, infer the conditional probabilities:

- Calculate the number of times a factor value is repeating in each node predictor from cad1 dataset.
- While creating cptables, values would be the ones that were obtained from above step.
- Use compile cpt function to create a list by combining results of all cptable and ortable functions.
- From the created list, conditional probabilities are obtained.

- Sex

```
     sex
male female
 0.8    0.2
```

- Smoker | sex

```
                  sex
smoker      male       female
  yes   0.7838983   0.8008475
  no    0.2161017   0.1991525
```

- SuffHeartF

```
          suffheartf
            yes   no
           0.29 0.71
```

- Inherit | smoker

```
                          smoker
   inherit            yes            no
     yes        0.5221239   0.7838983
     no         0.4778761   0.2161017
```

- Hyperchol | SuffHeartF:smoker

```
 smoker = yes
              suffheartf
hyperchol   yes no
     yes      1    1
     no       0    0

   smoker = no
              suffheartf
hyperchol   yes no
     yes      1    0
     no       0    1
```

- CAD|inherit:Hyperchol

```
hyperchol = yes
         inherit
   cad   yes   no
    yes   1     1
    no    0     0
hyperchol = no
          inherit
   cad    yes   no
    yes   1      0
    no    0       1
```

- Yes, d-separations were identified for some of the combinations. One example is ‘inherit’ and ‘suffheartf’.

b) After all the above process is done, compile the existing plist, propagate it and calculate probabilities using query grain function. The obtained joint probability for variables suffheartf and cad is

```
                          cad
 suffheartf        yes               no
      yes       0.2900000     0.1068771
      no        0.6773631     0.03263685
```

After the evidence is taken into consideration i.e; sex is female and high cholesterol is yes, the joint probability obtained is

```
                    cad
suffheartf      yes       no
      yes   0.3377585    0
      no    0.6622415    0
```

Here, we can clearly observe that there are zero persons without coronary artery disease when we consider high cholesterol.

c) Simulate the data with only five observations using simulate function and based on previous evidence.

```
$pred

$pred$smoker
           yes         no
[1,] 1.0000000  0.0000000
[2,] 0.8989127  0.1010873
[3,] 1.0000000  0.0000000
[4,] 1.0000000  0.0000000
[5,] 0.7281437  0.2718563

$pred$cad
      Yes   no
[1,]    1     0
[2,]    1     0
[3,]    1    0
[4,]    1    0
[5,]    1    0

$pEvidence
[1] 0.05434423 0.02469310 0.05937611 0.05937611 0.03330690
```

This is the output obtained for predict function for the simulated data set of five observations based on the new evidence.

d) Now, simulate the dataset with 500 observations based on the given evidence and predict the probabilities of smoker and cad given the other variables in the model.

The misclassification rate for cad is 35/500.

```
table(predict_500_class$pred$cad, sim.find_500$cad)
         yes   no
 no      19     0
 yes    465   16
```

The misclassification rate for smoker is 142/500.

```
table(predict_500_class$pred$smoker, sim.find_500$smoker)
          yes    no
 no        60     18
 yes     340     82
```

Conclusion: The performance of the underlying network is good as the size of the clique is small and there are not many variables, if it is large it would slow down and also effects the gRain package. This network of cad data seems optimal as we can see from the misclassification rate, no changes are needed to improve it.

3) Considering the given directed acyclic graph, using the dSep function, the results for the given nodes are determined as follows in terms of true or false,

a) C and G nodes are d-separated - False (They are not d-separated).

b) C and E nodes are d-separated - True (They are d-separated).

c) C and E are d-connected given evidence about G – False (They are d-connected).

d) A and G are d-connected given evidence about D and E – True (They are not d-connected).

e) A and G are d-connected given evidence on D – False (They are d-connected).