STA 546

STATISTICAL DATA MINING 2

HOMEWORK 4

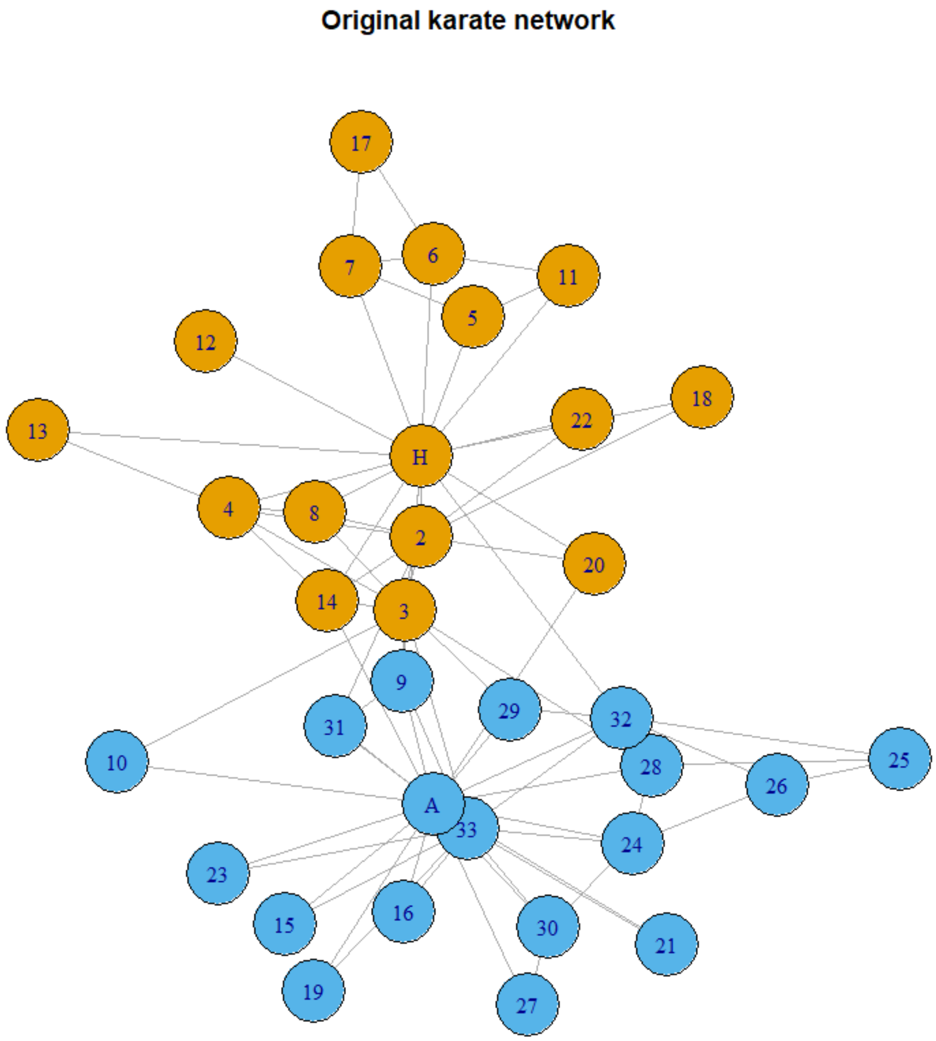
Abinesh Senthil Kumar

#50320934

Question1

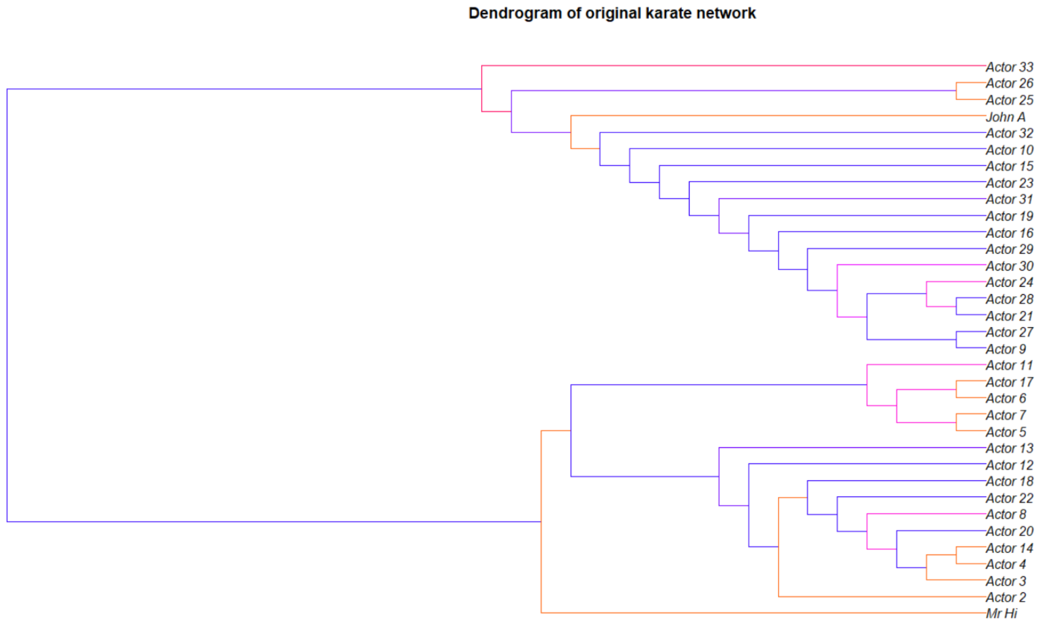
(a) Focus on the karate network. Create noisy datasets. Do this by deleting 5% of the edges randomly (track which ones they are). Perform MCMC for a random graph model (as in Clauset et al.) on this data followed by linkprediction. Are you able to predict the edges that you deleted?

Considering the karate network,


Original karate network

- The above picture represents the original karate network.

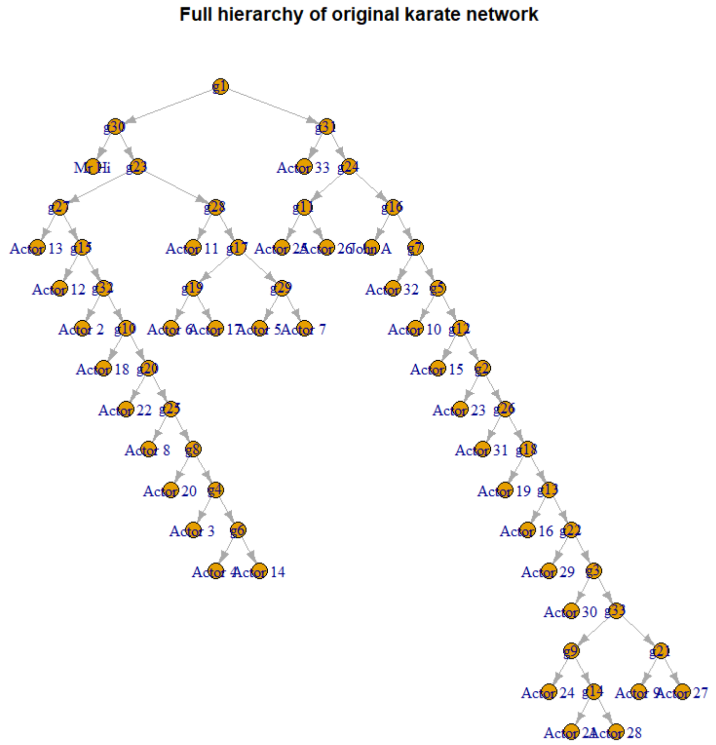
Using fit_hrg() function to fit dendrogram we get,



Dendrogram of original karate network

- We can clearly see there are two groups in the network as expected.

Plotting the full hierarchy of the original Karate network,



Full hierarchy of original karate network

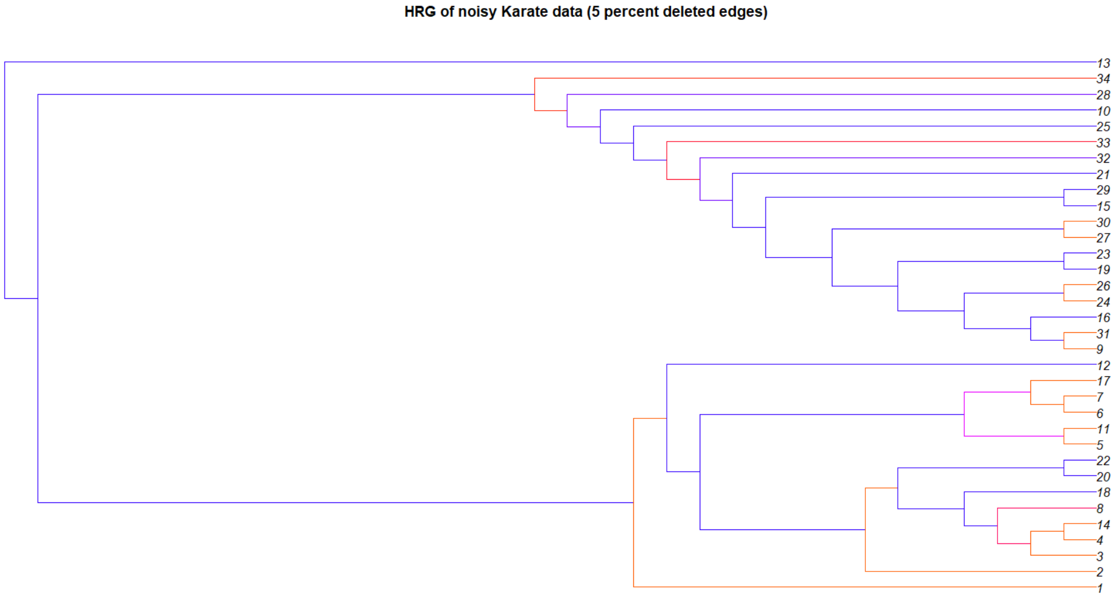
- Again, we can clearly see two groups in the karate network.

