Insanity check of the k-mean algorithm

The estimates of alpha

The blue line is the MLE global estimates for alpha, the red line is the E2 global estimates for alpha. The black dots are E2 estimates for the clusters. The blue dots are the MLE estimates, which contain several negative values that were not shown in the figure.

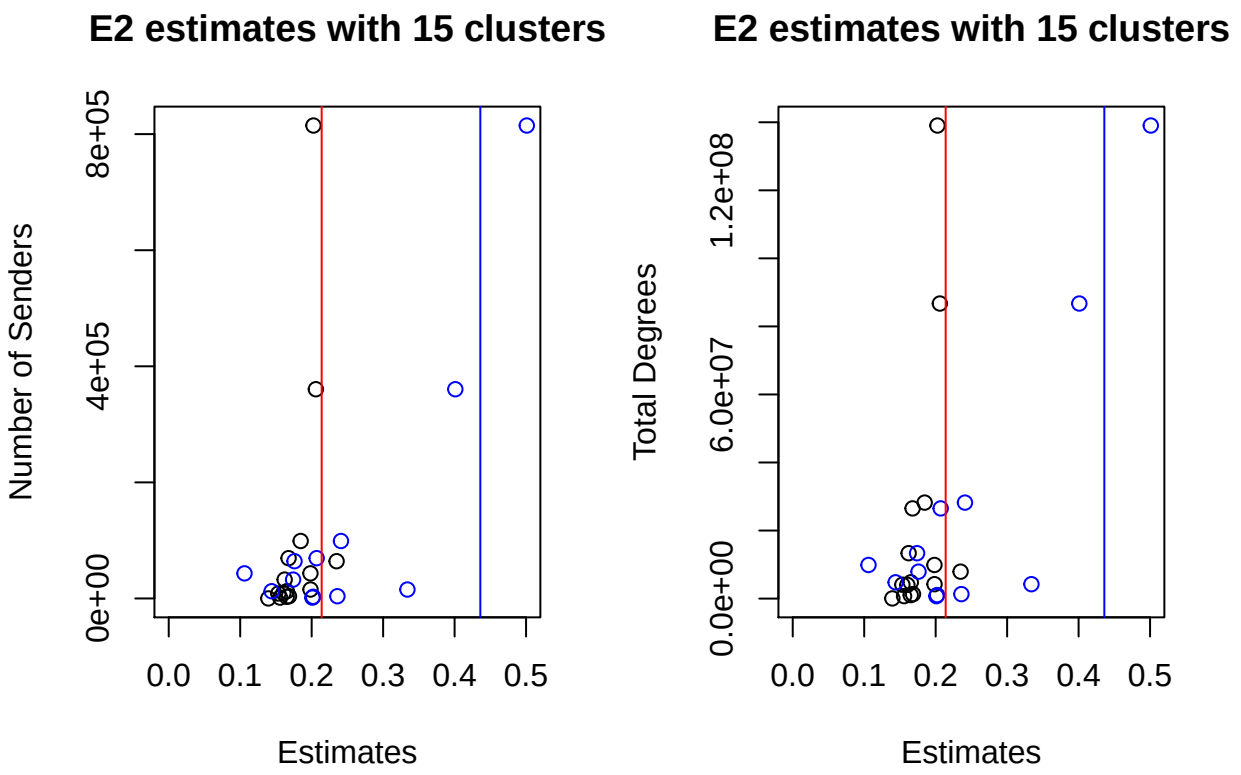
The MLE estimates were calculated based on $p(x) \propto (x + x_0)^{-\alpha} e^{-\beta x}$

```
est=c(0.2059,0.2025,0.1622,0.1677,0.1652,0.1846,0.2349,0.1604,0.1683,0.1983,0.1651,0.1534,0.1394,0.1558
ss=c(360469, 814799,32498,69411,12443,98990,64172,9424,3875,43248,2876,8577,65,1464,15415)
rs=c(86729503,139042471,13341127,26504656,4780666,28216142,7914563,4022417,1315008,9884241,1031520,4077
true=0.214

file=read.table(file="KR_k_mean2.log",header=F,sep="\n")
index1=seq(1,423,3)
index2=seq(2,423,3)
index3=seq(3,423,3)
est2=file[index3,]
tmp_str<-unlist(strsplit(as.character(est2)," "))
tmp_str<-tmp_str[which(tmp_str!="]"&tmp_str!="["&tmp_str!="")]
library(stringr)
tmp_str<-str_replace(tmp_str,"\\[","")
tmp_str<-str_replace(tmp_str,"\\]","")
tmp_str<-matrix(as.numeric(tmp_str),ncol=2,byrow = T)
est2<-c(tmp_str[,1])
ss2<-as.numeric(file[index1,])
rs2<-as.numeric(file[index2,])

MLE<-c(0.401,0.501,0.174,0.207,0.144,0.241,0.176,-1.063,0.236,0.106,0.202,-3.1,-0.49,0.201,0.334)
MLE2<-read.table(file="KR_MLE2.log",header=F,sep="\n",stringsAsFactors = FALSE)
MLE2[is.na(MLE2),1]="[NA NA NA]"
MLE2<-MLE2[index3,]
tmp_str<-unlist(strsplit(as.character(MLE2)," "))
tmp_str<-tmp_str[which(tmp_str!="]"&tmp_str!="["&tmp_str!="")]
tmp_str<-str_replace(tmp_str,"\\[","")
tmp_str<-str_replace(tmp_str,"\\]","")
tmp_str<-matrix(as.numeric(tmp_str),ncol=3,byrow = T)
```

