Link Prediction

Predicting the missing links using hierarchical random graph

Hierarchical random graph is a method which is used for analyzing the well-formed network. In this method we convert a network into a hierarchical dendrogram and with sampling of dendrograms using MCMC – Markov chain and Monte Carlo method by changing edges and finding the best possible fit dendrogram for the network with this the resulted dendrogram helps us to understand the behavior of the network and also gives valuable insights of possible linkages in the network with the probabilities.

Packages Used: igraph, igraphdata
Dataset: karate and kite
Methods: predict_edges,add_edges,delete_edges

Goal: Delete the edges from the network and use the MCMC of the predict_edges method to check for the deleted edges in the predicted list of edges and verify its results

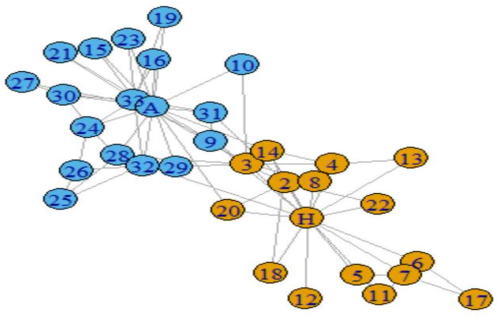
Newly created functions:
fnseqedgesdel – which returns the random sequence of edge indexes that to be deleted on the input of graph and percent of edges to be deleted

fnlistofedgesdel – which returns the edges with vertices which will be deleted based on the input of the graph and sequence of edge ids

ffinddeledgesinpred – which returns the deleted edges found in predicted edges with their probabilities which will be helpful to understand the accuracy of the MCMC method to predict the new edges

faddedgestograph – which adds predicted edges to the graph if the edge is one of the deleted edges then it adds as green to show the edge was recovered from predicted edges if its new edge then it will be added as red

KarateDataset: Social network which has two fighting clubs A and H and how its members are connected.



Original dataset vertices and edge count:

```
> vcount(karate)
```

```
[1] 34
> ecount(karate)
[1] 78
```

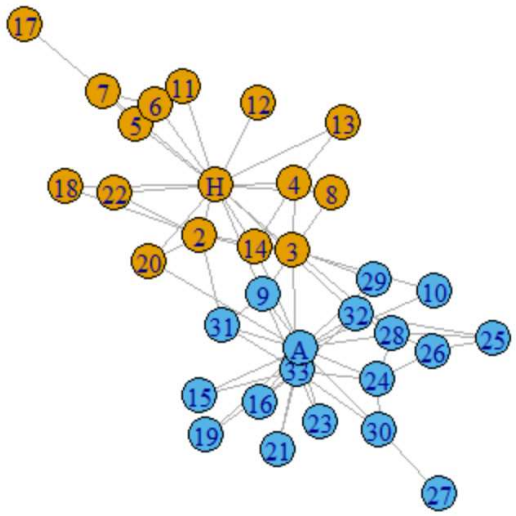
Deleting 5% of edges to create a noisy network:

```
> seqedges<-fnseqedgesdel(karate,5)
> listedgesdel<-fnlistofedgesdel(karate,seqedges)
> karate5<-delete.edges(karate,seqedges)
> vcount(karate5)
[1] 34
> ecount(karate5)
[1] 74
```

Edges deleted:

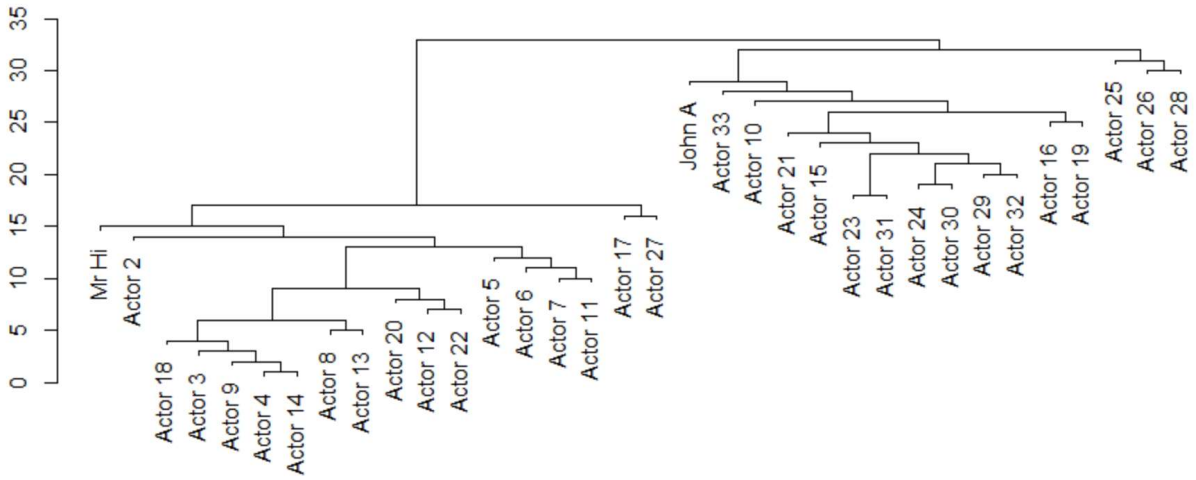
```
> listedgesdel
+ 4/78 edges from 4b458a1 (vertex names):
[1] Actor 6 --Actor 17 Actor 2 --Actor 8  Actor 27--John A   Actor 2 --Actor
4
```