Deleting specific percent of edges randomly,

- A function is created to delete the specific percent of edges and keep track of them.
- Also, the function gives out the deleted edges and the respective probabilities from the total predicted edges.

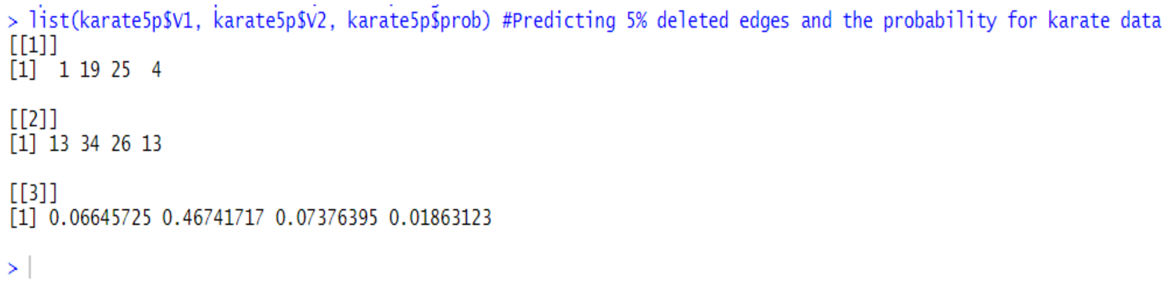
Karate network – 5% deleted data

Plotting the dendrogram of 5% deleted edges from the original network,



- We can still see that there are two groups present.

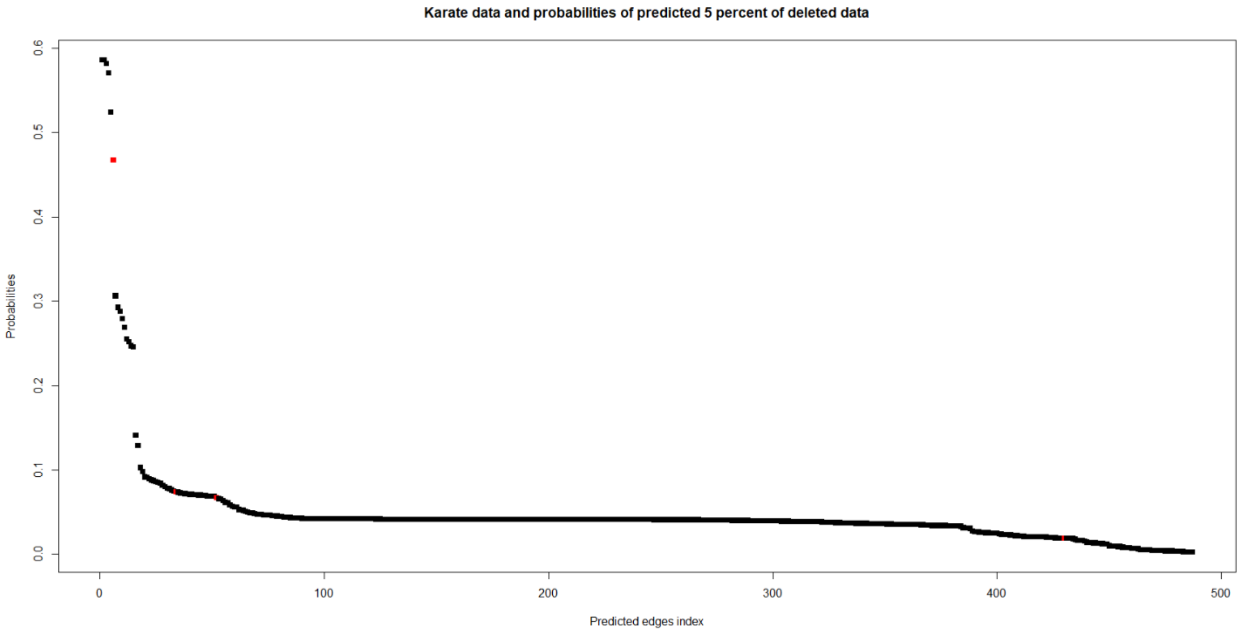
The deleted edges and the probability are given as,

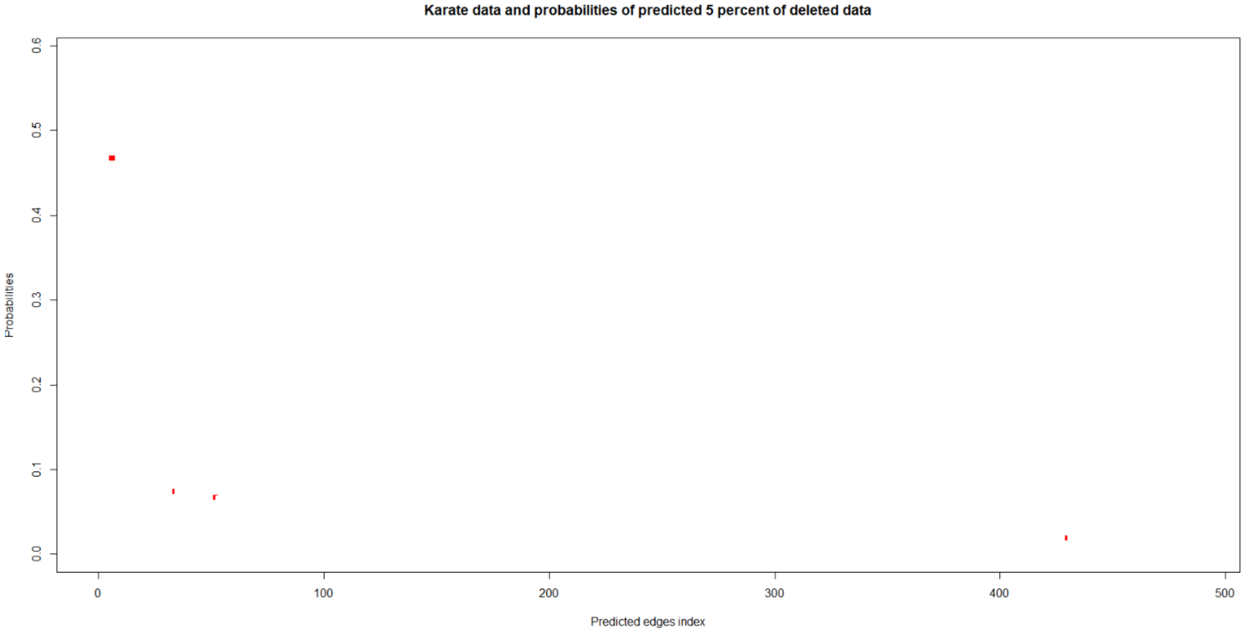


```
> list(karate5p$V1, karate5p$V2, karate5p$prob) #Predicting 5% deleted edges and the probability for karate data
[[1]]
[1]  1 19 25  4

[[2]]
[1] 13 34 26 13

[[3]]
[1] 0.06645725 0.46741717 0.07376395 0.01863123

> |
```

- The three lists represent the specific edge and their respective probability. i.e. nodes 1 and 13 has a predicted probability of 0.06645725