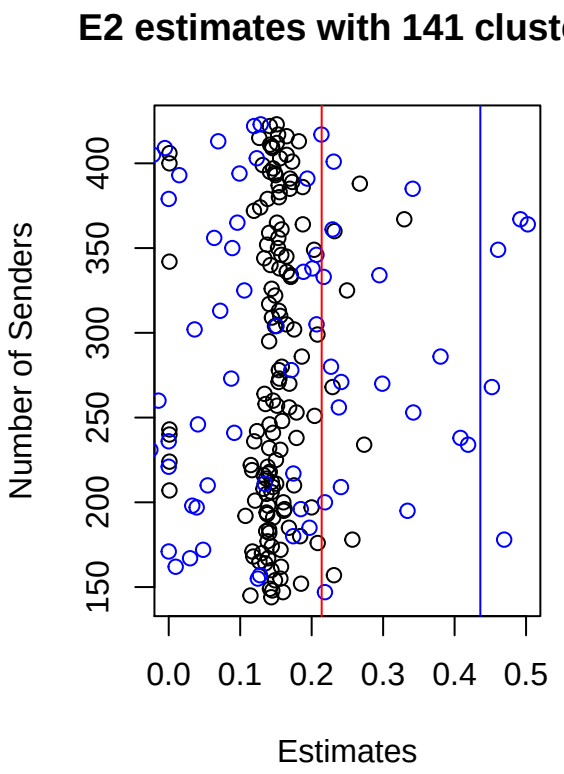
```
## Warning in matrix(as.numeric(tmp_str), ncol = 3, byrow = T): NAs introduced
## by coercion
```

