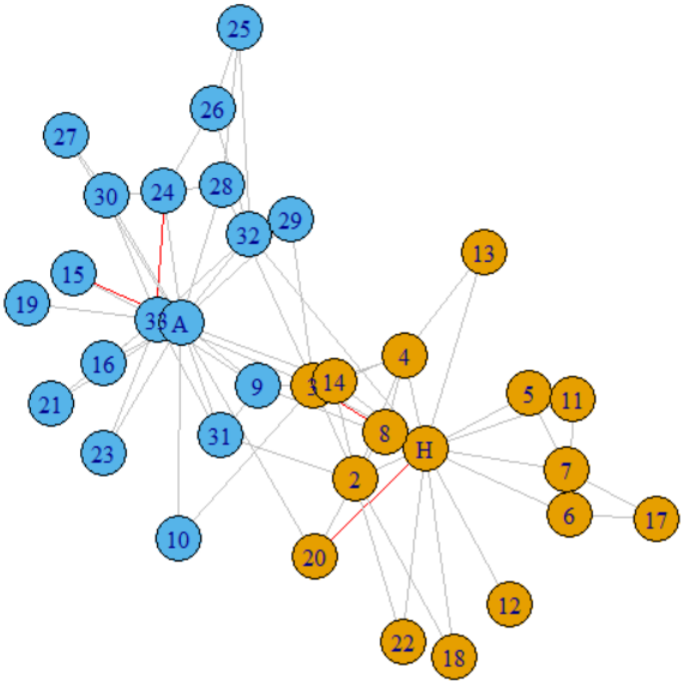
1)

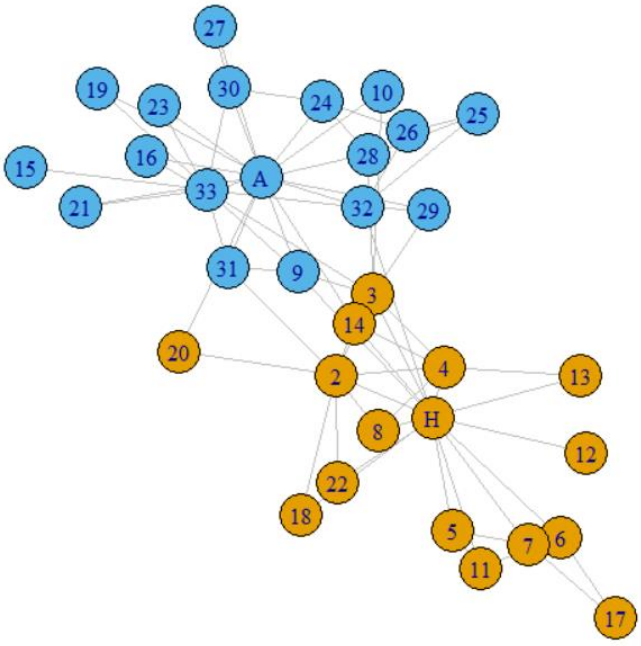
(a) The complete karate graph where the 4 red edges are the random 5% that are deleted to create the noisy dataset.



The edges to be removed -

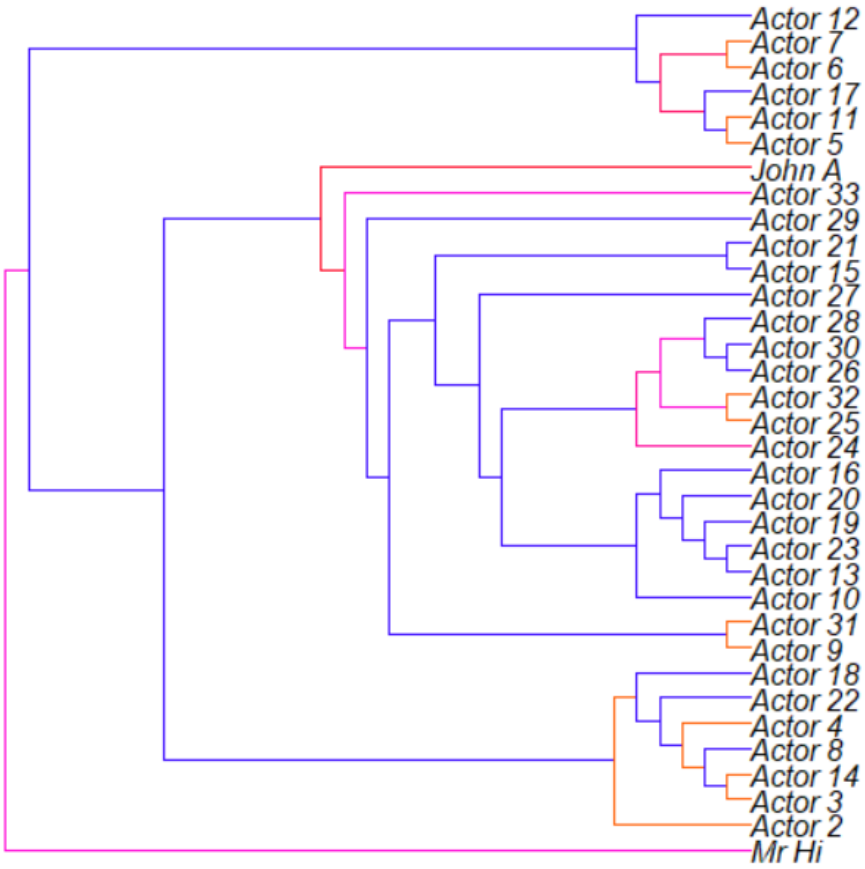
```
+ 4/78 edges from 4b458a1 (vertex names):
[1] Actor 3 --Actor 8  Actor 15--John A
[3] Mr Hi   --Actor 20 Actor 24--Actor 33
```

The graph after removing the edges -

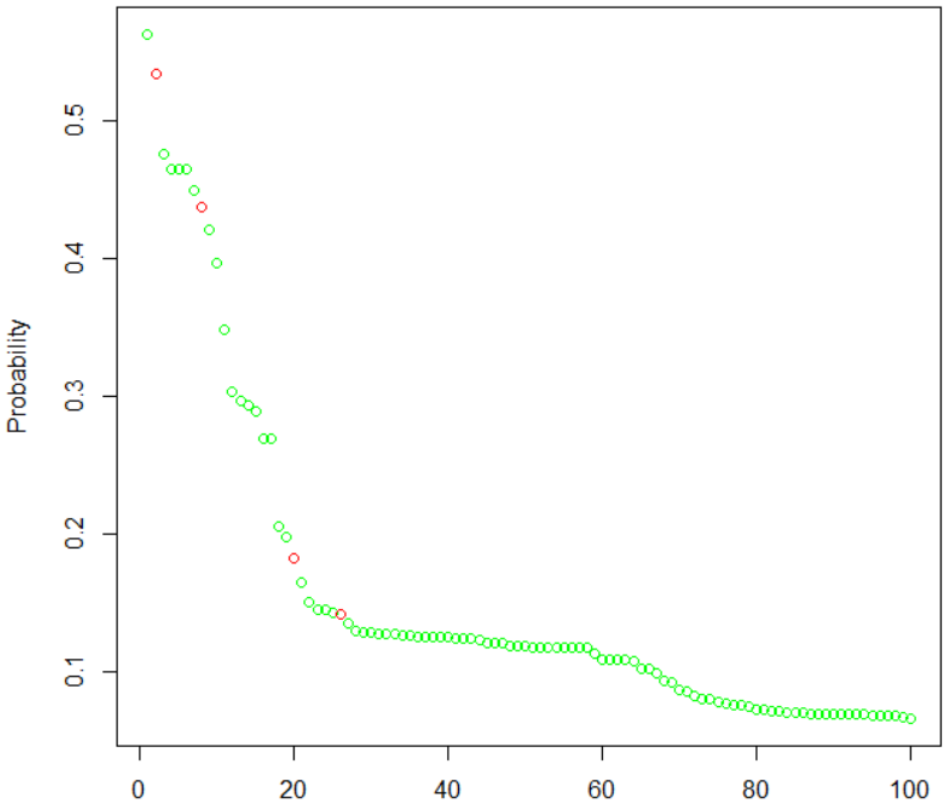


Although the graph looks different in appearance, it clearly has the correct structure.