Plotting the deleted edges probability in the full predicted edges,



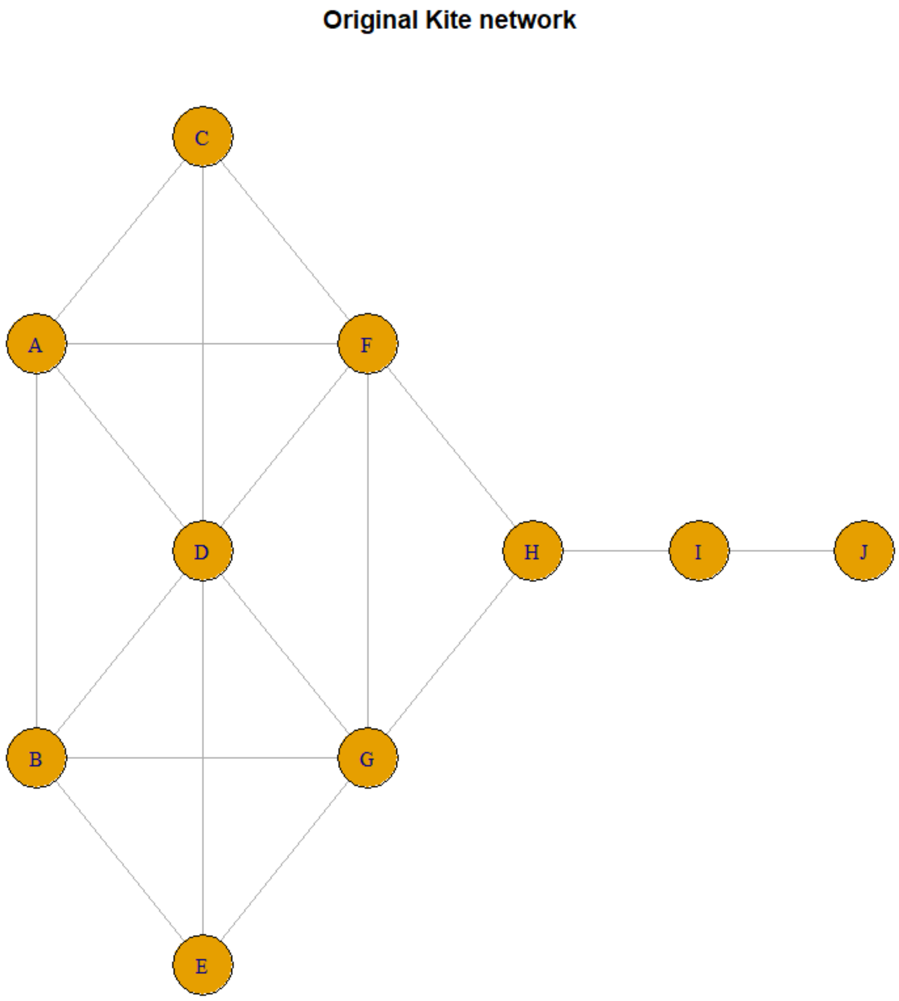


- The red dots represent our randomly deleted edges from the graph.
- As seen from the values we can see that only one edge has a high probability of 0.46 out of four predicted edges.
- The predicted (deleted) edges are spread throughout the range of predicted edges.

Yes, we are able to predict the randomly deleted edges. But most of them as less probability

(b) Focus on the yeast network (or kite network). Create noisy datasets. Do this by deleting 5% of the edges randomly (track which ones they are). Perform MCMC on this data followed by link-prediction. Are you able to predict the edges that you deleted at random well?

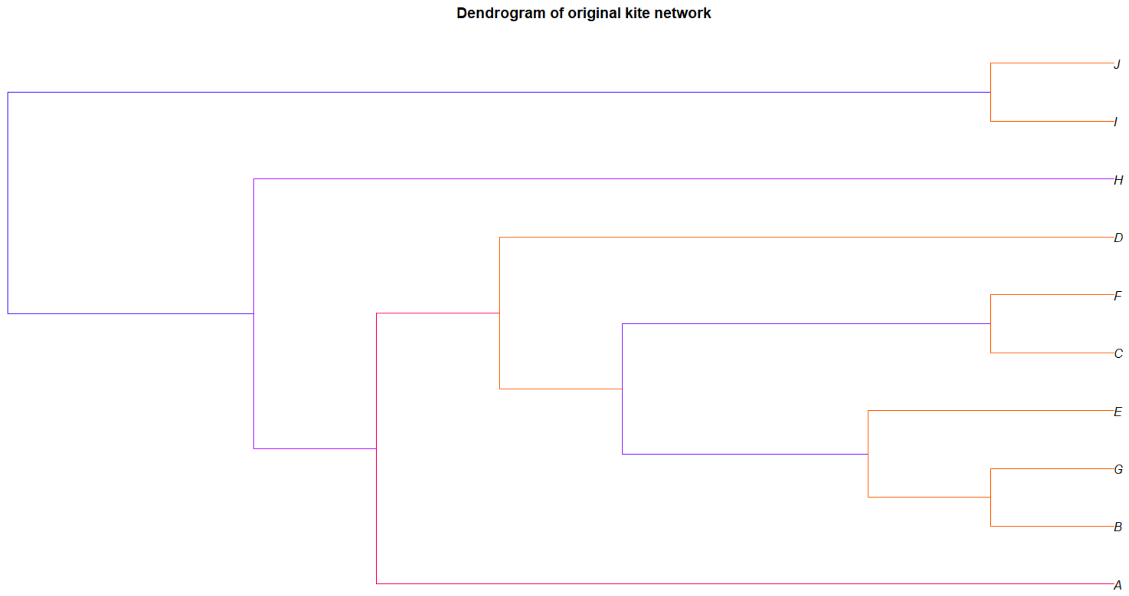
Considering the **kite** network



Original Kite network

- The above picture represents the original kite network.

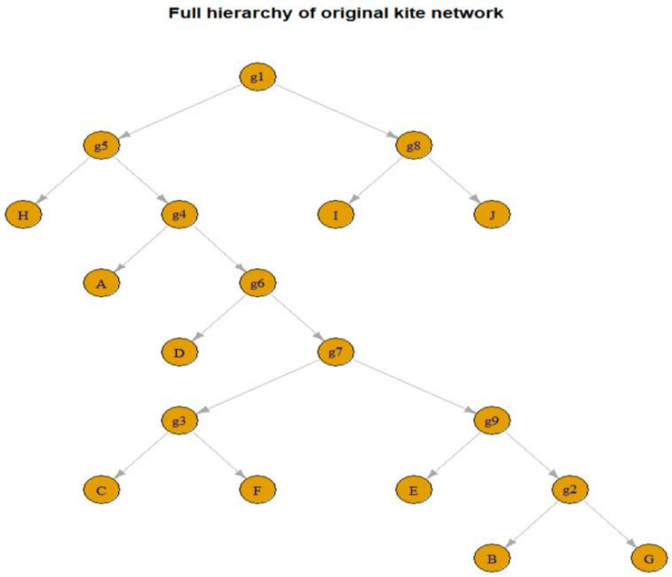
Using fit_hrg() function to fit dendrogram we get,



Dendrogram of original kite network

- From this we can see the network kind of has two subgroups but one of them has less nodes.

Plotting the full hierarchy,



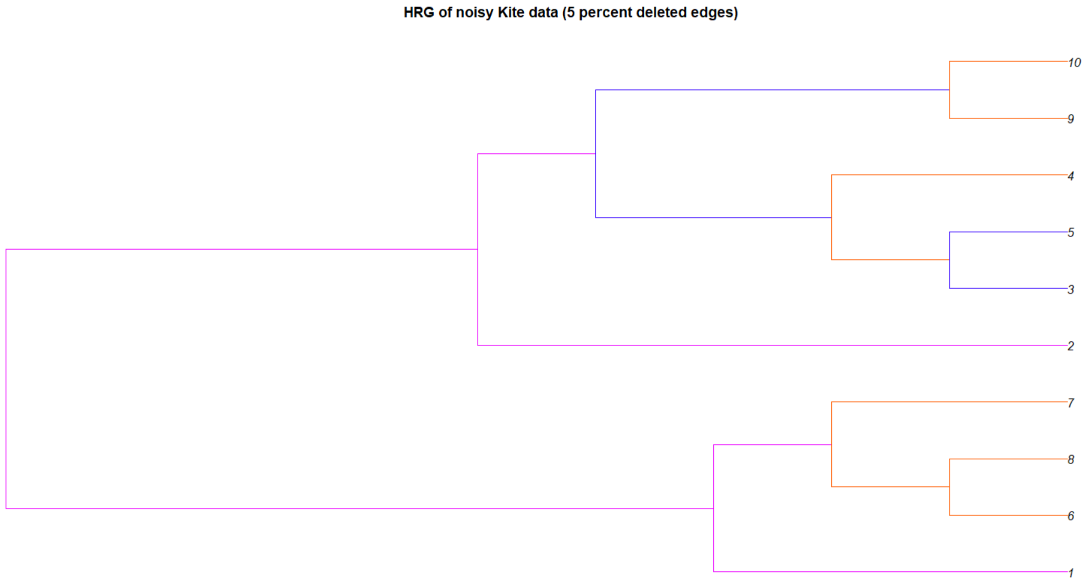
Full hierarchy of original kite network

Deleting specific percent of edges randomly,

- A function is created to delete the specific percent of edges and keep track of them.
- Also, the function gives out the deleted edges and the respective probabilities from the total predicted edges.

