STAT 2221 - Homework 5 - Spectral clustering & random graph models

true

October 2021

Part I (stochastic blockmodels)

```
# Preliminaries
library(igraph) # package for network analysis & random graphs
library(irlba) # package for numerical linear algebra routines
library(mclust) # package for Gaussian mixture model clustering
library(rgl)
options(scipen = 999)
```

Consider the following setup that generates a random stochastic blockmodel (SBM) graph, saved as *my.graph*.

```
# SBM model 1

n <- 5000
rho <- log(n)/n
aa <- 4.5
bb <- 0.25
cc <- 4.5
block.sizes = c(1/2, 1/2)*n

# above - specified inputs
########################################
# below - automated inputs/outputs

pref.matrix <- rho*rbind(c(aa, bb),
                         c(bb, cc)) #this is matrix B - prob of edge between group i and j

set.seed(1234)

my.graph <- sample_sbm(n, pref.matrix, block.sizes, directed = FALSE, loops = TRUE)


#2 leading eigenvals and vectors of adjacency matrix
adj_ed <- embed_adjacency_matrix(my.graph, 2, which="lm", scaled = F)
adj_ed$D
```

```
## [1] 21.37299 19.31311
```

```
head(adj_ed$X)
```

```
##             [,1]        [,2]
## [1,] 0.01666704 -0.01757226
## [2,] 0.01138190 -0.01014940
## [3,] 0.01363640 -0.01374932
## [4,] 0.01345882 -0.01551609
## [5,] 0.01687230 -0.01567405
## [6,] 0.01426462 -0.01418789
```

```
#2 smallest eigenvals and vectors of laplacian
lap_ed <- embed_laplacian_matrix(my.graph, 2, which = "sa", scaled = F)
lap_ed$D
```

```
## [1] 0.000000 1.950716
```

```
head(lap_ed$X)
```

```
##             [,1]        [,2]
## [1,] 0.01414214 -0.01459176
## [2,] 0.01414214 -0.01387669
## [3,] 0.01414214 -0.01499625
## [4,] 0.01414214 -0.01675683
## [5,] 0.01414214 -0.01291442
## [6,] 0.01414214 -0.01393429
```

```
#testing different egein decomp functions
# test2 <- spectrum(my.graph, which = list(pos="LM", howmany=2))
#test3 <- irlba(as_adj(my.graph),3,smallest = F)
```

(0) Examine the eigenstructure of the adjacency matrix and of graph Laplacians corresponding to *my.graph*. Discuss what aspects of the graph are revealed by its eigenvalues and eigenvectors. Namely, what is the 'ground truth' underlying the generated data?

Based on the way the data is generated from the given code, we can see that he ground truth about the generated data is that the first 2500 vertices are from first cluster, and the last 2500 vertices are from the second cluster. We have 2 clusters of equal size in this first model. Therefore, it's enough to look at only the 2 leading (largest) eigenvalues and their associated eigenvectors for the adjacency matrix, or the 2 smallest eigenvalues and their associated eigenvectors for the graph Laplacian. Those eigenvalues and (head of) eigenvectors are displayed above.

For all SBM models provided below, including the one above, do the following:

(1) Cluster the eigenvectors of the graphs' adjacency matrices using *kmeans* and *Mclust*.

Below is the process I went through for this problem. I defined a master function that takes inputs of $n, \rho, aa, bb, cc, block.sizes$. From the inputs the function will form the B matrix of independent Bernoulli, and then generate a sample of stochastic block model. The adjacency matrix is generated and its 2 leading eigenvectors are then calculated. After that, the function will apply both k-means method and Gaussian mixtrue model (Mclust) on the matrix formed by those 2 eigenvectors.

(2) Compute the adjusted rand index (ARI) between the clusterings you obtain and the ground truth clusterings.

The master function I defined above calculates the ARI between the clusterings from ground truth versus the clusterings from k-means and MClust. It will output these ARI scores in 2-column table.
Below you can find the detail of the function, as well as the results when it's applied to all of the models we have here.

(3) Are your findings surprising? Discuss. Describe the ways in which the SBM models are related to and differ from one another.

