Network type: Peer to peer network

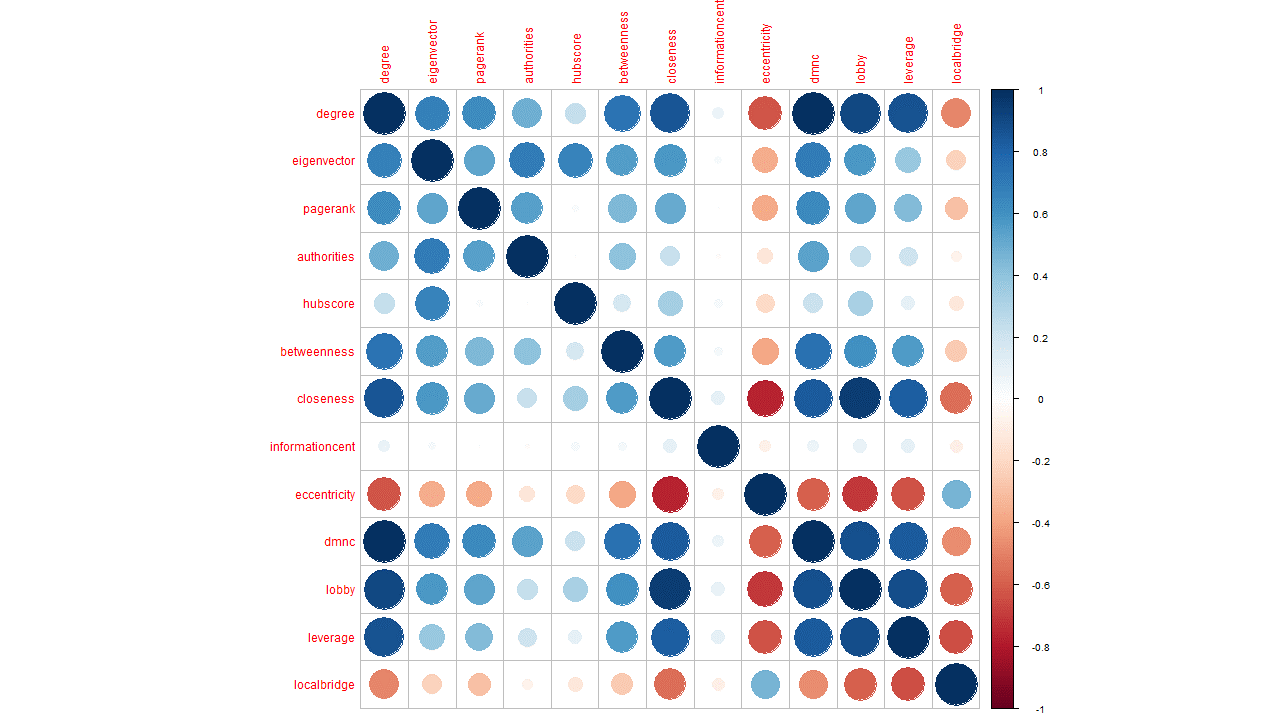
Characteristics:

* "peers" are computer systems which are connected to each other via the Internet.
* Files can be shared directly between systems on the network without the need of a central server.
* In other words, each computer on a P2P network becomes a file server as well as a client.