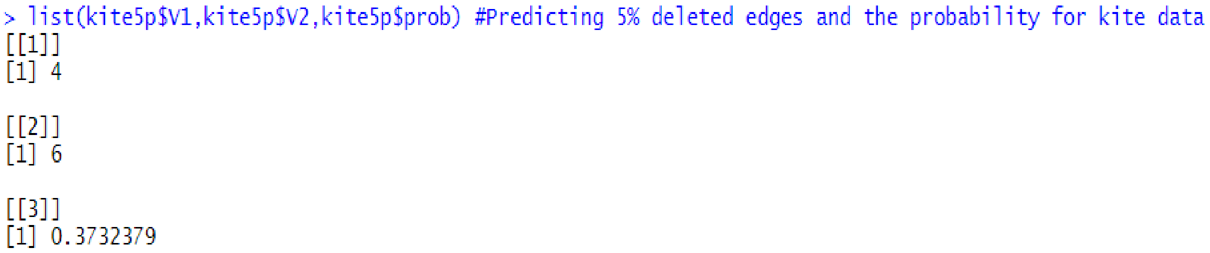
Kite network – 5% deleted data

Plotting the dendrogram of the noisy network,



- After deleting 5% of the edges, we can see that the network groups into two sub groups

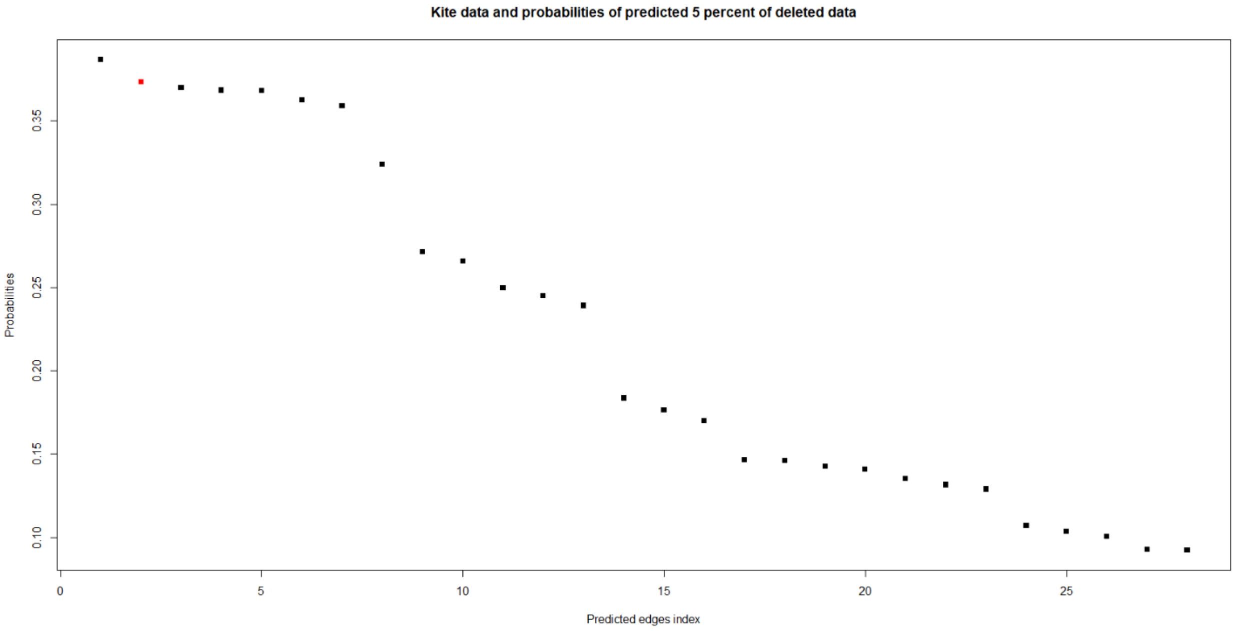
The deleted edges and the probability is given by,

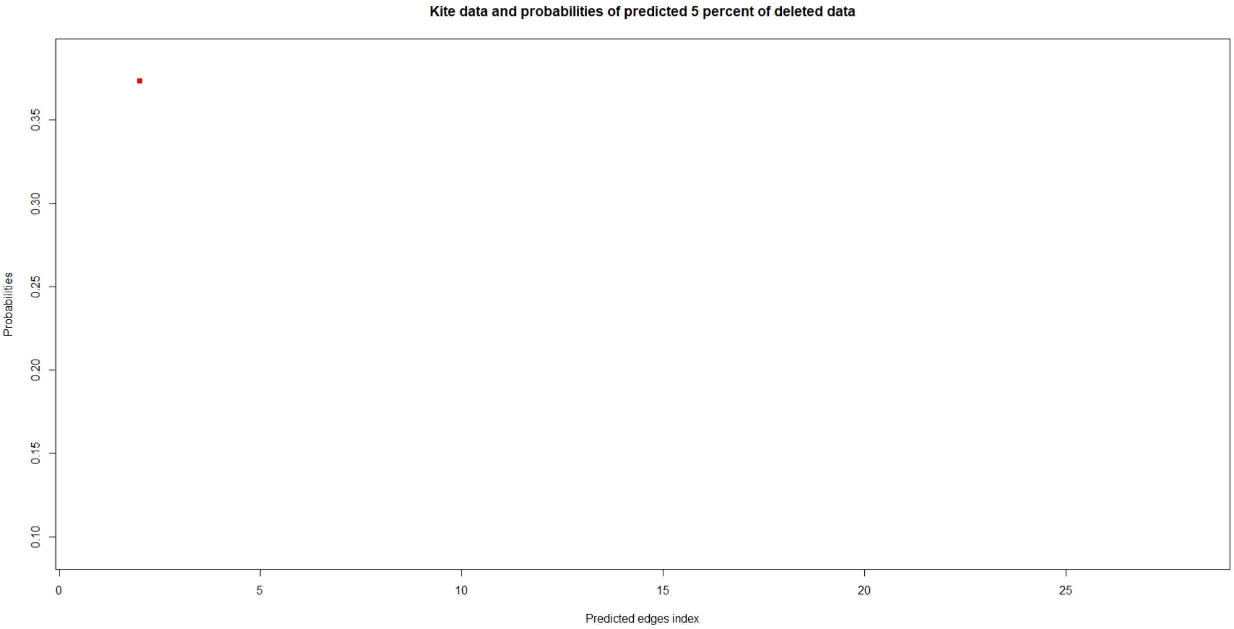


```
> list(kite5p$V1,kite5p$V2,kite5p$prob) #Predicting 5% deleted edges and the probability for kite data
[[1]]
[1] 4

[[2]]
[1] 6

[[3]]
[1] 0.3732379
```

- From this we can see that 5% includes only one edge and the deleted edge is node 4 to node 6 and it is predicted with a probability of 0.3732379

Plotting the deleted edges probability with the full predicted edges,





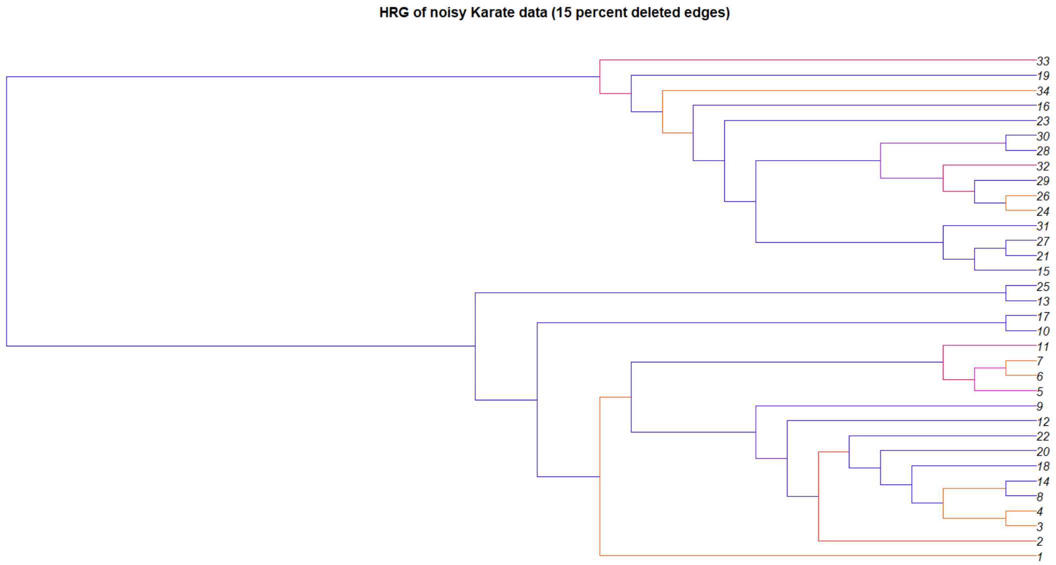
- The red dot are the deleted edges and from this we can see that the predicted probability of our deleted edge is quite high compared to other predicted edges.

Yes, we are able to predict the deleted edges and it has a high probability

(c) Repeat the exercise in part (a) and (b) after deleting 15%, and 40% of the edges. Comment on your findings.

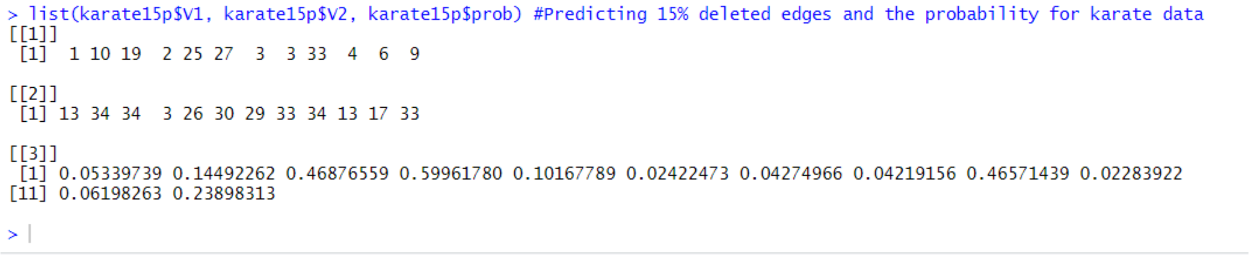
Karate network – 15% deleted data

Plotting the dendrogram of noisy network after deleting 15% of data,



- We can see that the network groups into two even after deleting 15% of edges.