Dendrogram after performing MCMC on the noisy dataset -

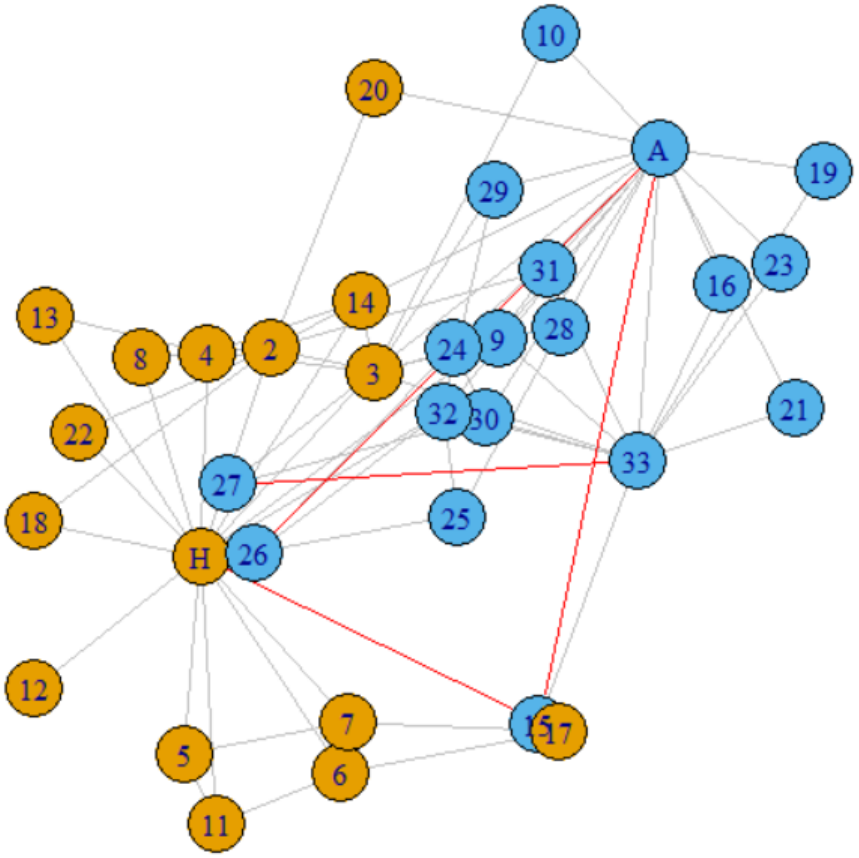


Probability of the top 100 edges predicted with the removed edges in red –



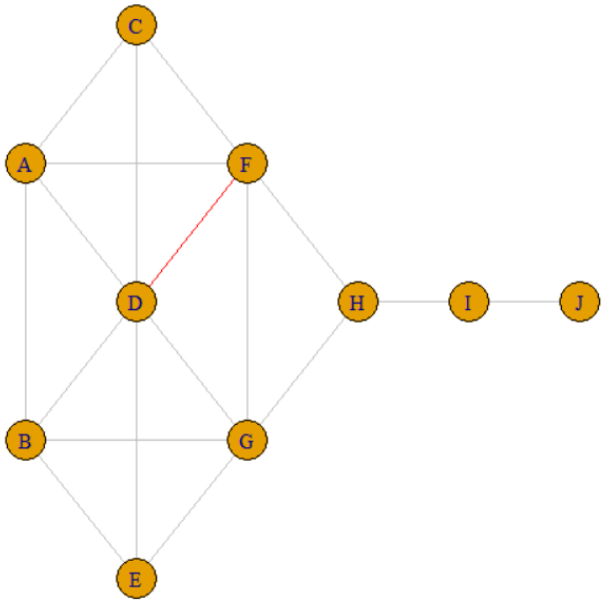
The removed edges are the 2nd,8th,20th and 26th predicted edges with a probability of 0.534, 0.437, 0.182 and 0.142 respectively. We are able to predict 2 of the edges rather well while all 4 of the removed edges are in the top 30 predictions.

The graph after adding the top 4 predicted edges –



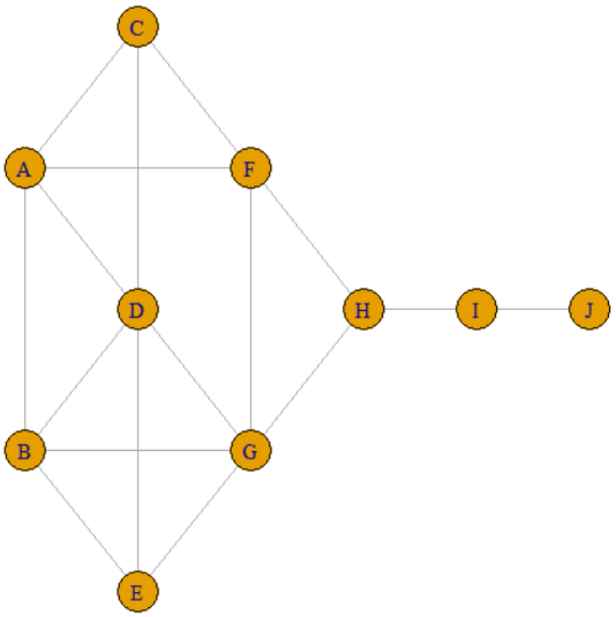
The graph still retains elements of its structure before deleting the edges and the separation between the two factions is still comprehensible.

(b) The complete kite graph where the 1 red edge is the random 5% that are deleted to create the noisy dataset.