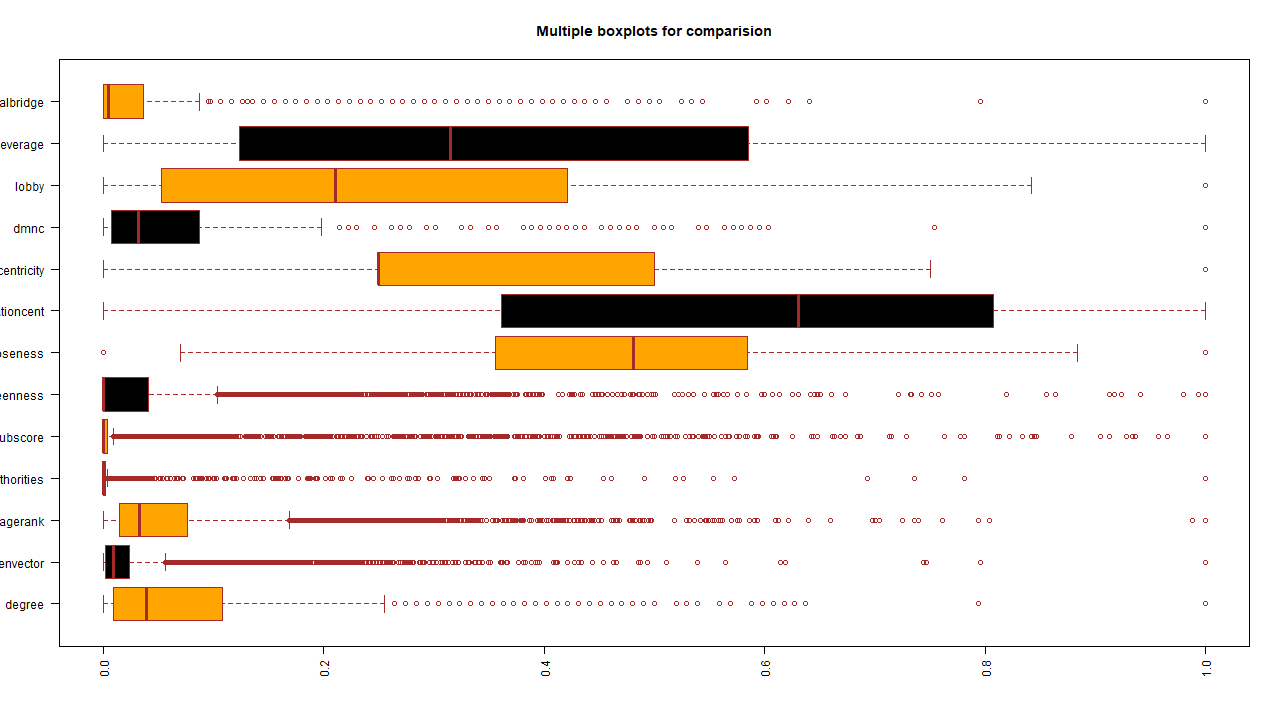
Results:

Peer-to-peer (P2P) networks are networks in which all nodes have the possibility of being connected to each other. This is due to the nature of the network, as edges are formed between nodes based on the availability of files and file segments. Users of P2P networks often share all kinds of files in these networks, hence it is one big community where all nodes can potentially connect to each other, rather than small cliques or closely connected communities.

Correlation plot:



From the correlation matrix it can be seen that, degree and DMNC are strongly positively correlated, while closeness and lobby index are also strongly positively correlated. On the other hand eccentricity is strongly negatively correlated to both closeness and lobby. This makes sense as nodes that have more files to share will have higher degree and hence be a part of a denser community structure. Nodes that have higher closeness value can be thought of as a more active node in the network which is supplying more files compared to others, hence it should also have higher lobby index as it has a bigger probability of having strongly connected neighbors.



The boxplot shows the distributions of the variables. It is quite evident that most of the variables are skewed, and only closeness seems to have a somewhat normal distribution. Eccentricity is heavily right skewed as the values are mostly 0.25 and 0.5 while a few are 0 and 0.75. As these are discrete values they are not well represented in the box plot. Hubscore and Authority clearly have very similar distributions which makes sense considering they are derived using almost similar formulae. Degree, DMNC and pagerank have similar distributions. (*Didn’t include the exact numbers i.e. range but can be included obviously)*

*m1\_all\_var13<-ncentrality*

*m2\_without\_dg\_var12<-within(ncentrality, rm(degree))*

*m3\_without\_comm\_var7<-within(ncentrality, rm(informationcent,eccentricity,dmnc,lobby,leverage,localbridge))*

*m4\_without\_nodes\_var6<-within(ncentrality, rm(degree,eigenvector,pagerank,authorities,hubscore,betweenness,closeness))*

*m5\_without\_ranks\_var9<-within(ncentrality, rm(eigenvector,pagerank,authorities,hubscore))*