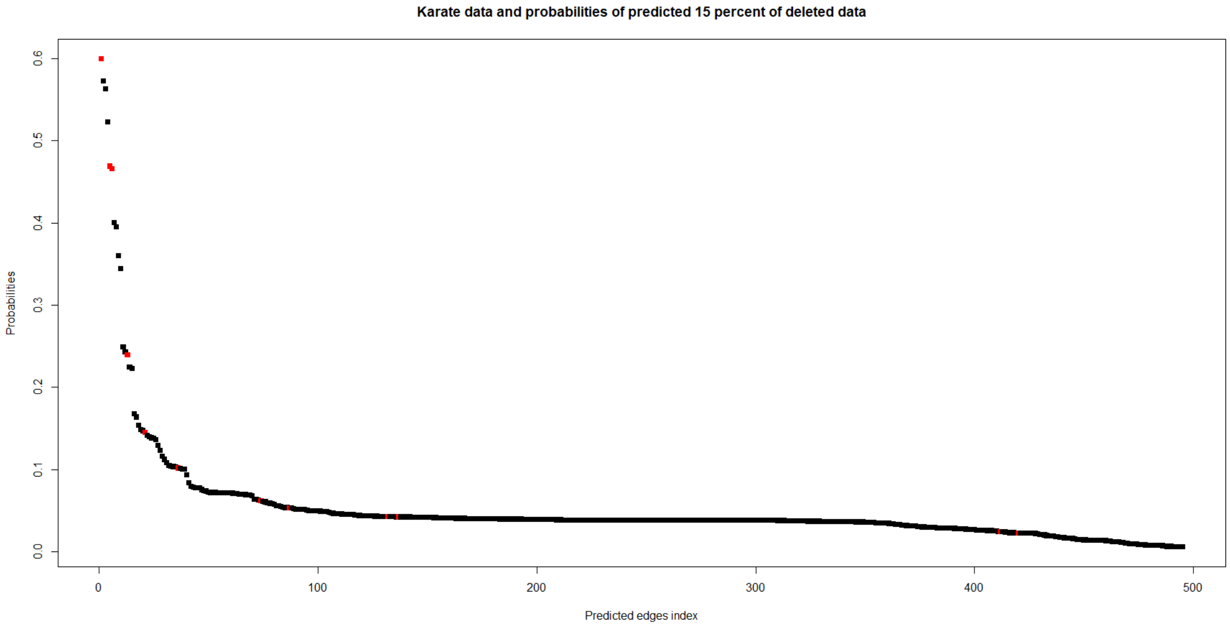
The deleted edges with probability is given by,

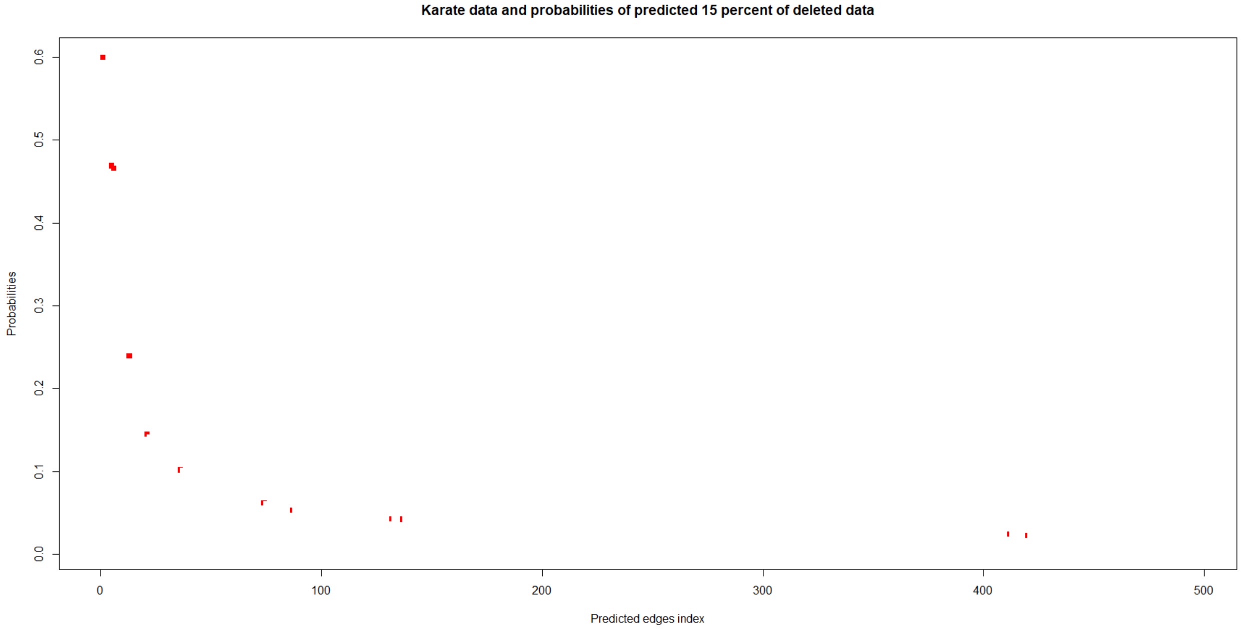


```
> list(karate15p$V1, karate15p$V2, karate15p$prob) #Predicting 15% deleted edges and the probability for karate data
[[1]]
 [1]  1 10 19  2 25 27  3  3 33  4  6  9

[[2]]
 [1] 13 34 34  3 26 30 29 33 34 13 17 33

[[3]]
 [1] 0.05339739 0.14492262 0.46876559 0.59961780 0.10167789 0.02422473 0.04274966 0.04219156 0.46571439 0.02283922
[11] 0.06198263 0.23898313

> |
```

- From this we can see that three of the edges (19-34, 2-3, 33-34) are predicted with high probabilities compared to other predicted edges.

Plotting the deleted edges probability with full predicted probability,

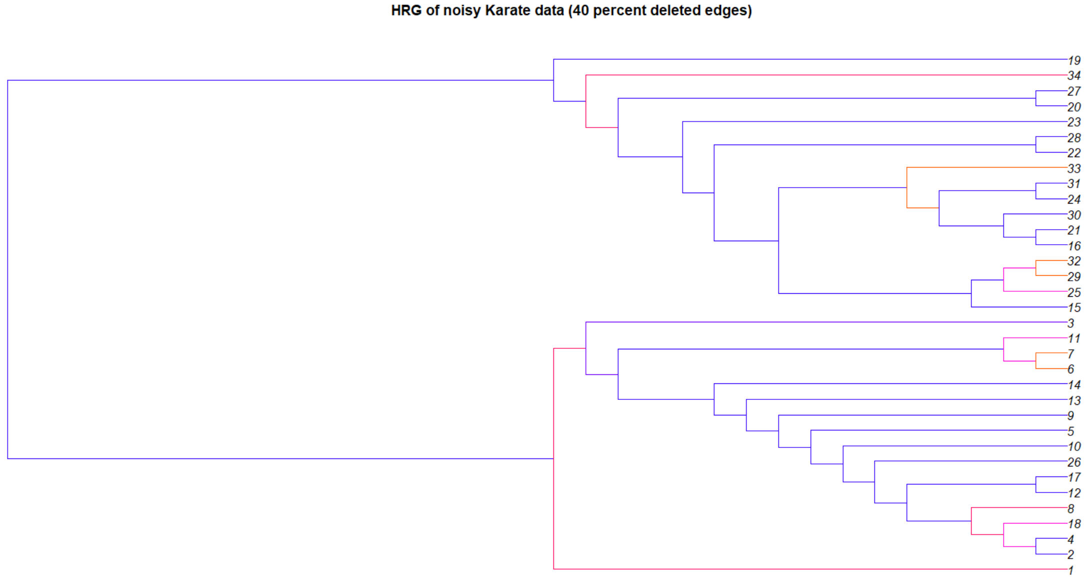




- The red dots are the randomly deleted edges and as said before only three of the edges have very high probability out of all predicted edges.
- The predicted (deleted) edges are spread throughout the range of predicted edges.

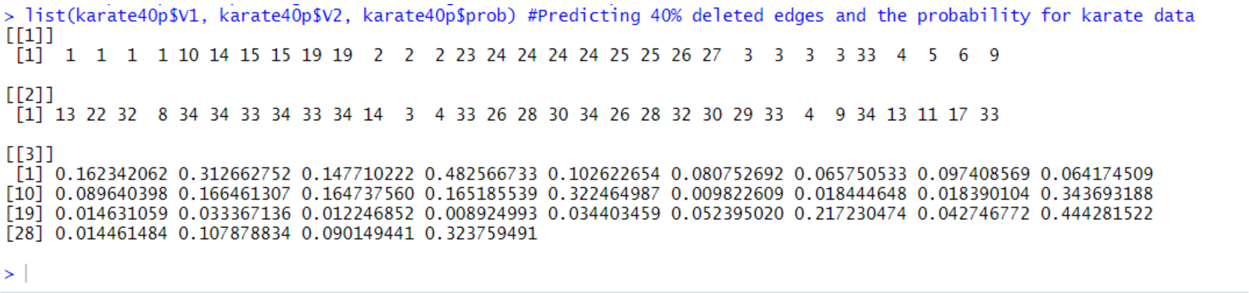
Karate data – 40% deleted edges.