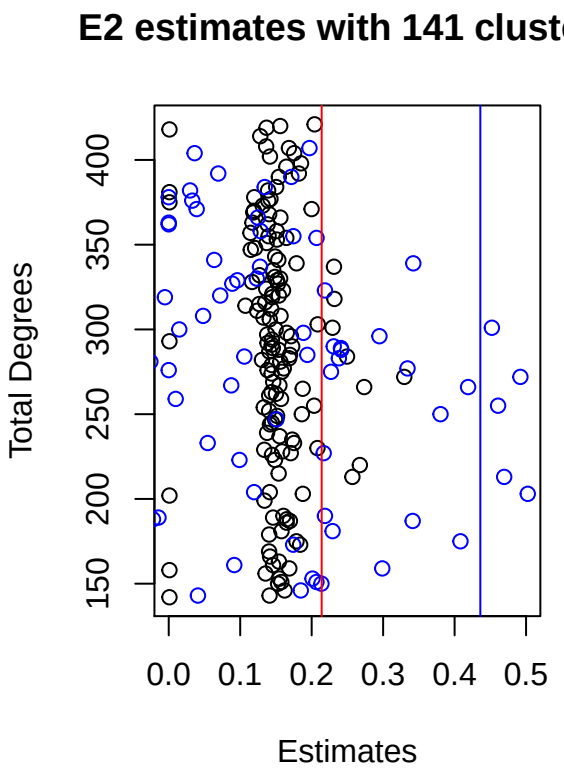
```
MLE2<-c(tmp_str[,1])-1

par(mfrow=c(1,2))
plot(est,ss,xlim=c(0,0.5),xlab="Estimates",ylab="Number of Senders",main = "E2 estimates with 15 cluste
points(MLE,ss,xlab="Estimates",col="blue")
abline(v=0.214,col="red")
abline(v=0.436,col="blue")

plot(est,rs,xlim=c(0,0.5),xlab="Estimates",ylab="Total Degrees",main = "E2 estimates with 15 clusters")
points(MLE,rs,xlab="Estimates",col="blue")
```

```
abline(v=0.214,col="red")
abline(v=0.436,col="blue")
```



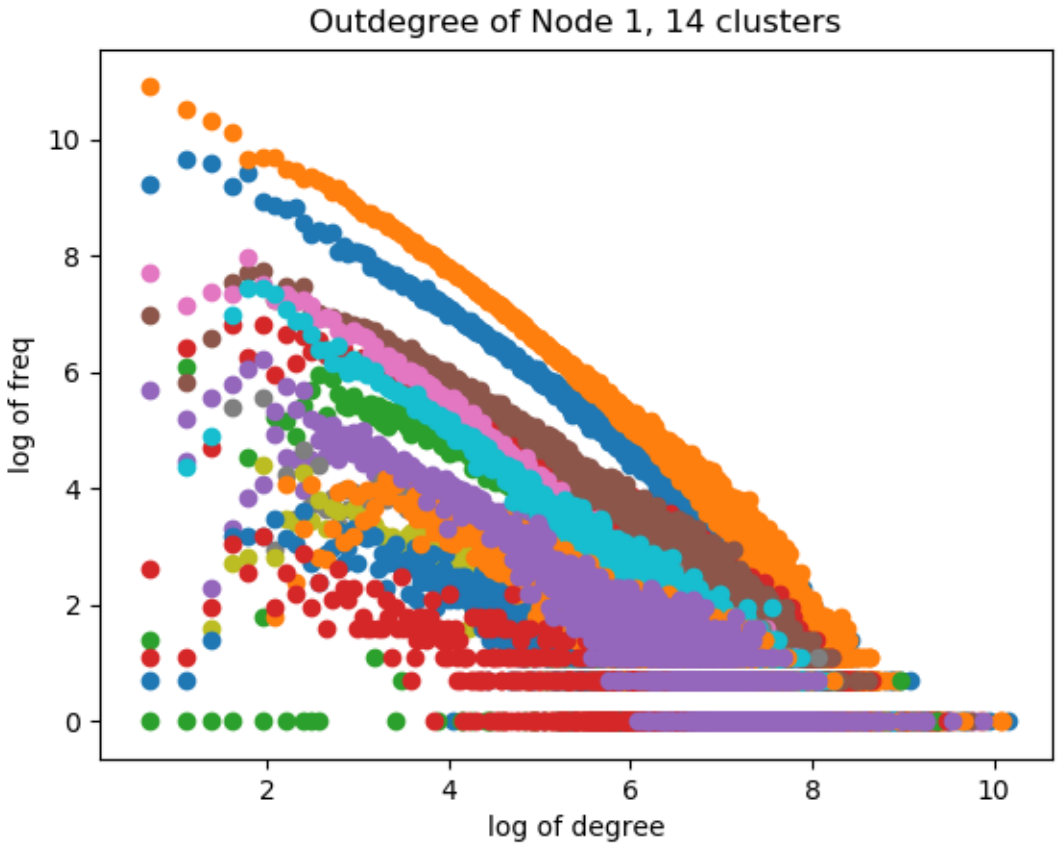

```
plot(est2,ss2,xlim=c(0,0.5),xlab="Estimates",ylab="Number of Senders",main = "E2 estimates with 141 clus
points(MLE2,ss2,xlab="Estimates",col="blue")
abline(v=0.214,col="red")
abline(v=0.436,col="blue")

plot(est2,rs2,xlim=c(0,0.5),xlab="Estimates",ylab="Total Degrees",main = "E2 estimates with 141 clusters
points(MLE2,rs2,xlab="Estimates",col="blue")
abline(v=0.214,col="red")
abline(v=0.436,col="blue")
```