*m6\_without\_dist\_var9<-within(ncentrality, rm(betweenness,closeness,informationcent,eccentricity))*

*m7\_mix\_match1\_var6<-within(ncentrality, rm(eigenvector,closeness,informationcent,dmnc,lobby,leverage,localbridge))*

*m8\_mix\_match2\_var6<-within(ncentrality, rm(degree,authorities,hubscore,betweenness,informationcent,eccentricity,leverage))*

*m9\_mix\_match3\_var6<-within(ncentrality, rm(closeness,eigenvector,authorities,hubscore,betweenness,leverage,informationcent))*

*m1<-m1\_all\_var13*

*m2<-m2\_without\_dg\_var12*

*m3<-m3\_without\_comm\_var7*

*m4<-m4\_without\_nodes\_var6*

*m5<-m5\_without\_ranks\_var9*

*m6<-m6\_without\_dist\_var9*

*m7<-m7\_mix\_match1\_var6*

*m8<-m8\_mix\_match2\_var6*

*m9<-m9\_mix\_match3\_var6*

PCA:

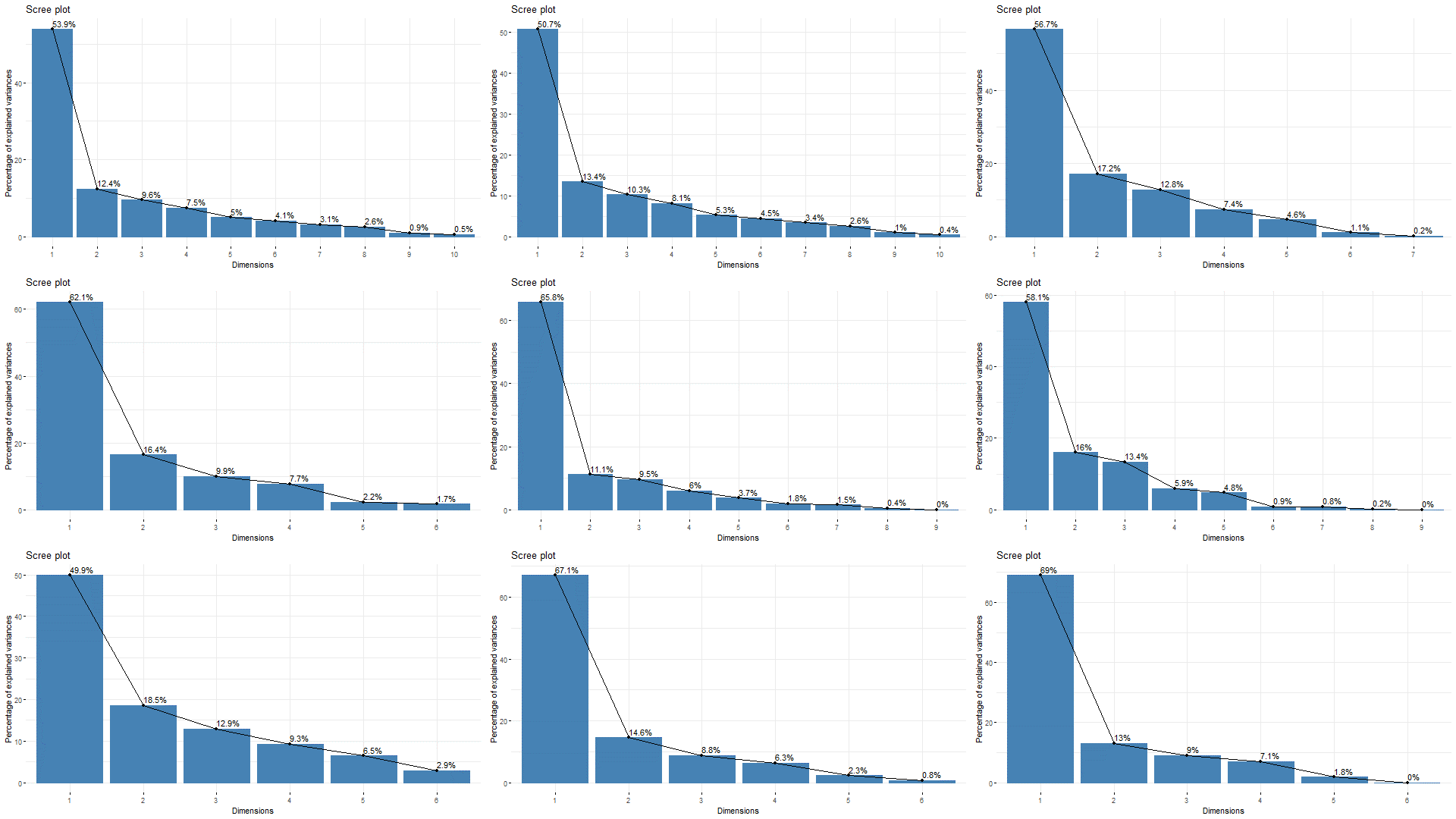
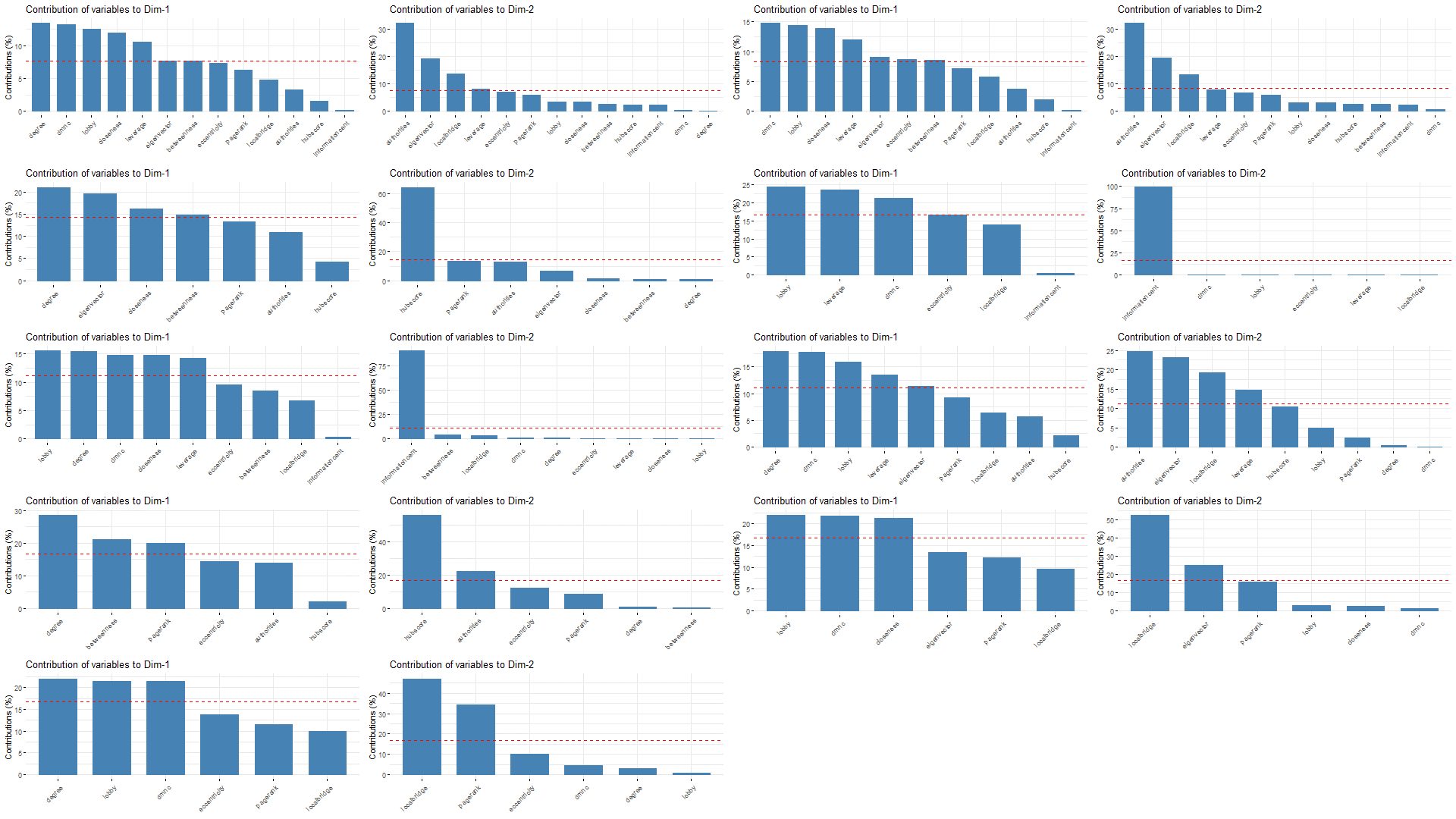
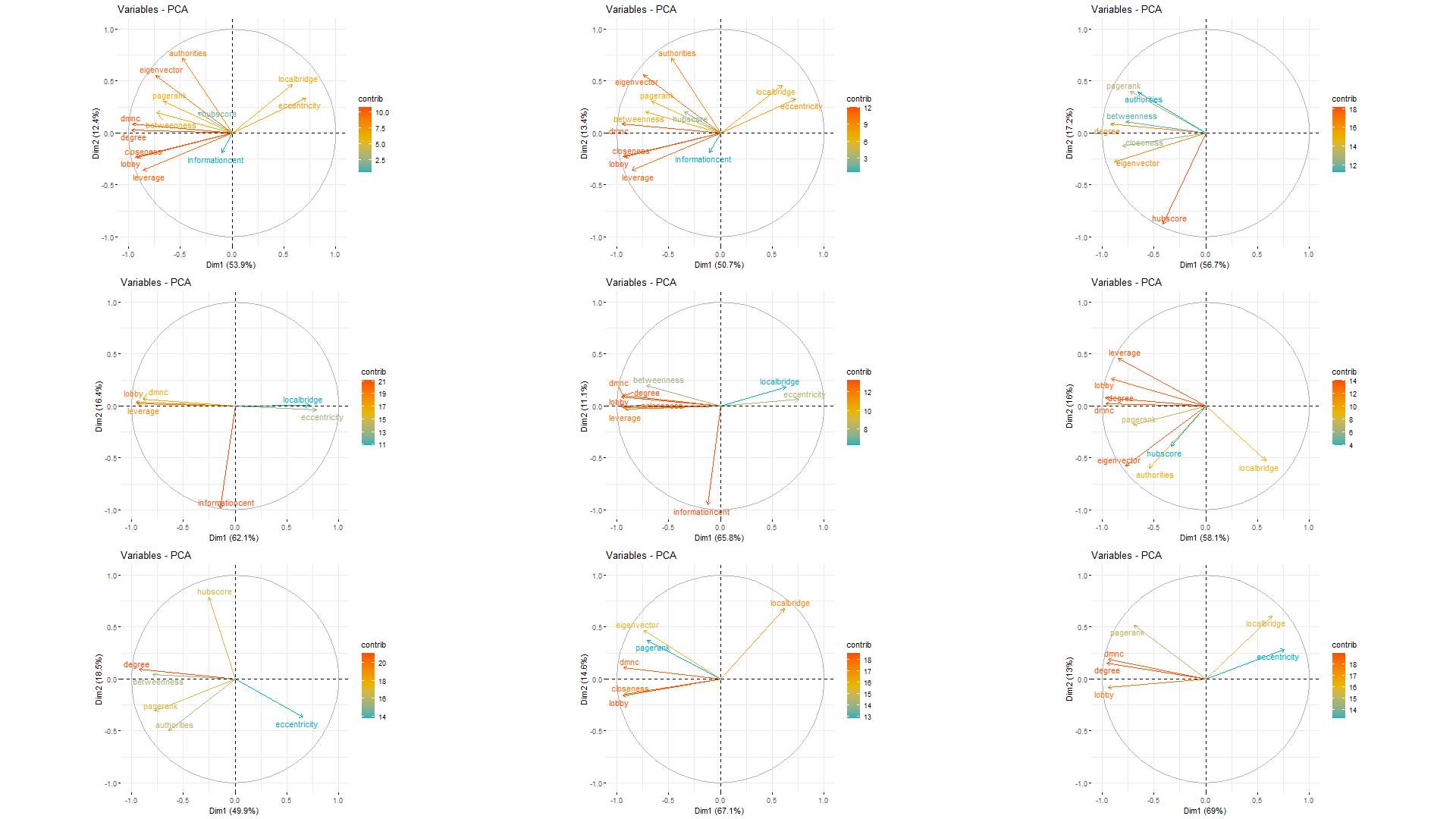