Plotting the dendrogram of noisy network after deleting 40% of edges,



- We can see that the network still groups into two groups even after deleting 40% of edges.

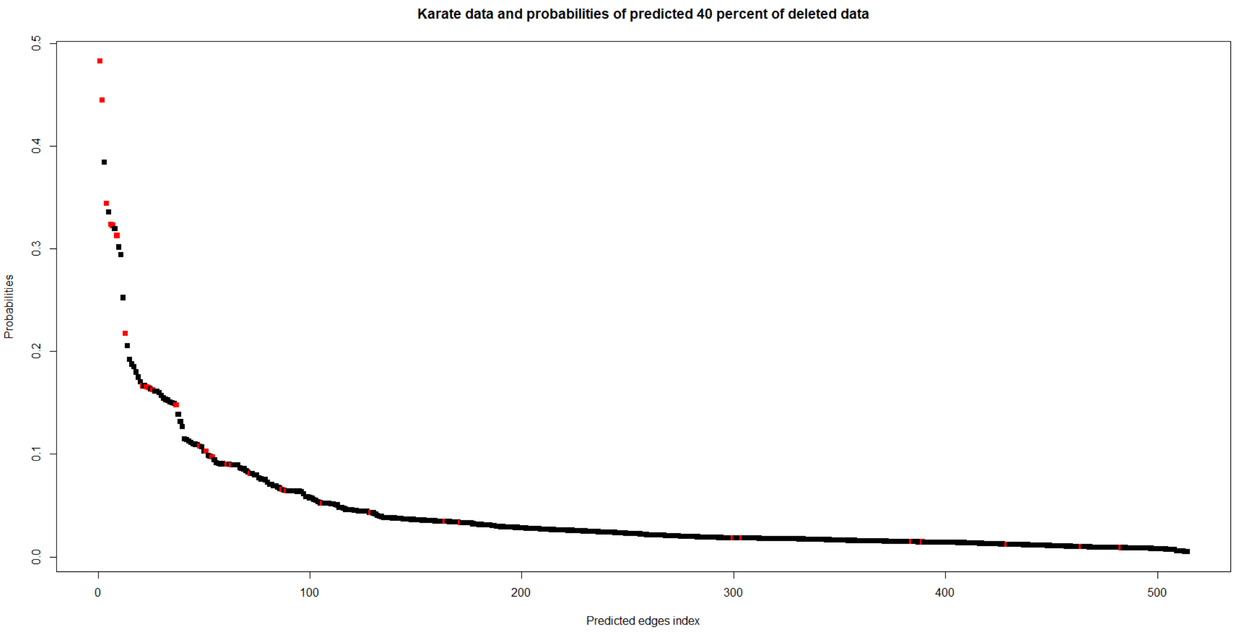
The deleted edges with probability are given by,

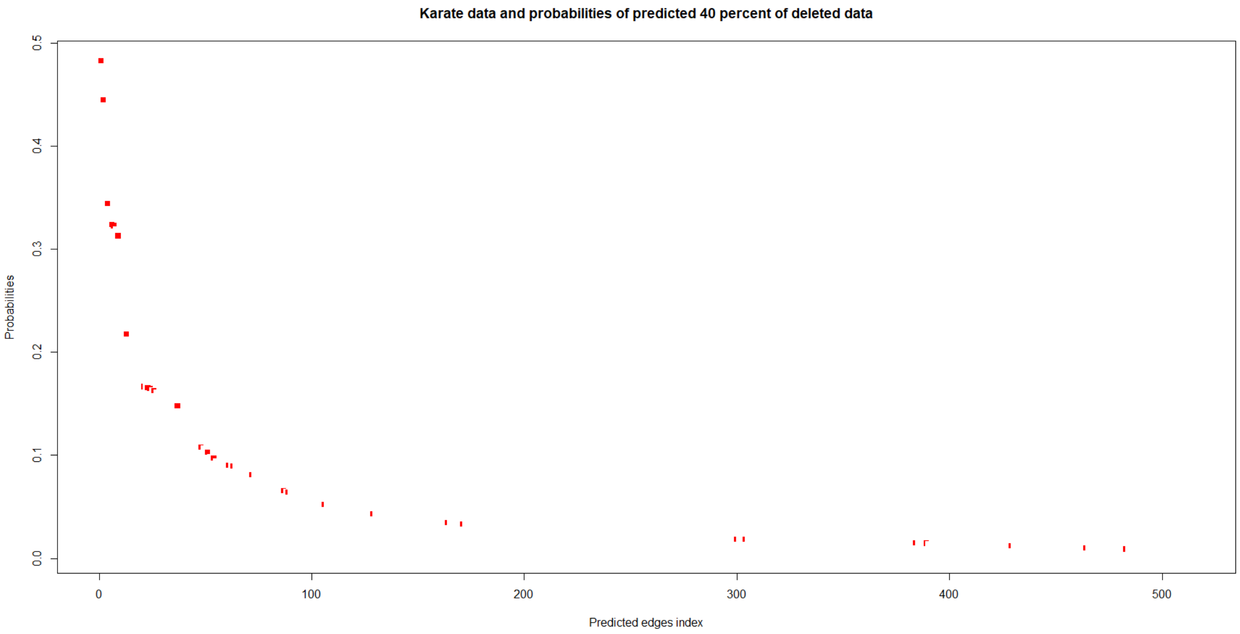


```
> list(karate40p$V1, karate40p$V2, karate40p$prob) #Predicting 40% deleted edges and the probability for karate data
[[1]]
 [1]  1  1  1  1 10 14 15 15 19 19  2  2  2 23 24 24 24 24 25 25 26 27  3  3  3  3 33  4  5  6  9

[[2]]
 [1] 13 22 32  8 34 34 33 34 33 34 14  3  4 33 26 28 30 34 26 28 32 30 29 33  4  9 34 13 11 17 33

[[3]]
 [1] 0.162342062 0.312662752 0.147710222 0.482566733 0.102622654 0.080752692 0.065750533 0.097408569 0.064174509
[10] 0.089640398 0.166461307 0.164737560 0.165185539 0.322464987 0.009822609 0.018444648 0.018390104 0.343693188
[19] 0.014631059 0.033367136 0.012246852 0.008924993 0.034403459 0.052395020 0.217230474 0.042746772 0.444281522
[28] 0.014461484 0.107878834 0.090149441 0.323759491

> |
```

- From this we can see the deleted edges and their predicted probability.

Plotting the deleted edges with full predicted probability,

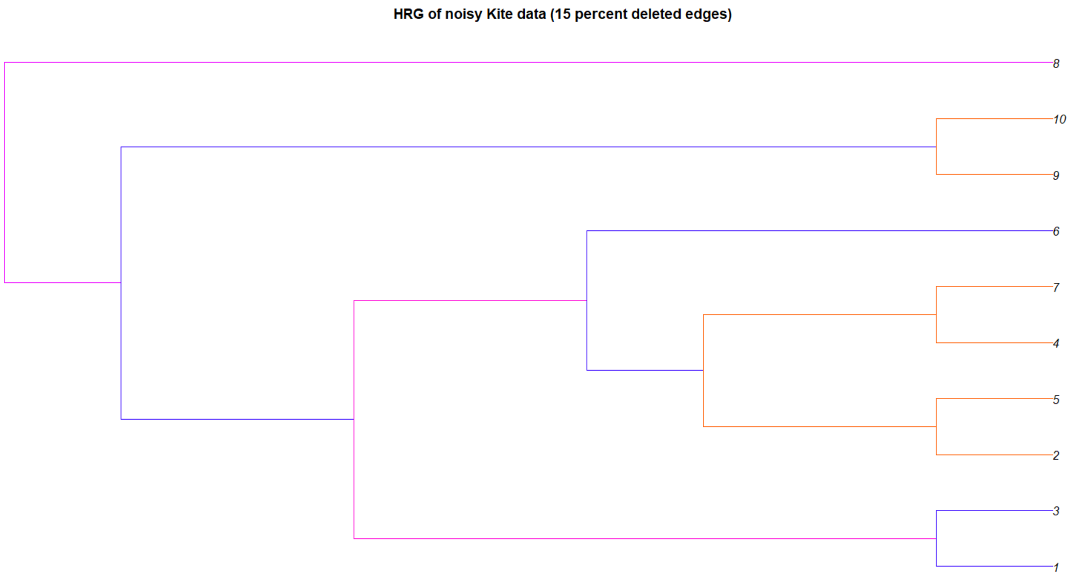




- The red dots represent the predicted (deleted) edges from the original network and we can see that few edges are infact in top compared to other edges.
- The predicted (deleted) edges are still spread throughout the range of predicted edges.

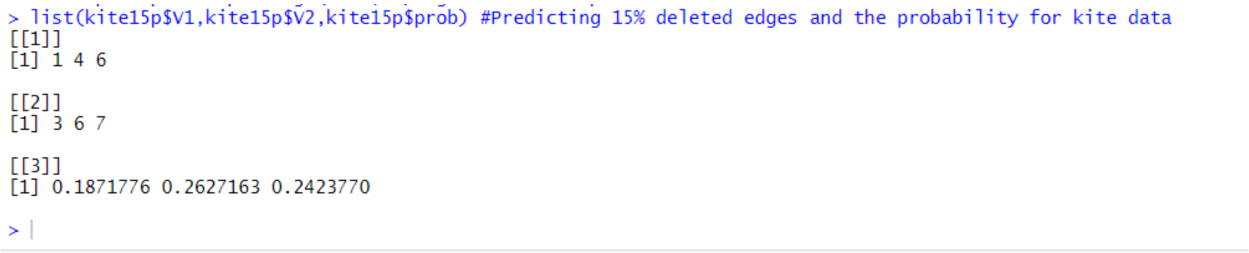
Kite network – 15% deleted data,

Plotting the dendrogram of noisy network,



- From this dendrogram we can see that the network kind of has two subgroups but one of the groups has only one point.