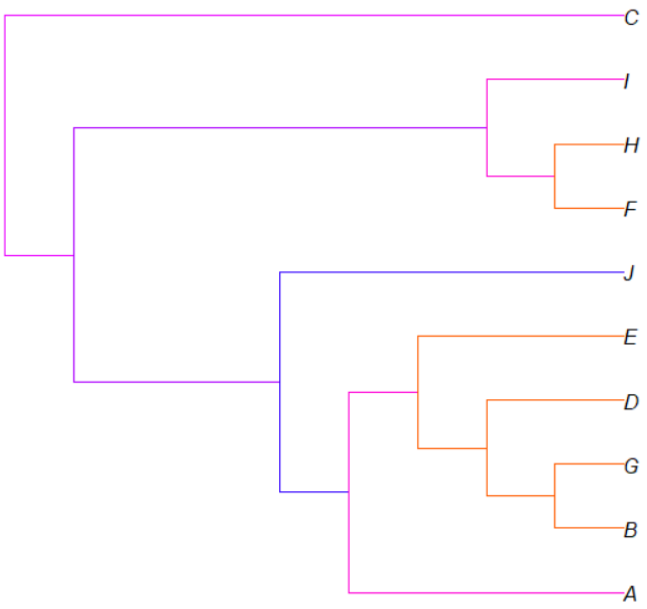


The edge to be removed is D -- F

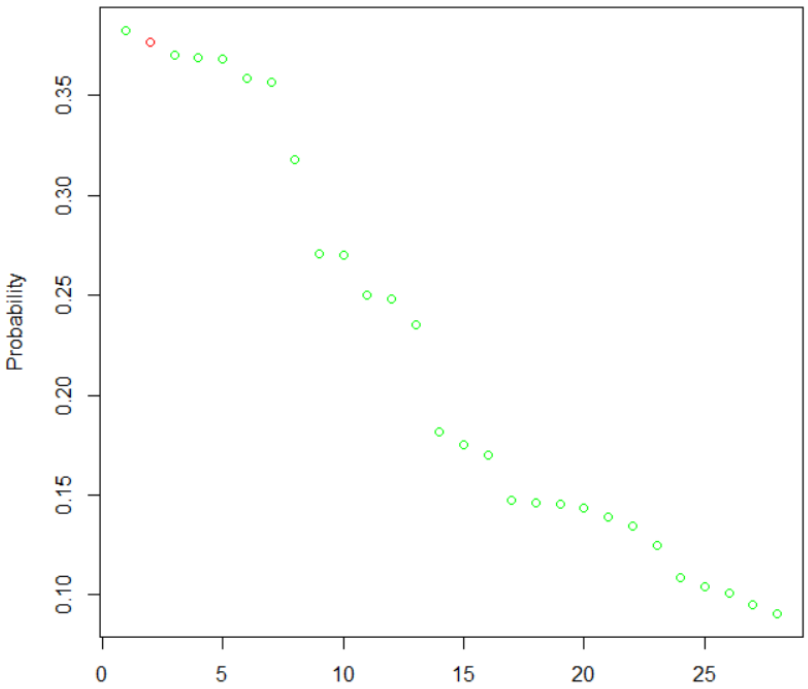
The graph after removing the edge –



Dendrogram after performing MCMC on the noisy dataset –

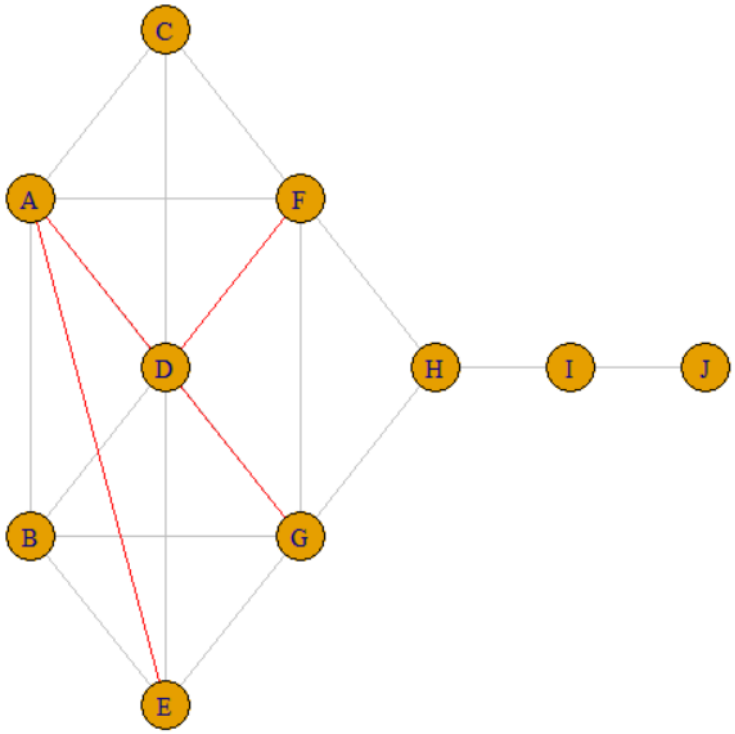


Probability of the edges predicted with the removed edge in red –



The removed edge is the 2nd predicted edge with a probability of 0.376 and is fairly well predicted.

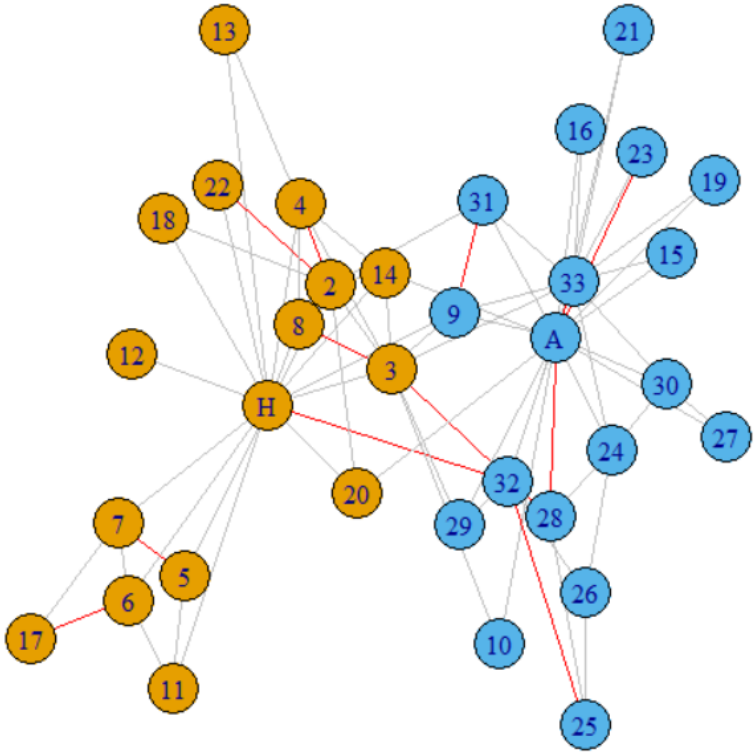
The graph after adding the top 3 predicted edges –



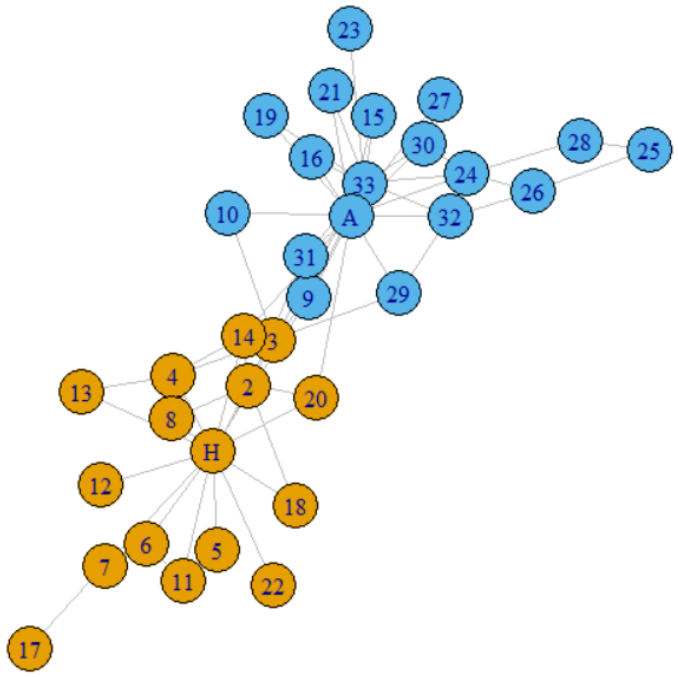
The graph's structure resembles that of the original indicating that deleting 5% of the edges and predicting them will result in a similar graph likely due to the small size of this particular graph.