After deletion:



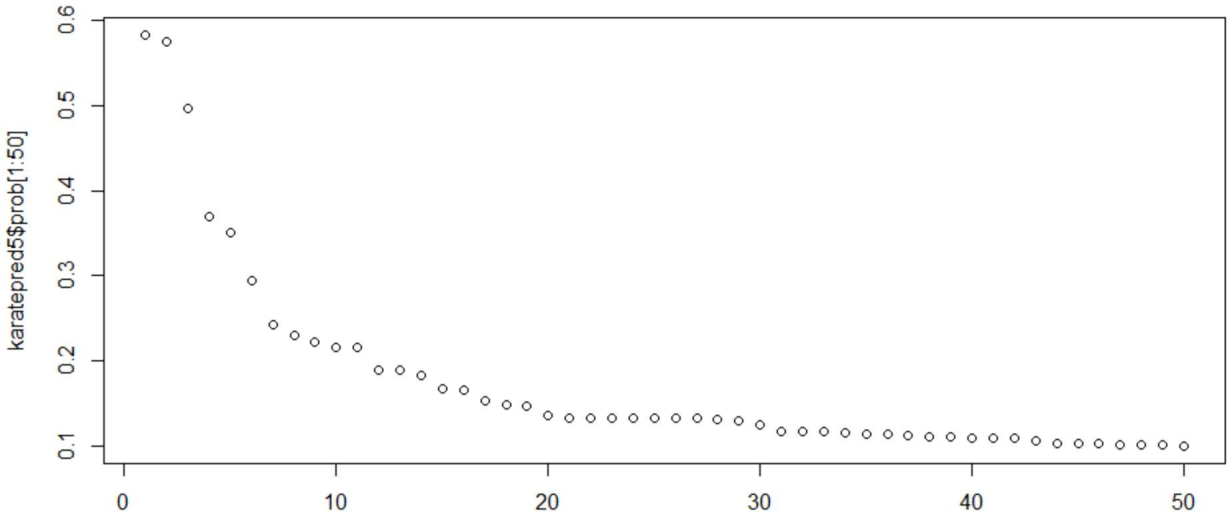
Fitting a network into a dendrogram:

```
> hcg<-fit_hrg(karate5)
>
> #it stricly separated two club members - A and H
> plot_dendrogram(hcg)
```



Predicting Edges:

It has predicted 487 edges and each edge with certain probability. Let's plot the top 50 edges.



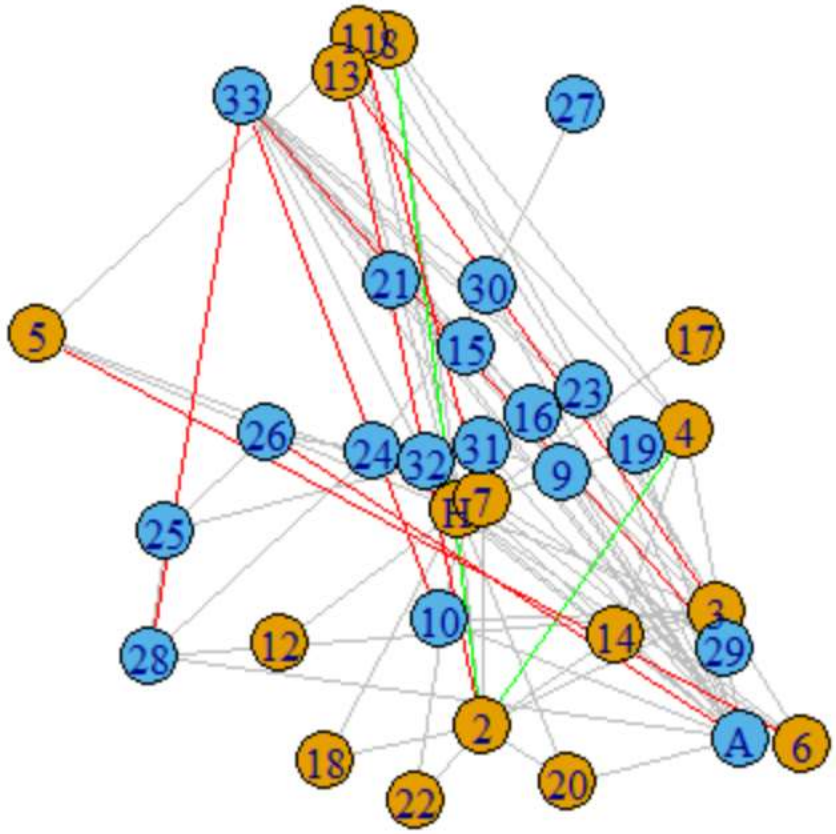
We see the elbow for top 10 edges lets add top 10 edges to the graph.