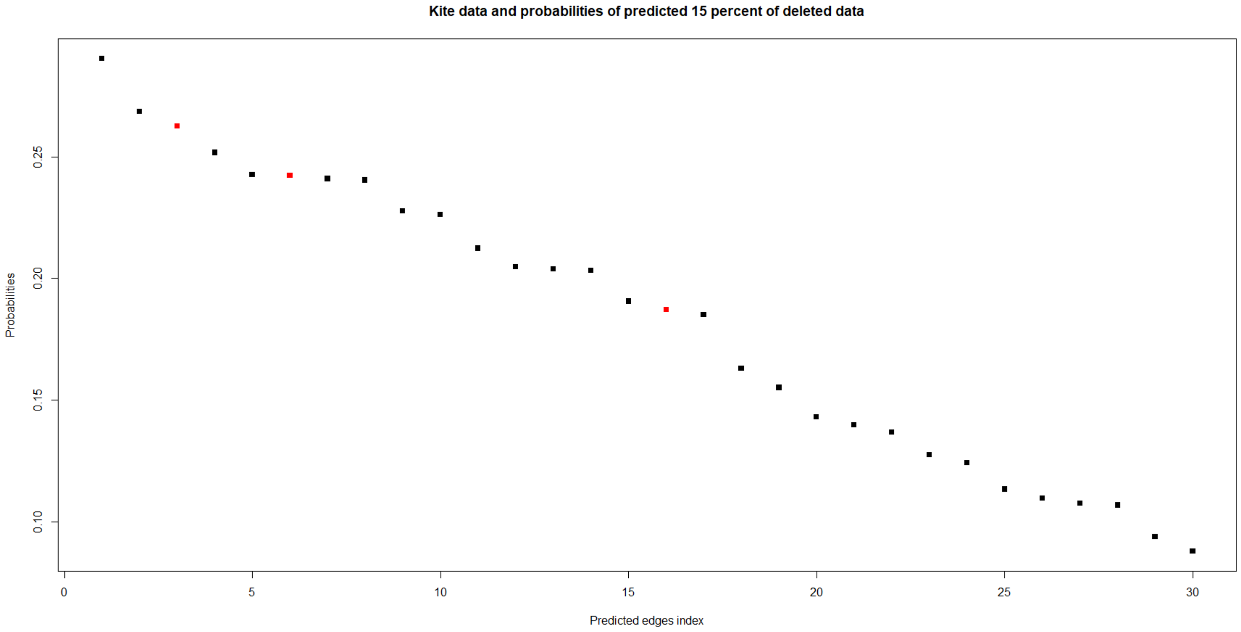
The deleted edges with probability is given by,

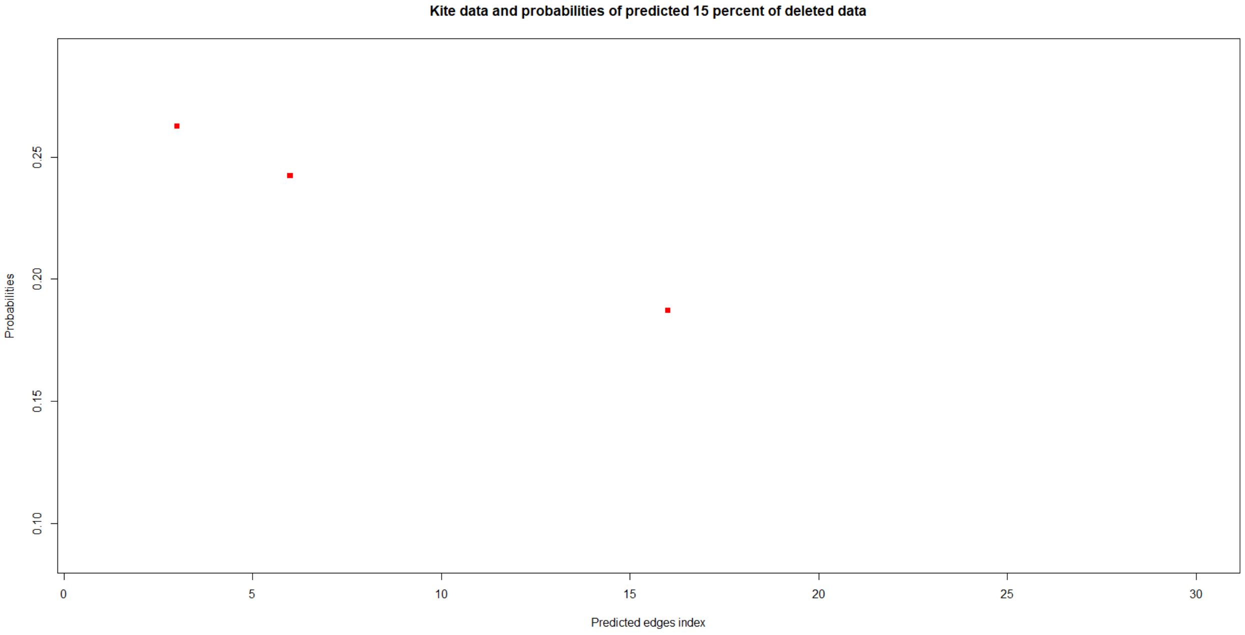


```
> list(kite15p$V1,kite15p$V2,kite15p$prob) #Predicting 15% deleted edges and the probability for kite data
[[1]]
[1] 1 4 6

[[2]]
[1] 3 6 7

[[3]]
[1] 0.1871776 0.2627163 0.2423770

> |
```

- From this we can see that 3 edges account for 15% of deleted data and the edges (4-6, 6-7) has high probability when compared to the other one.

Plotting the deleted edges with full predicted probability,

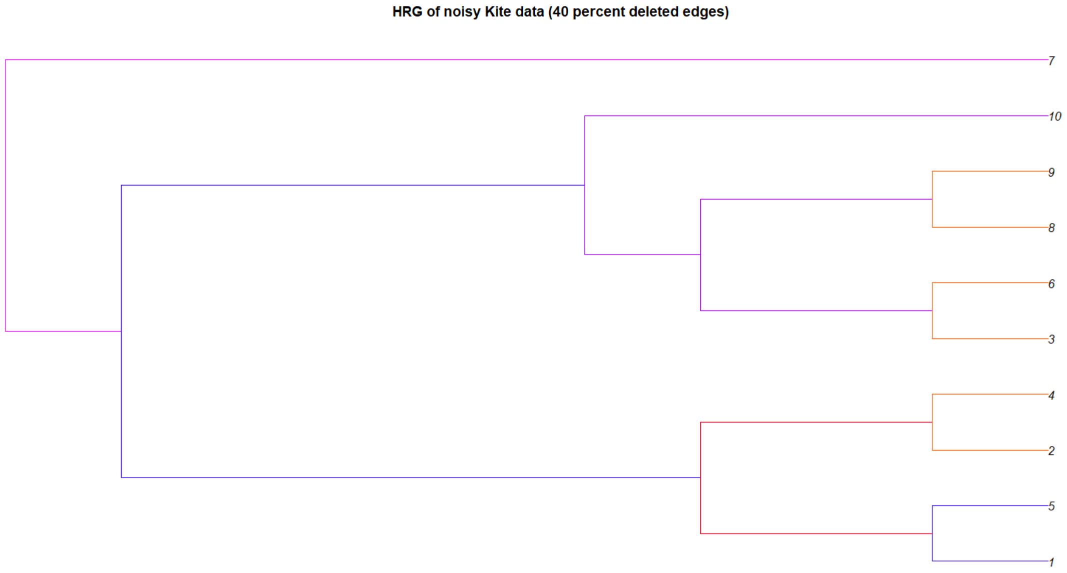




- The red dots represent the predicted(deleted) edges and as said before two of the three edges has quite high probability when compared to the other edge.

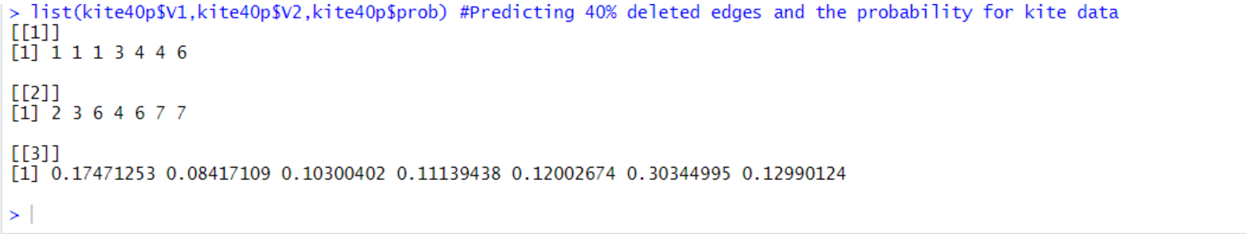
Kite network – 40% deleted edges.