Figure 1 main model

In the main model, the first principal component shows about 53.9% variation, while the first 2 principal components shows about 66.3% variation and finally, taking the first three principal components together capture about 76% variance of the dataset.

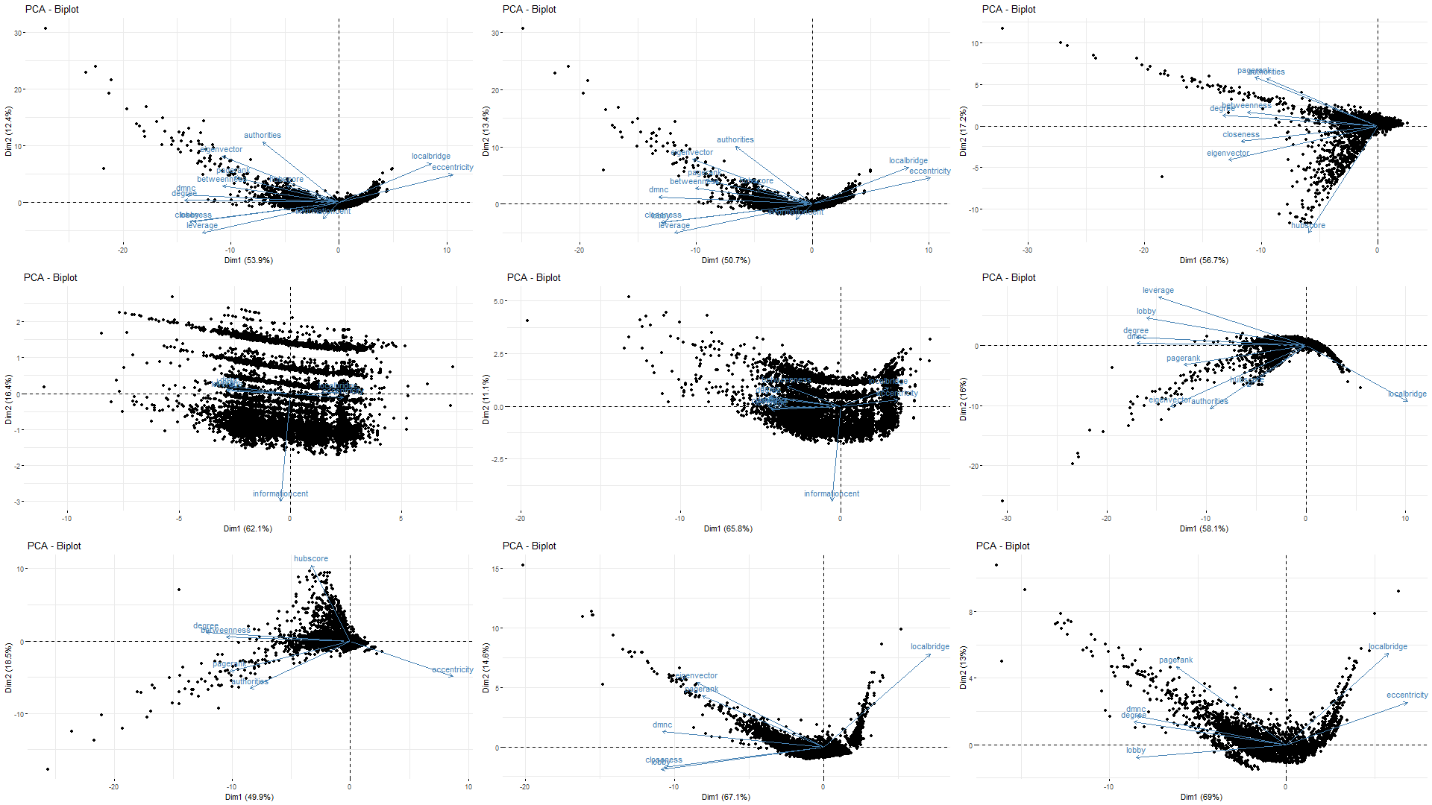