(c) Karate – 15%

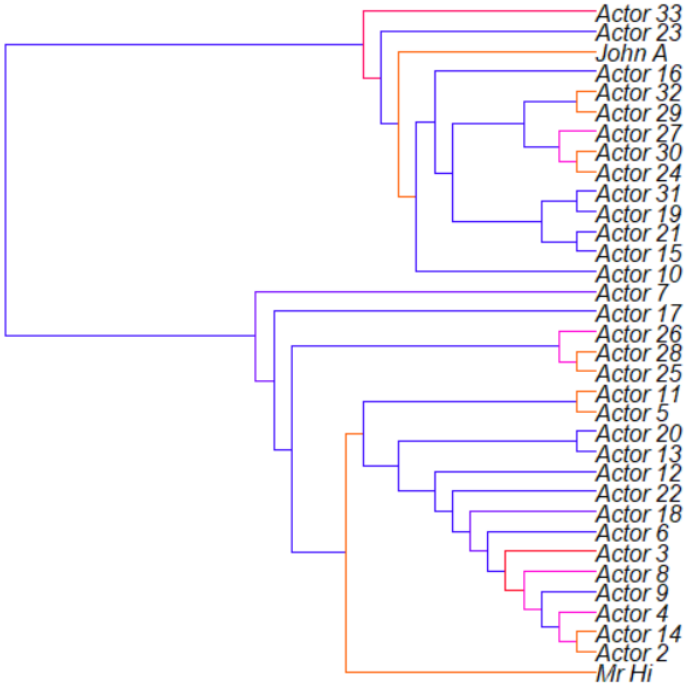
The complete karate graph where the 12 red edges are the random 15% that are deleted to create the noisy dataset–



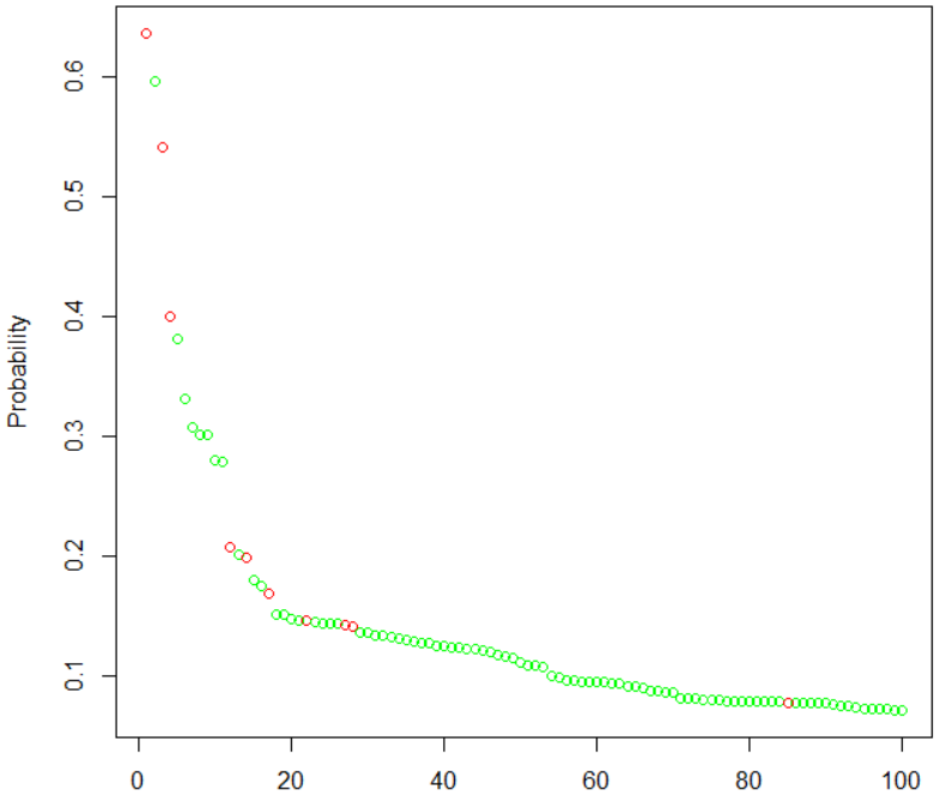
The graph after removing the 12 edges –



Dendrogram after performing MCMC on the noisy dataset-

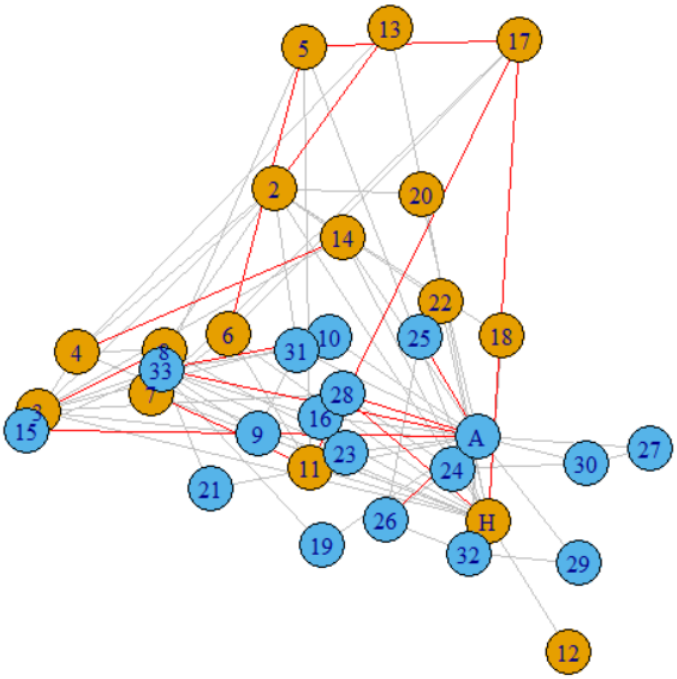


Probability of the edges predicted with the removed edges in red-



10 of the 12 removed edges are the 1st,3rd,4th,12th,14th,17th,22nd,27th,28th and 85th of the top 100 predicted edges with probabilities as shown in the graph. 3 of the 12 edges are predicted very well with 10 of the edges in the top 90 predictions but 2 of the edges are not even in the top 100 predictions.

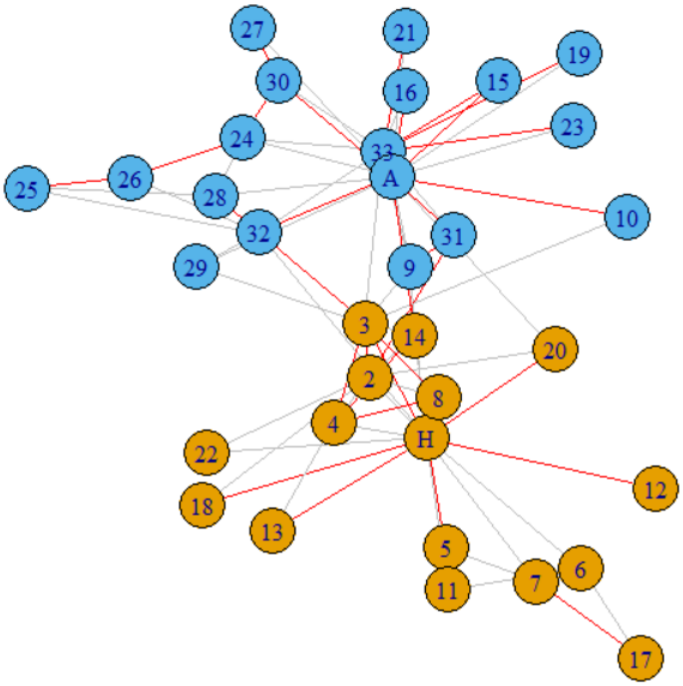
The graph after adding 15 of the top predicted edges -