We had deleted four edges and their probabilities in the predicted edges:

```
> ffinddeledgesinpred(karate5,listedgesdel,karatepred5)
Edge =  Actor 6|Actor 17  Probability= 0.07496459
Edge =  Actor 2|Actor 8  Probability= 0.3509742
Edge =  Actor 27|John A  Probability= 0.1252366
Edge =  Actor 2|Actor 4  Probability= 0.3697002
```

Except 1 edge, rest are highly possible edges with high probability, its part of top 10 edges. Let's add and check the graph.

We can observe 2 deleted edges are recovered from prediction with high probability and are part of the graph after adding **(edges in green)**.

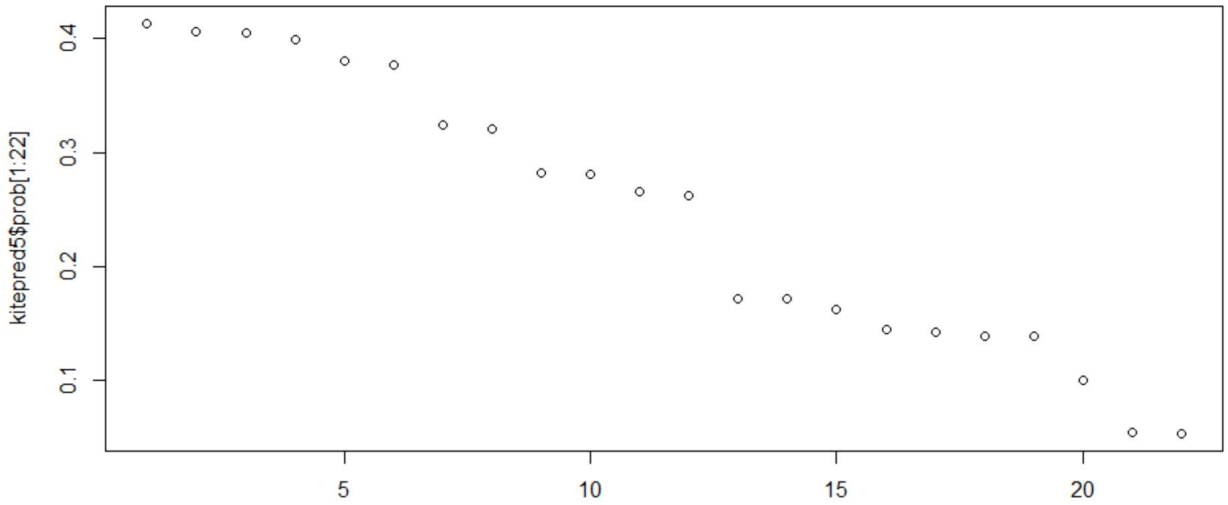


Similarly performing the deletion of 5% of edges for kite dataset:

Edges deleted: only one edge deleted

```
> vcount(kite)
[1] 10
> ecount(kite)
[1] 18
> seqedges<-fnseqedgesdel(kite,5)
> listedgesdel<-fnlistofedgesdel(kite,seqedges)
> kite5<-delete.edges(kite,seqedges)
> vcount(kite5)
[1] 10
> ecount(kite5)
[1] 17
> listedgesdel
+ 1/18 edge from 6b7ddad (vertex names):
[1] I--J
```

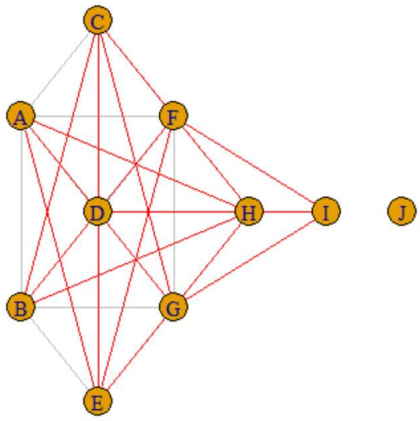
Predicted edges probability:



Adding top 15 edges from predicted edges and see whether we were able to recover the deleted edge.

Our deleted edge has less probability:

```
> ffinddeledgesinpred(kite5,listedgesdel,kitepred5)
Edge =  I|J  Probability= 0.03647065
```



We were not able to recover the deleted edge(no green edges in graph) between I and J as its probability is really less.

Similarly, we repeated the same process for 15% and 45% edges, there we could see prediction was higher rate for deleted edges.

Observation: Prediction of edges using the hierarchical random graph which will open new possibilities when applied onto real network as it can predict possible new linkages and with probabilities so analyzing such links will give some new insights.

Possible usage in finding new chemicals: When analyzing network of chemical structure of complex compounds, using this method will find new links between molecular structures which will be able to produce new chemicals compounds without any real experiments.