From the dimension contribution plots it can be seen that Degree, DMNC, Lobby, Closeness and leverage are the more significant factors contributing to principal component 1. In principal component 2, authorities, eigenvector and local bridging centrality contributes more.

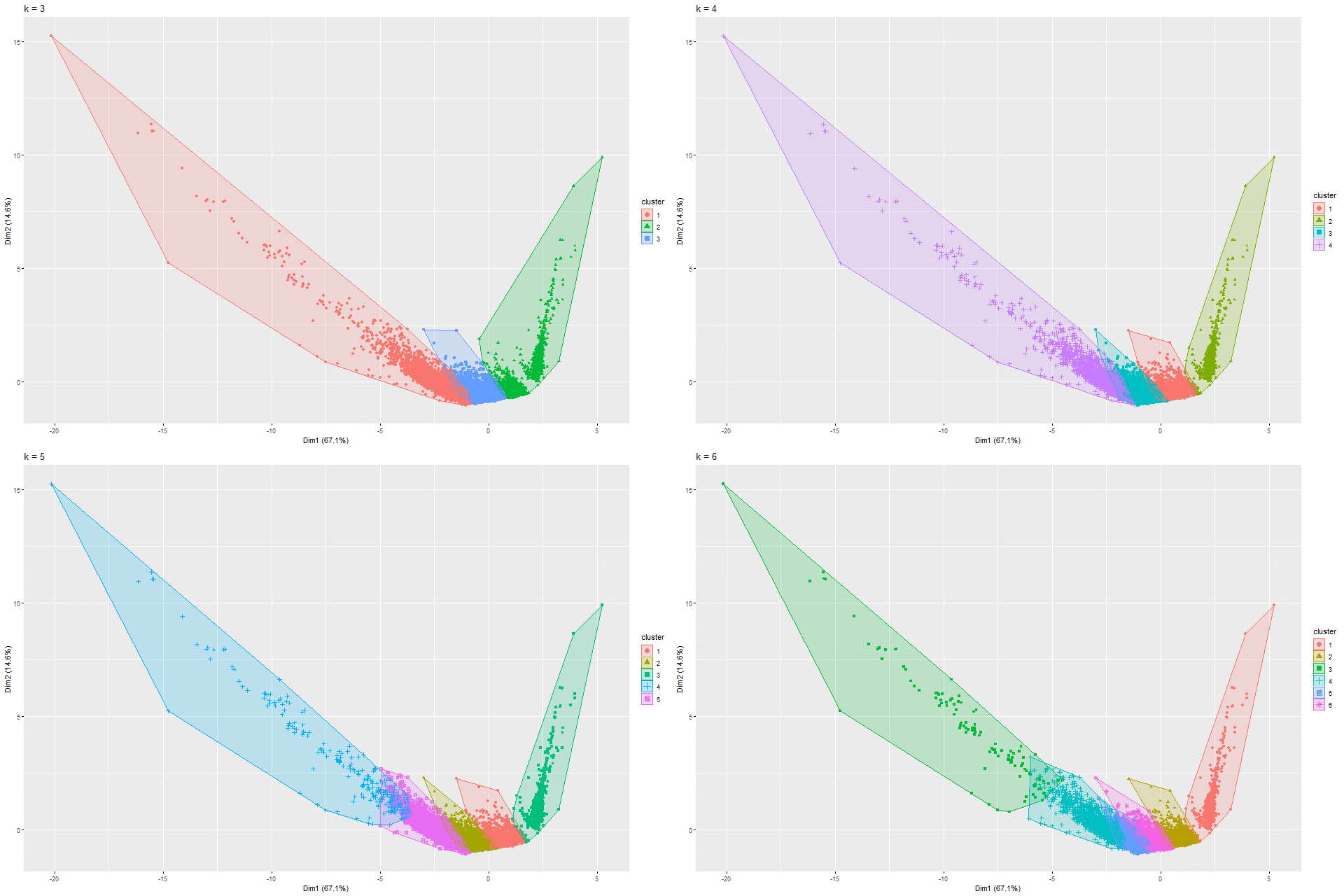
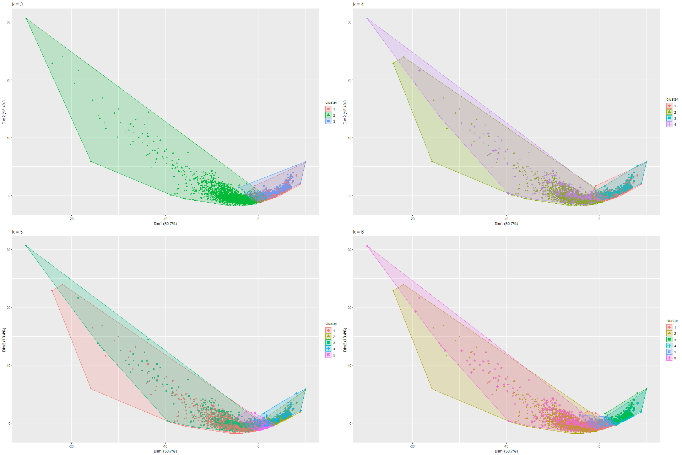
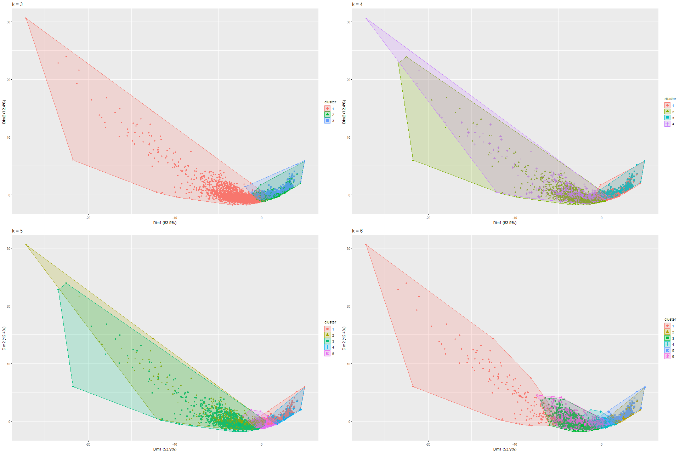