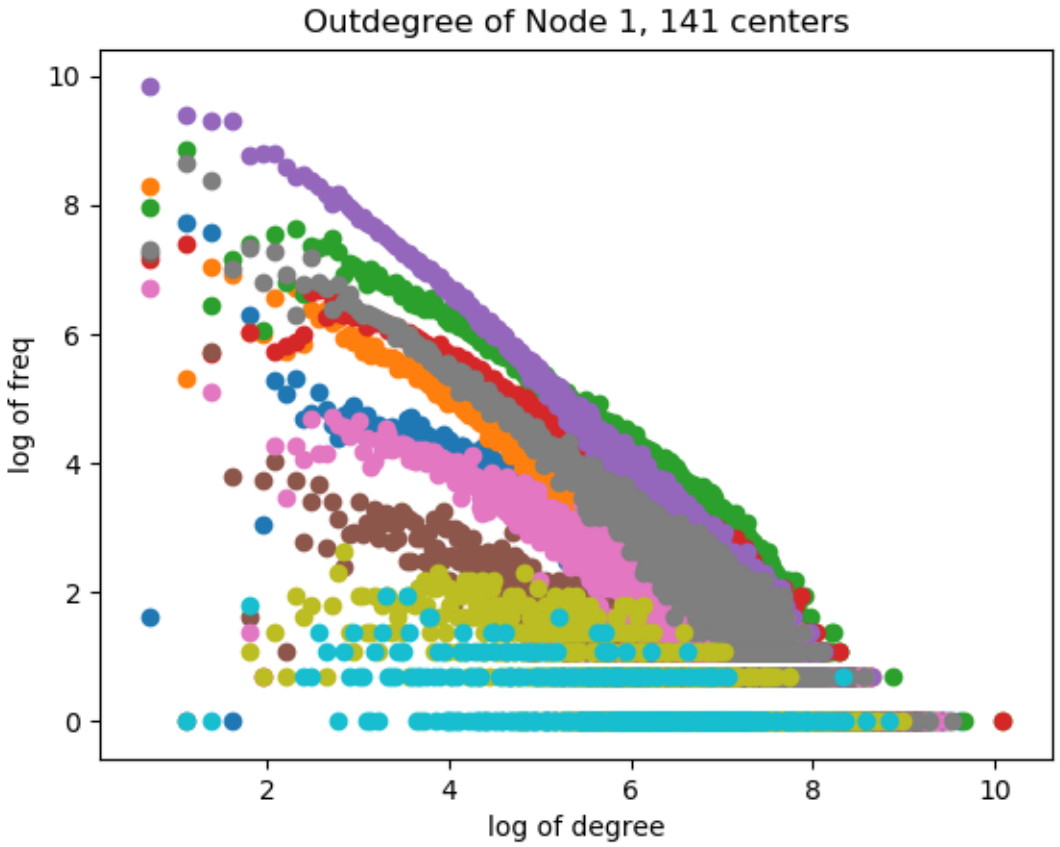
The degree distribution

log-log plots of the 15 clusters:



{placeHere}

log-log plots of the 141 clusters, the first 10 clusters:



{placeHere}

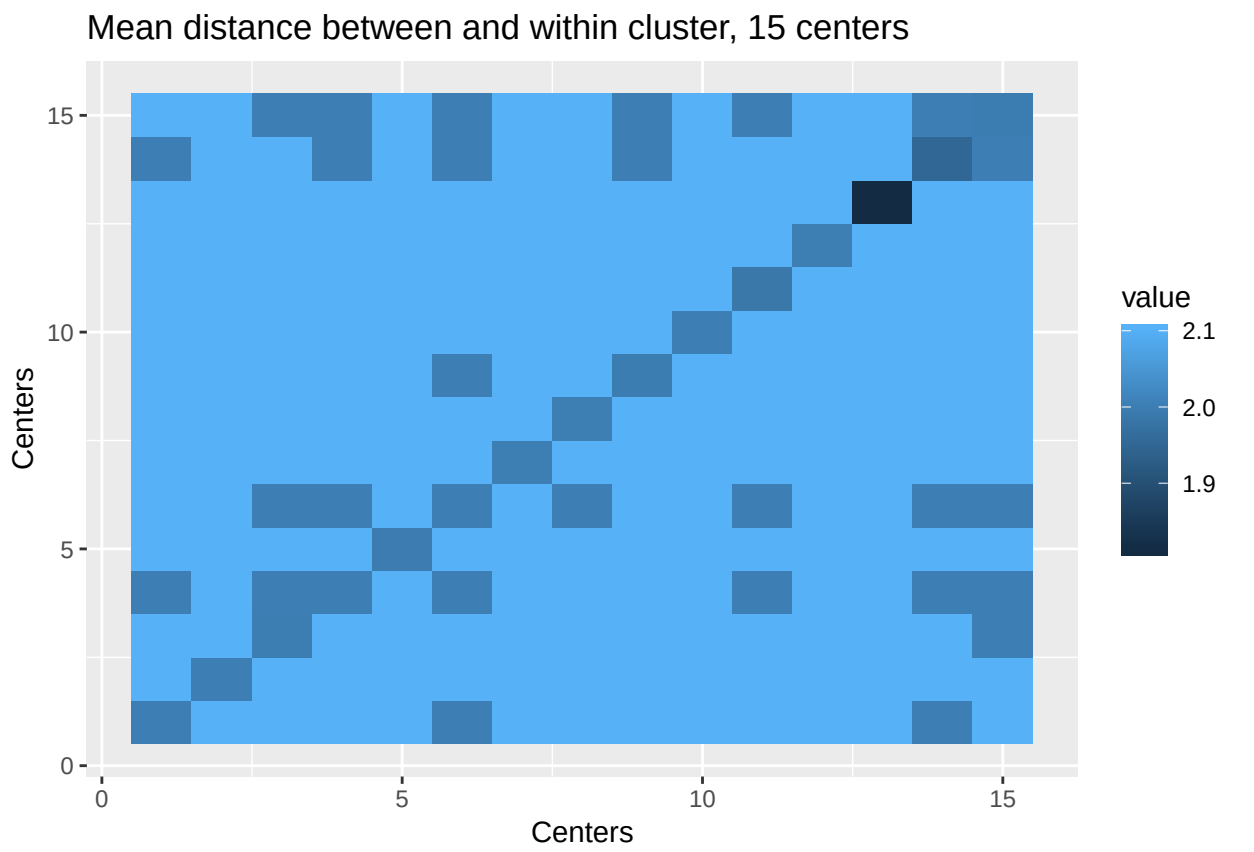
Mean distance within and between clusters:

The distance is defined as the distance between the focal node and the nodes that were selected to be the centers. The mean was defined as the average distance over the nodes within a cluster. Notice **that most of the nodes actually are not similar to any of the centers**. In this case, they are mostly likely to be assigned to the first clusters based on my implementation of the algorithm. That's why the size of the first two clusters are very large.