Plotting the dendrogram of the noisy network,



- After deleting 40% of edges we could see that the network groups into two subgroups.

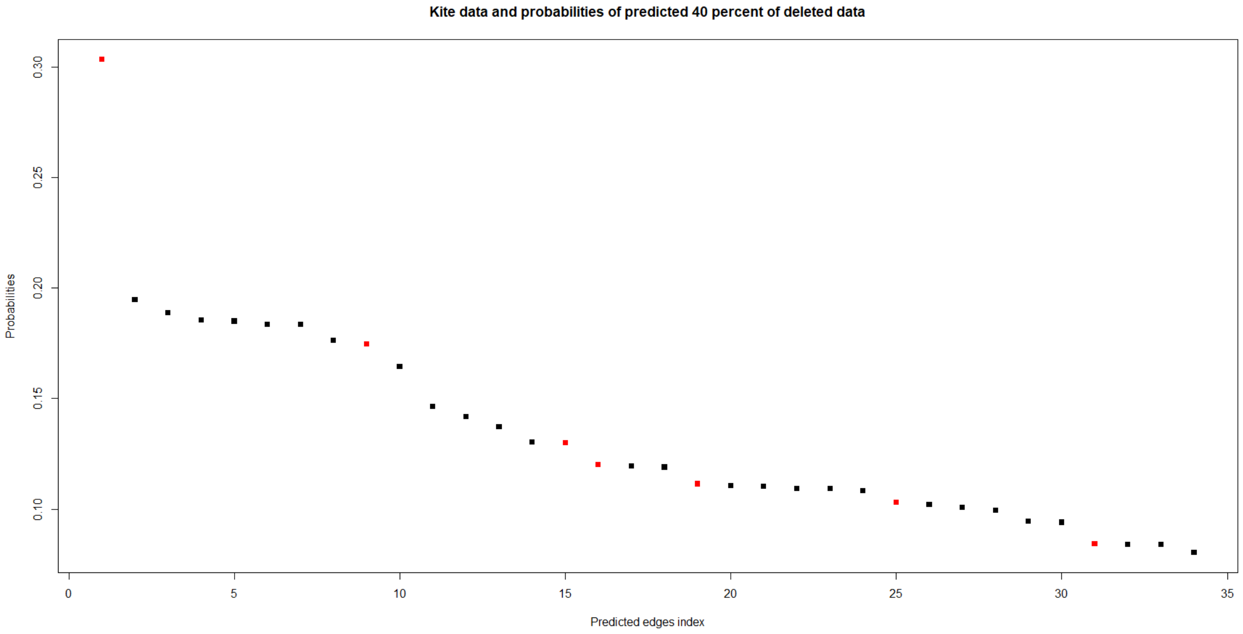
Deleted edges with probability is given by,

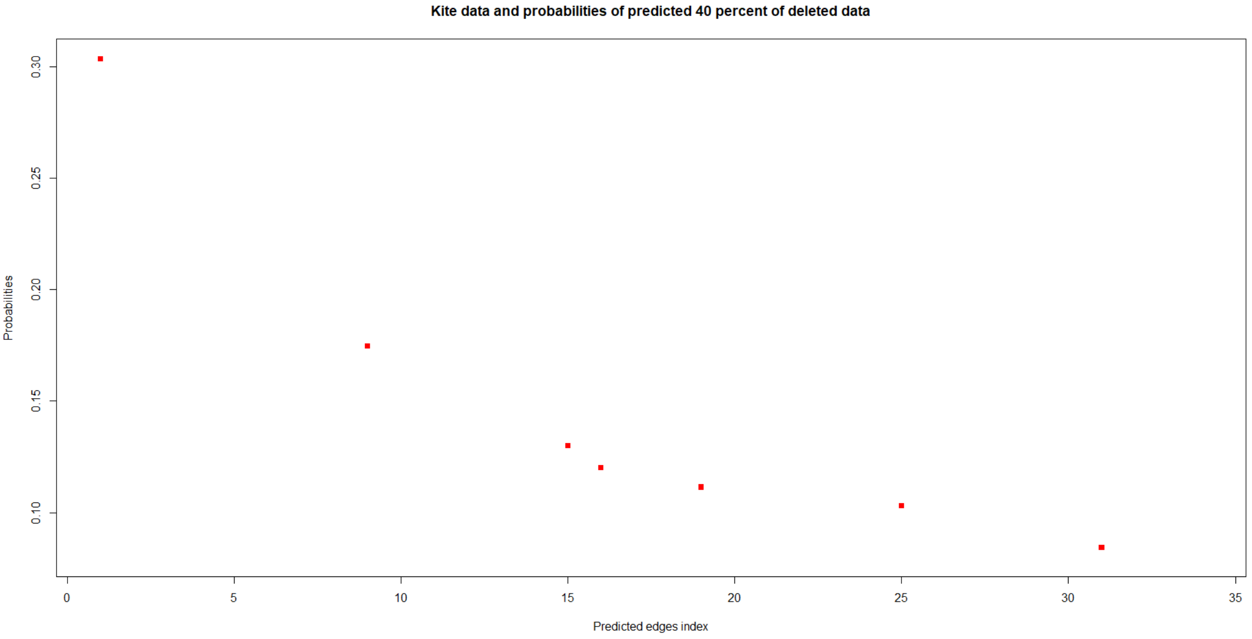


```
> list(kite40p$V1,kite40p$V2,kite40p$prob) #Predicting 40% deleted edges and the probability for kite data
[[1]]
[1] 1 1 1 3 4 4 6

[[2]]
[1] 2 3 6 4 6 7 7

[[3]]
[1] 0.17471253 0.08417109 0.10300402 0.11139438 0.12002674 0.30344995 0.12990124

> 
```

- We can see that only one of the predicted(deleted) edge (4-7) has high probability when compared to others.

Plotting the deleted edges with full predicted probability,

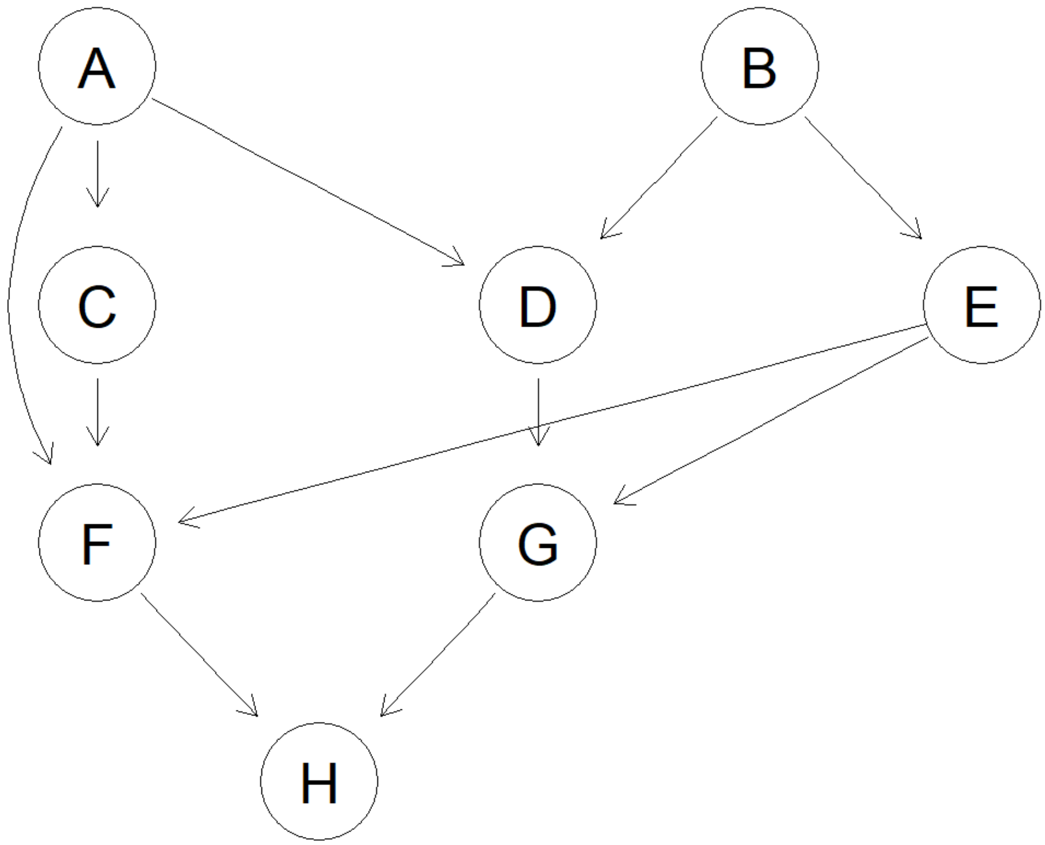




- Red color represents the predicted (deleted) edges
- As said before we could see that only one of the predicted (deleted) edge has high probability when compared to other predicted(deleted) edges.

Question 2

Constructing the given graph on R using dagList function we get,



Using dsep function we can specify the condition to get if the given path is dseperated or not.

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**

Since 3rd question is optional I just attempted it in R and I have attached the R code but, I did not write any report for that.