From the cos2 contribution plot it can be seen that the principal component 1 captures the most variation of Degree, DMNC, Lobby, Closeness and leverage while principle component 2 captures most variation of authorities, eigenvector and local bridging centrality which is consistent with the dimension contribution plots. The length and color of the variable signifies the importance/contribution of the variables, and hence in the main model degree, DMNC, lobby, closeness, leverage, eigenvector and authorities contribute the most.

It cannot be expected to identify a model that captures the variation and structure of the main model based on scree plots of cos2 contribution plots of the ablation study model as these are dependent on the centralities being used in the PCA.



From the scree plot it can be seen that the main model and model 2 have similar structures, which is a given considering the only difference is model 2 does not have degree. However, model 8 and model 9 also have somewhat of a similar structure which is farther proved by the K-means on PCA plots below. This does make sense considering model 8 and 9 contain some community variables and some node centrality variables, hence retaining the actual structure of the main model. Hence they are possible options for feature selection.



From the calinhara index it is seen that the optimal number of clusters for the main model is 4, for model 8 and model 9 is 3. Following this, it can be seen the main model gives 2 distinct clusters with minimal overlap. Model 8 has 2 overlapping clusters which cannot be used for distinguishing, but model 9 gives 3 distinct cluster with virtually no overlap. This can prove to be useful in distinguishing P2P networks.