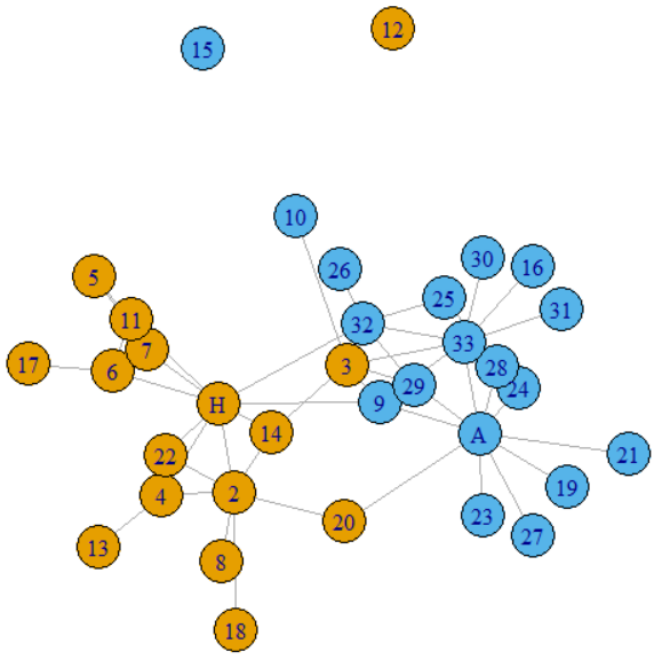
It is evident that removing 15% of the edges and adding 15 predicted edges has made the separation between the two factions hard to comprehend and significantly changed the graph from the original.

Karate – 40%

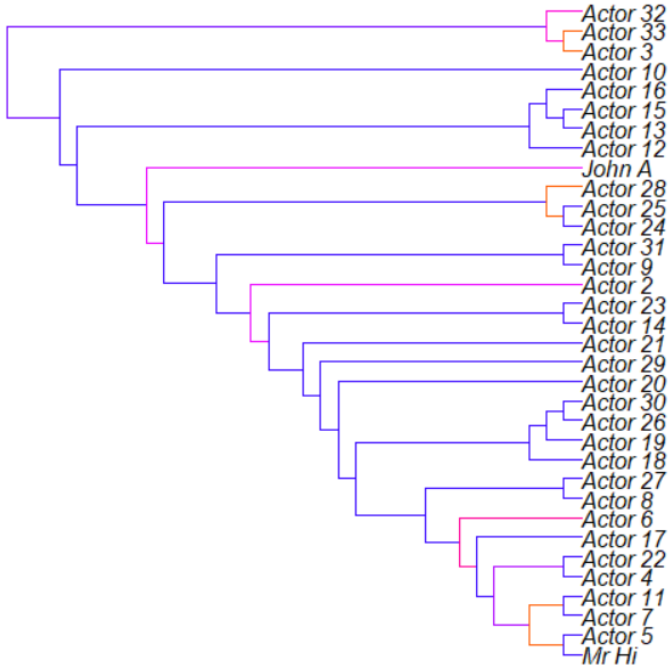
The complete graph where the 31 red edges are the 40% to be deleted to create the noisy dataset



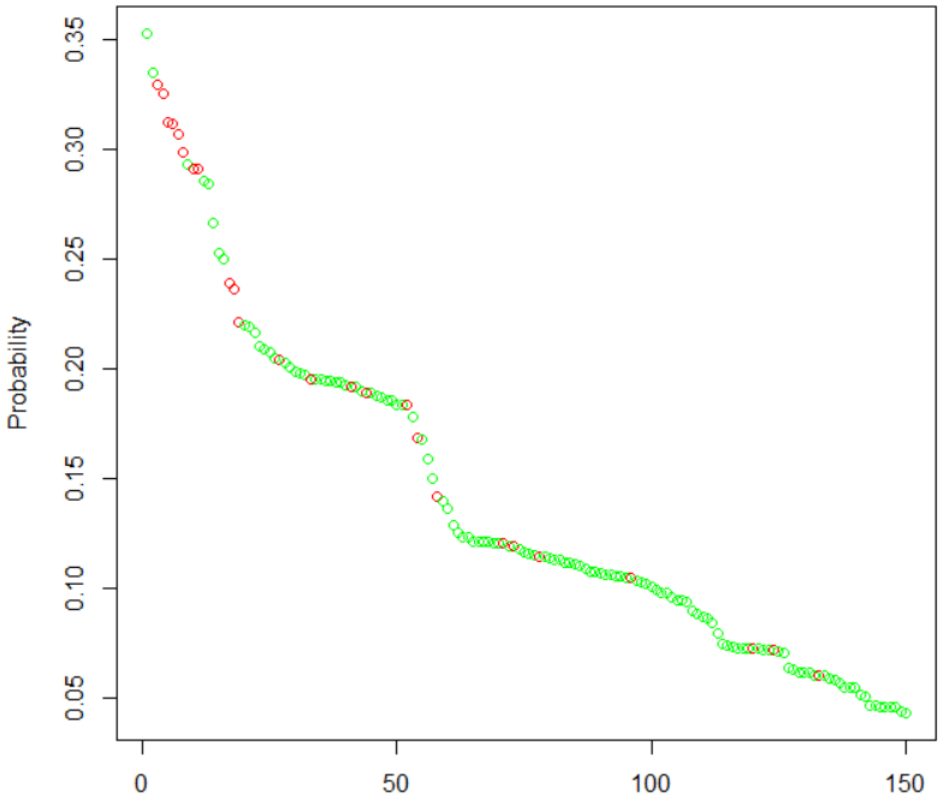
The graph after removing the edges –



Dendrogram after performing MCMC on the noisy dataset-

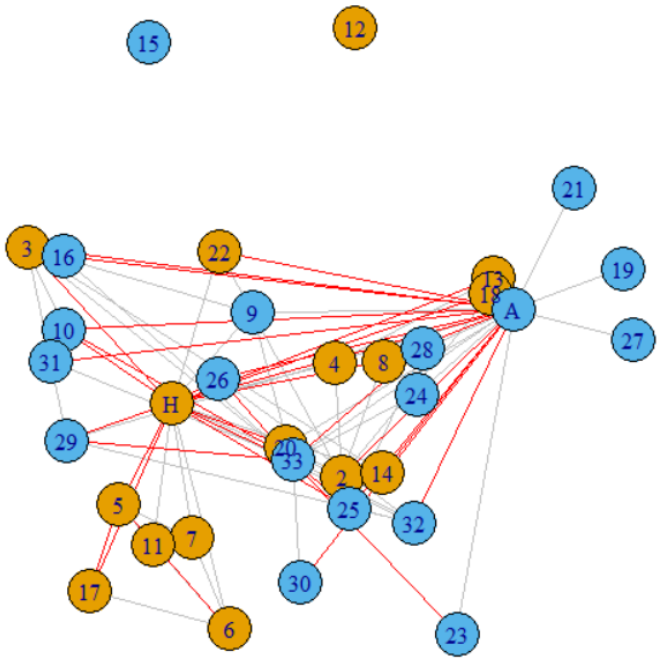


Probability of edges predicted with the edges removed in red -



25/31 removed edges are within top 150 of the predictions with 8 of the removed edges being predicted rather well. Interestingly, if the removed edges had a connection with Mr. Hi or John A they seem to show a relatively higher probability of being predicted.