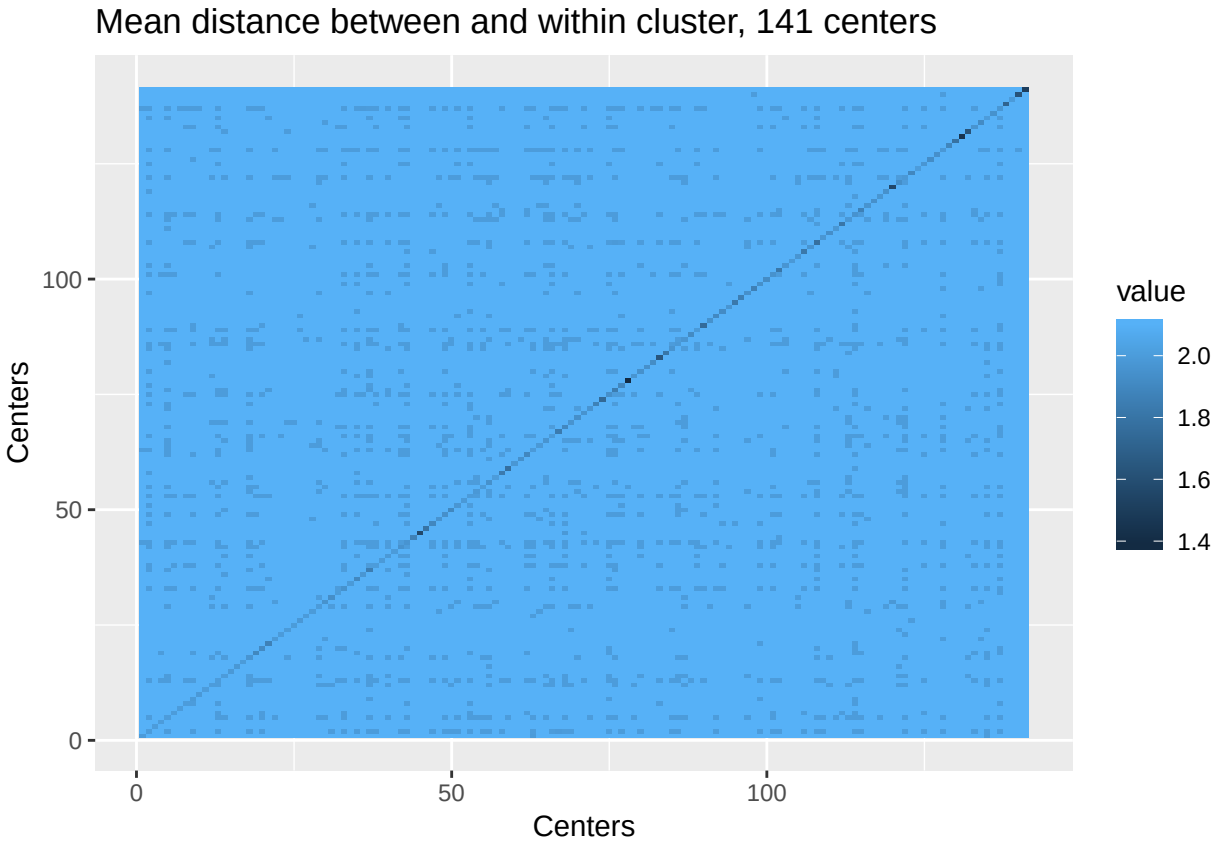
NAs and the distance exactly equals to 2 were set to be 2.1, because the within cluster distance is very close to 2 (The difference of two norm 1 vectors should range between 0 to 2), and the visualization is not very good if the value is not changed.

```
library(ggplot2)
library(gridExtra)
mat=read.table(file="mat1.txt",header=F)
mat[mat==2]=2.1
mat2=read.table(file="mat21,txt",header=F)
mat2[is.na(mat2)]<-2.1
mat2[mat2==2]=2.1
plot=data.frame(V1=rep(c(1:15)),V2=rep(1:15,each=15),value=c(unlist(mat[,1:15])))
p1<-ggplot(plot,aes(V1,V2,fill=value))+geom_tile()+xlab("Centers")+ylab("Centers")+ggtitle("Mean distan
plot=data.frame(V1=rep(c(1:141)),V2=rep(1:141,each=141),value=c(unlist(mat2[,1:141])))
p2<-ggplot(plot,aes(V1,V2,fill=value))+geom_tile()+xlab("Centers")+ylab("Centers")+ggtitle("Mean distan
```

```
p1
```


Mean distance between and within cluster, 15 centers
Centers
Centers
0
5
10
15
value
2.1
2.0
1.9

```
p2
```


Mean distance between and within cluster, 141 centers
Centers
100
50
0
0
50
100
Centers
value
2.0
1.8
1.6
1.4