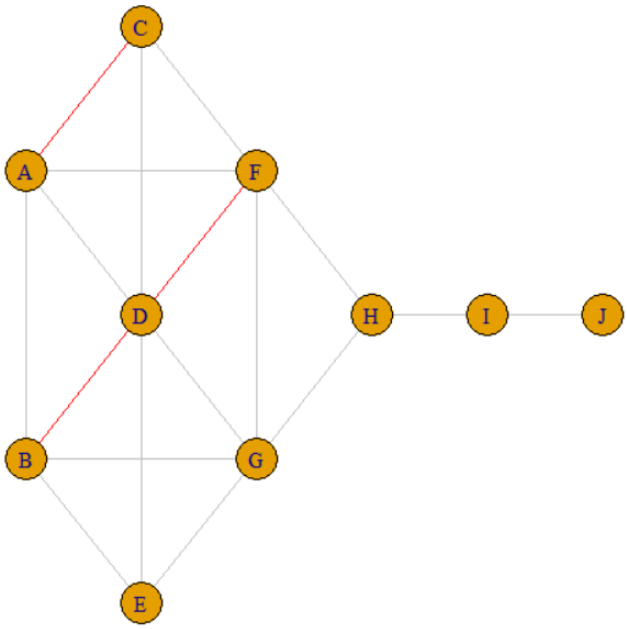
Graph after adding the top 35 edge prediction –



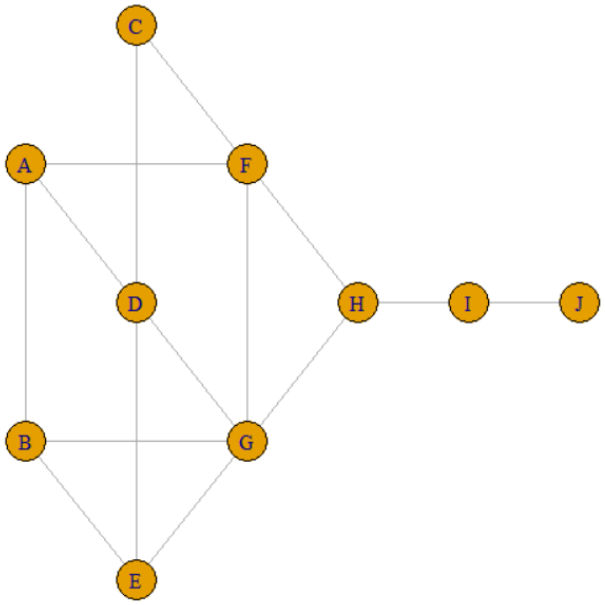
Similar to 15%, it is evident that removing 40% of the edges and adding according to prediction has made the separation between the two factions much harder to comprehend while leaving a few nodes unconnected from the rest of the graph.

Kite – 15%

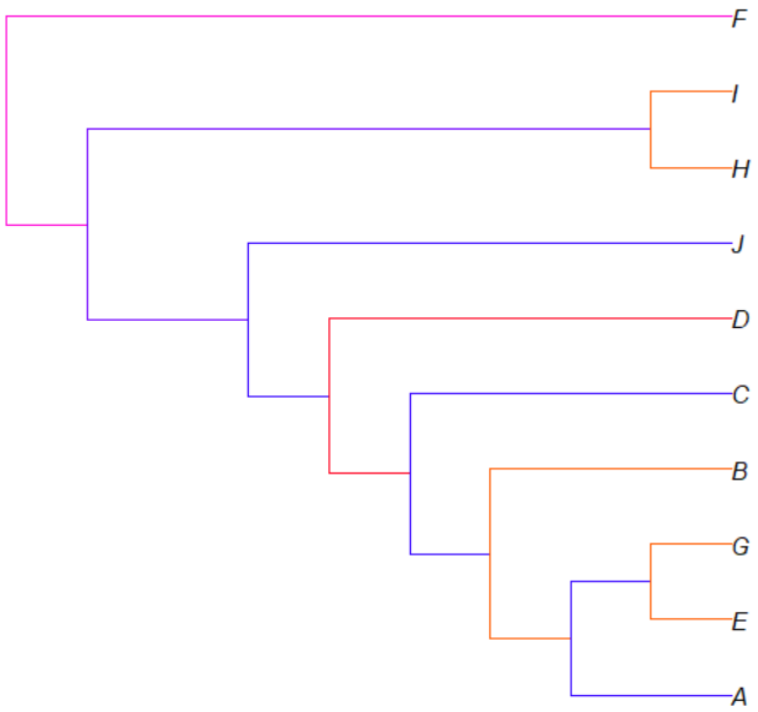
The complete kite graph where the 3 red edges are the 15% edges that are removed to create the noisy dataset –



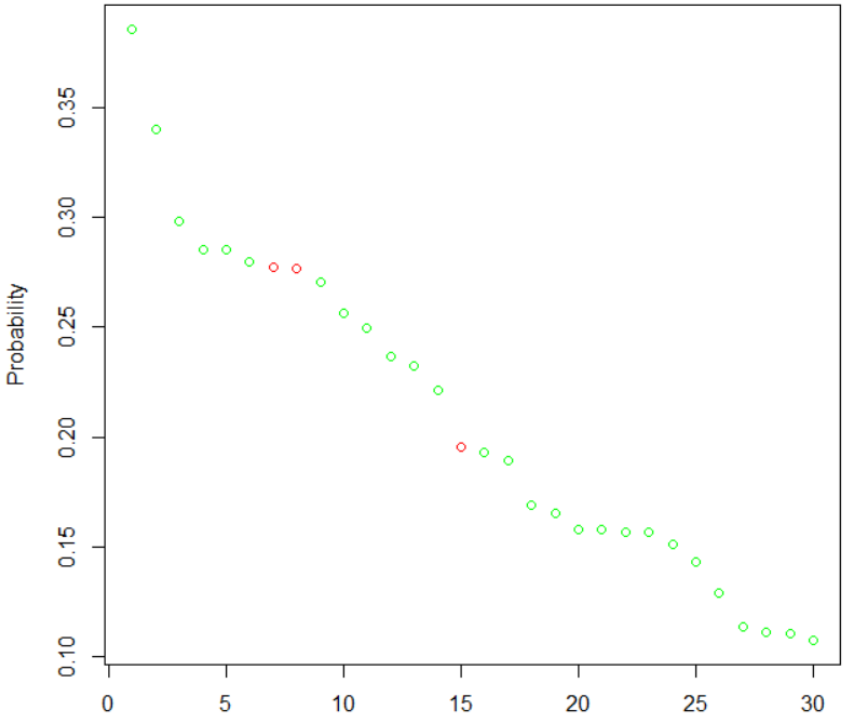
The graph after removing the edges –



The dendrogram after performing MCMC on the noisy database-

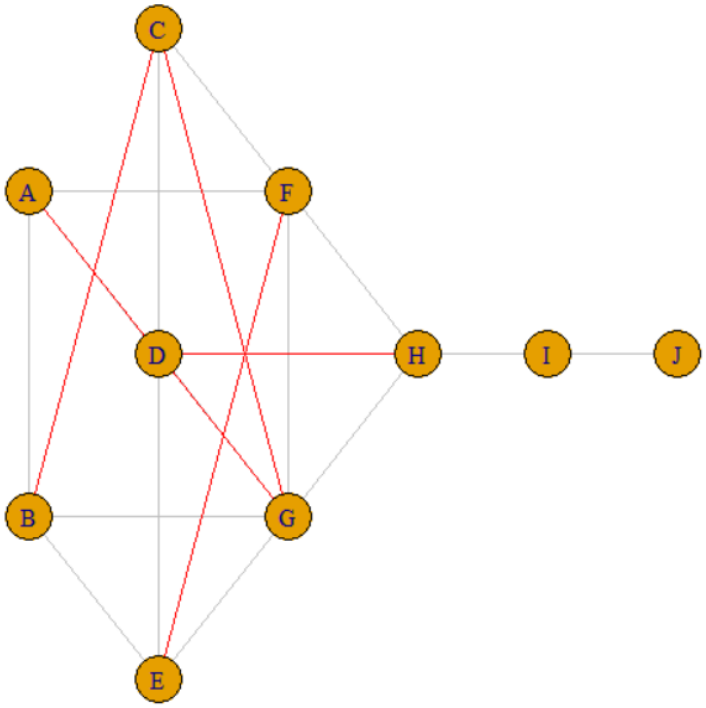


Probability of the predicted edges with the removed edges (D—F,B—D,A—C) in red –



2 of the edges are the 7th and 8th predictions out of 30 which indicates that none of the 3 edges are being predicted satisfactorily.

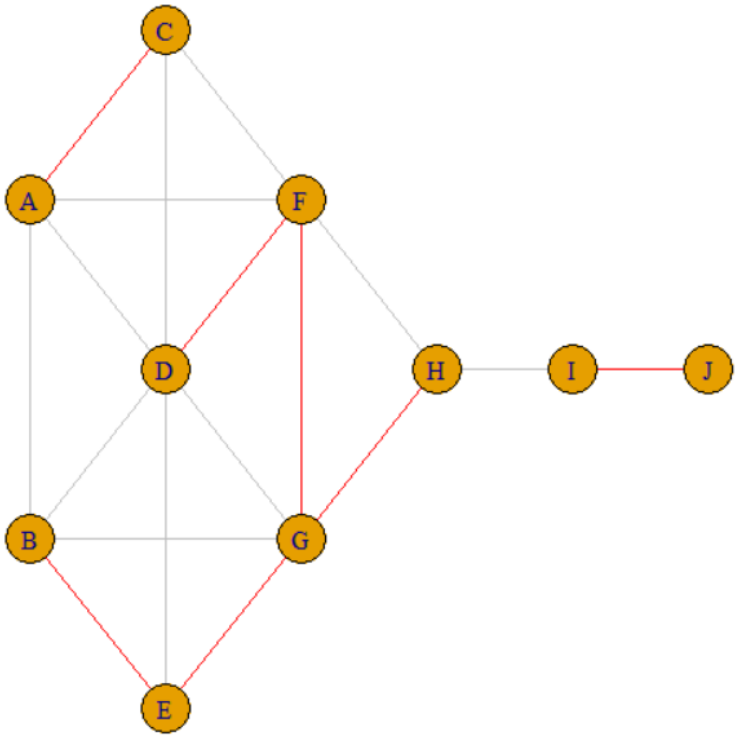
Graph after adding the top 5 predicted edges –



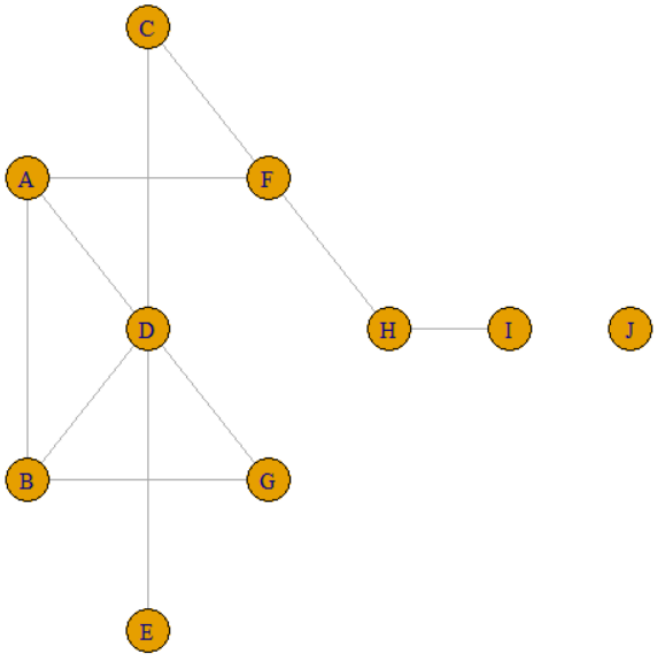
It appears that deleting 15% of the edges in this graph makes it hard to predict the original edges. Furthermore, the structure of the graph has changed substantially as compared to the 5% deletion.

Kite – 40%

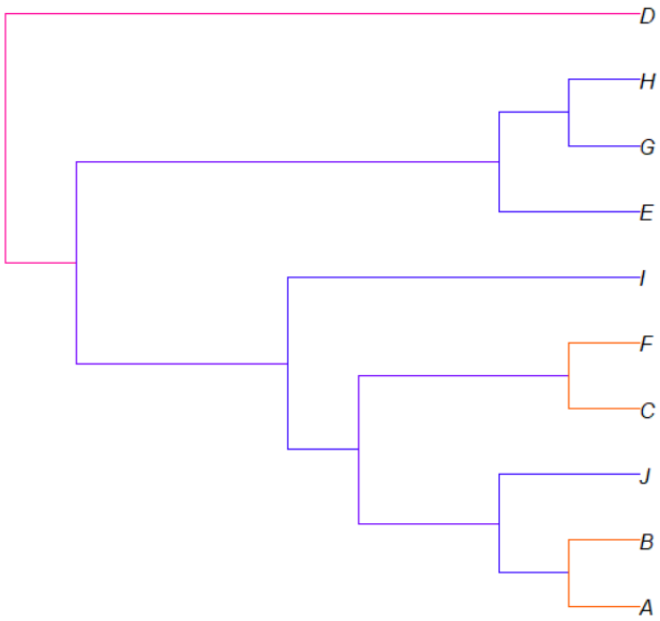
The complete graph where the 7 red edges are the 40% edges to be remove to create the noisy dataset-



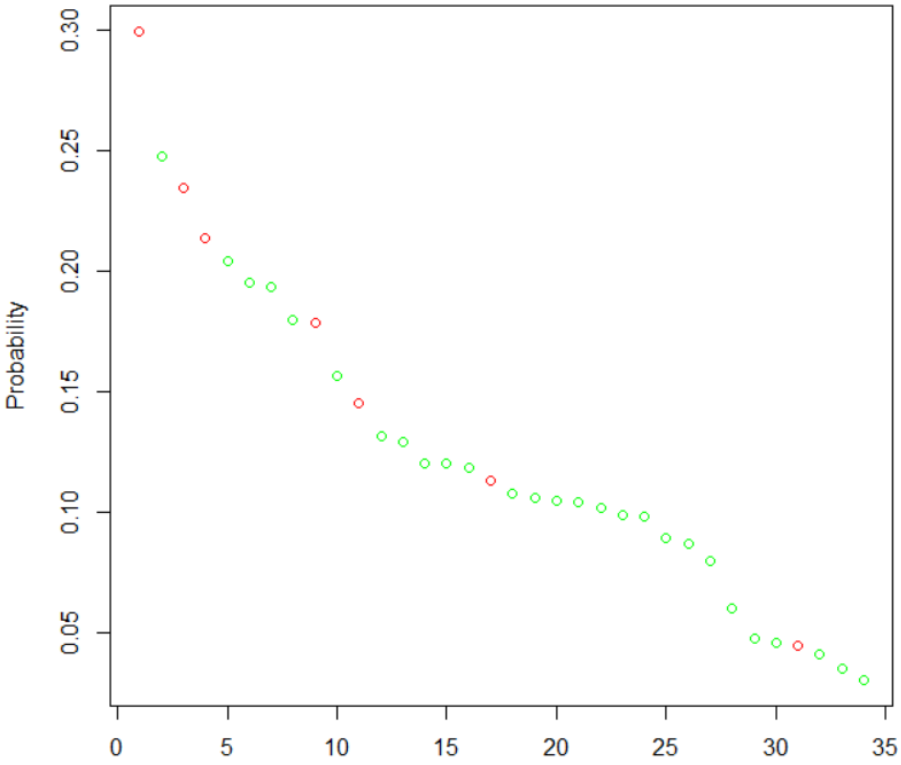
The graph after removing the edges –



Dendrogram after performing MCMC on the noisy dataset -

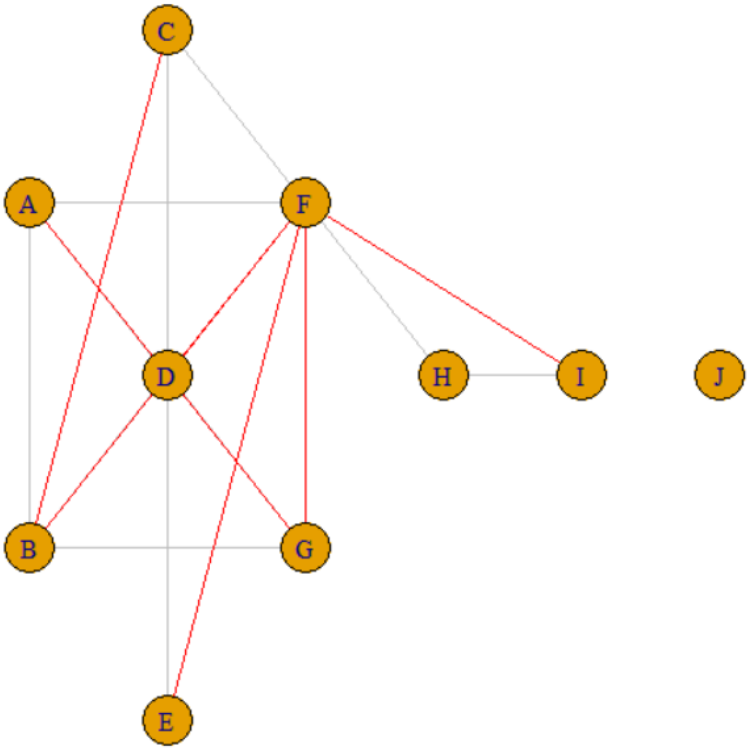


Probability of the predicted edges with the removed edges (A—C,B—E,G—H,I—J,F—G,D—F,E—G) in red -



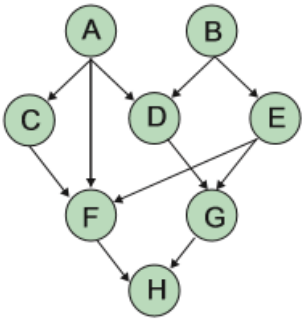
3 of the 7 removed edges are predicted rather well with 5 in the top 11 out of 34 edges.

Graph after adding top 7 of the predicted edges –



Once again, it is clear that deleting 40% of the edges and applying link prediction results in a significantly different graph structure as compared to 15% and 5% deletion.

2)



A) False – C and G are not d-separated as they're connected via a path through A.

B) True – C and E do not have any common ancestors and are marginally independent.

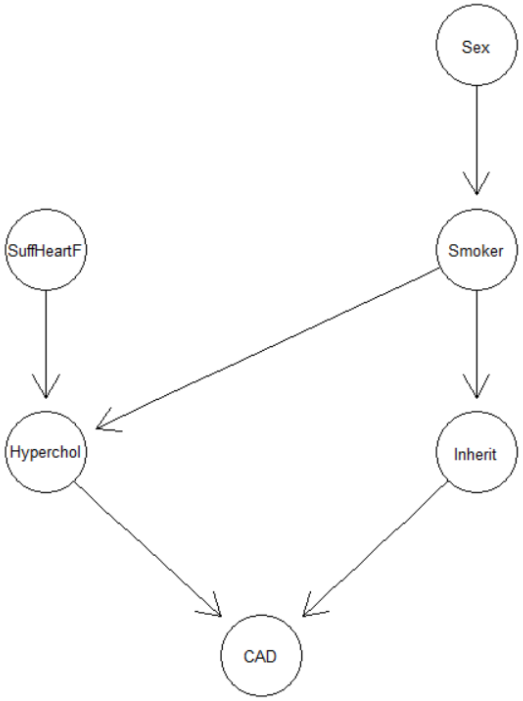
C) True – C and E do not have any common ancestors but they become dependent when given the node G due to a connection via a path through D.

D) False – A and G are not d-connected because both the parents of G are given and there is no connection after deleting them.

E) True – A and G are d-connected via a path through B and E after "moralizing" because only D is given.

3)

The graph of the “optimal network” with 6 variables -



Conditional probabilities obtained by using extractCPT() –

Sex –

```
Sex
   Female      Male
0.1991525 0.8008475
```

SuffHeartF -

```
SuffHeartF
       No       Yes
0.7076271 0.2923729
```

Smoker|Sex –

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

Hyperchol|Smoker,SuffHeartF –

```
, , Smoker = No

         SuffHeartF
Hyperchol        No       Yes
      No  0.6750000 0.2727273
      Yes 0.3250000 0.7272727

, , Smoker = Yes

         SuffHeartF
Hyperchol        No       Yes
      No  0.4645669 0.3275862
      Yes 0.5354331 0.6724138
```

Inherit|Smoker –

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

CAD|Inherit,Hyperchol –

```
, , Inherit = No

     Hyperchol
CAD           No       Yes
  No  0.8214286 0.4487179
  Yes 0.1785714 0.5512821

, , Inherit = Yes

     Hyperchol
CAD           No       Yes
  No  0.5000000 0.2600000
  Yes 0.5000000 0.7400000
```

D – Seperations:

Sex and SuffHeartF

Smoker and SuffHeartF

CAD and Sex given evidence about Smoker

Hyperchol and Sex given evidence about Smoker etc.

b) After compiling and propagating, querying the marginal, joint and conditional probabilities BEFORE and AFTER absorbing the evidence –

Marginal (Before, After)–

```
$SuffHeartF
SuffHeartF
       No       Yes
0.7076271 0.2923729

$CAD
CAD
       No       Yes
0.5401298 0.4598702
```

```
$SuffHeartF
SuffHeartF
       No       Yes
0.6162534 0.3837466

$CAD
CAD
       No       Yes
0.3924294 0.6075706
```

Joint (Before, After)–

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

```
           CAD
SuffHeartF        No       Yes
       No  0.2408676 0.3753858
       Yes 0.1515618 0.2321848
```

Conditional (Before, After)–

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

```
     SuffHeartF
CAD           No       Yes
  No  0.6137859 0.3862141
  Yes 0.6178472 0.3821528
```

It is clear that after absorbing the new evidence (female sex and high cholesterol) the marginal probability of suffering from heart failure increased along with significant increase in CAD.

The joint and conditional probabilities show a similar increase.

c) Simulating a new data set with 25 observations –

In this data set -

Smoker – No : 8, Yes: 17

CAD – No: 11, Yes: 14

The predict() function is predicting that all 25 observations are Smokers.

d) Simulating a new data set with 500 observations –

In this data set –

Smoker – No: 152, Yes: 348

CAD – No: 193, Yes: 307

The predict() function however is once again predicting that all 500 observations are Smokers!

I'm not able to figure out where exactly I went wrong. In the very unlikely case that I didn't make a mistake, I'm uncertain if this is due to a bug or if the seed is resulting in such a prediction. I couldn't find much help online.