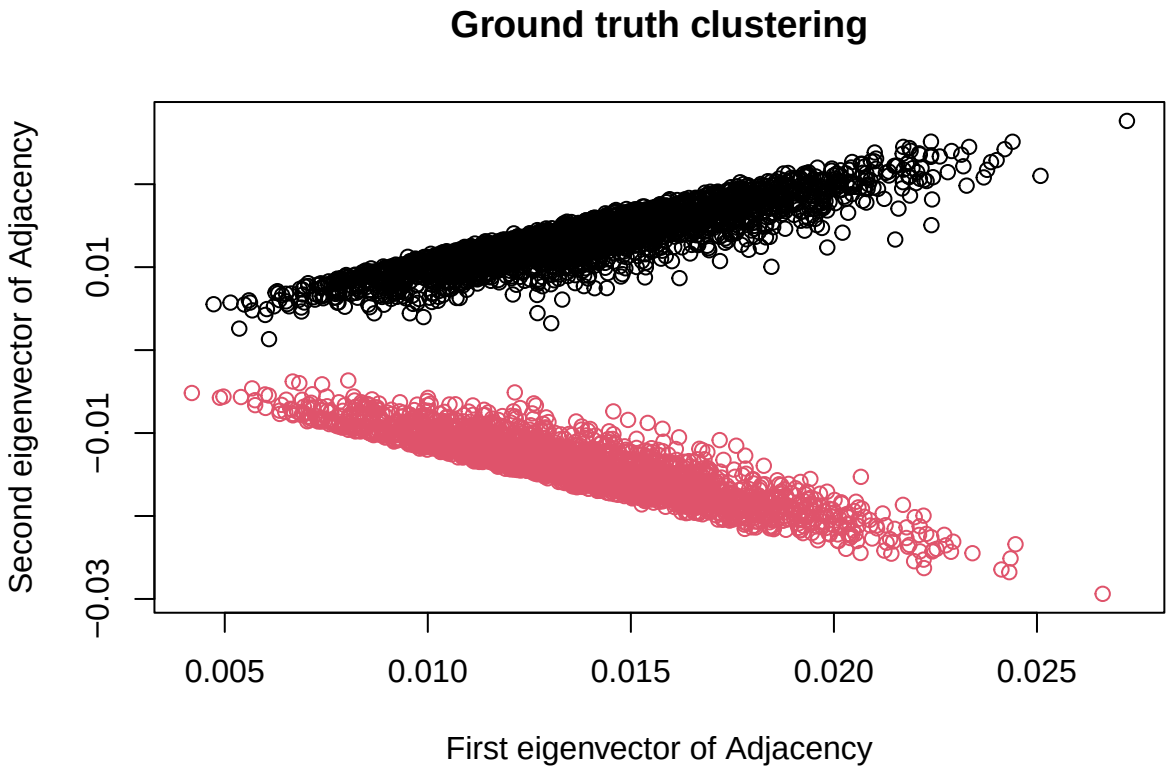
My findings are not surprising, because they agree with what I learnt in lecture (Lei, Rinaldo 2015).
When we apply k-means on the matrix form by the 2 leading eigenvectors of the adjacency matrix, we are carrying out community discovery using spectral clustering. From the lecture, we learnt that whether the discovery is easy or not depends on the number of communities and the community size imbalance.

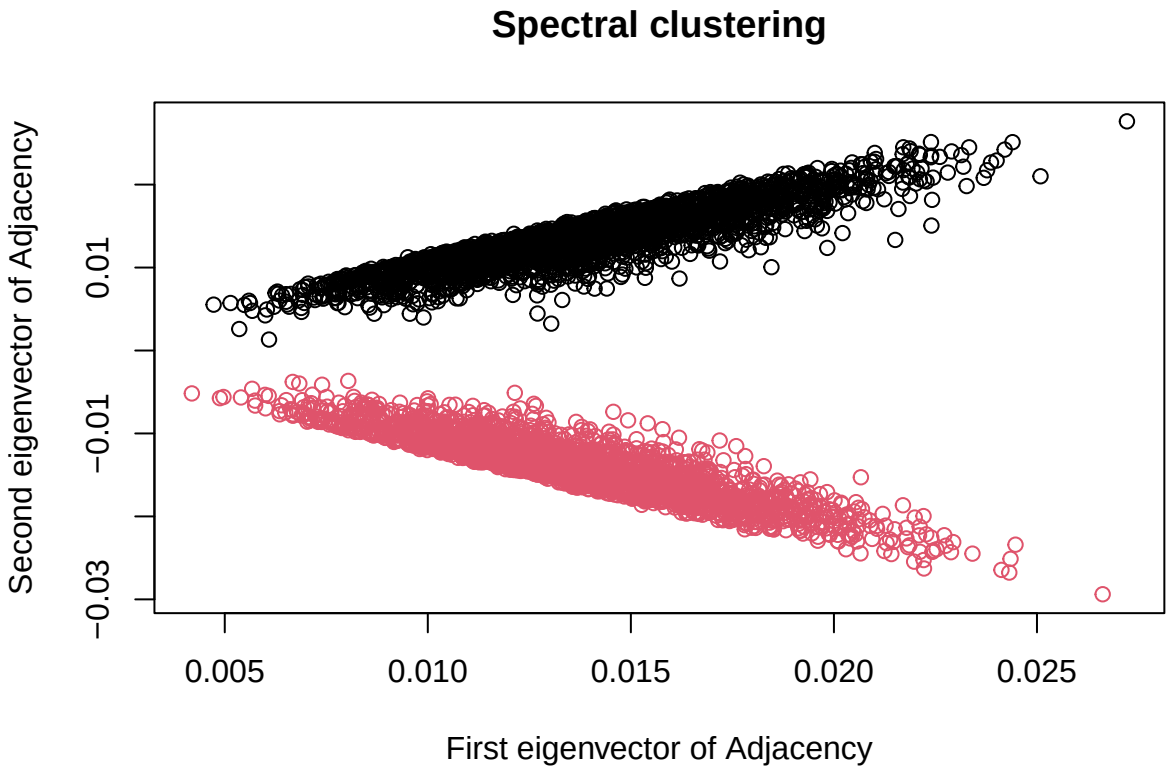
(4) Notice that the methodology in Priebe et al. (2019) differs from the von Luxburg survey! Explain and discuss, with examples.

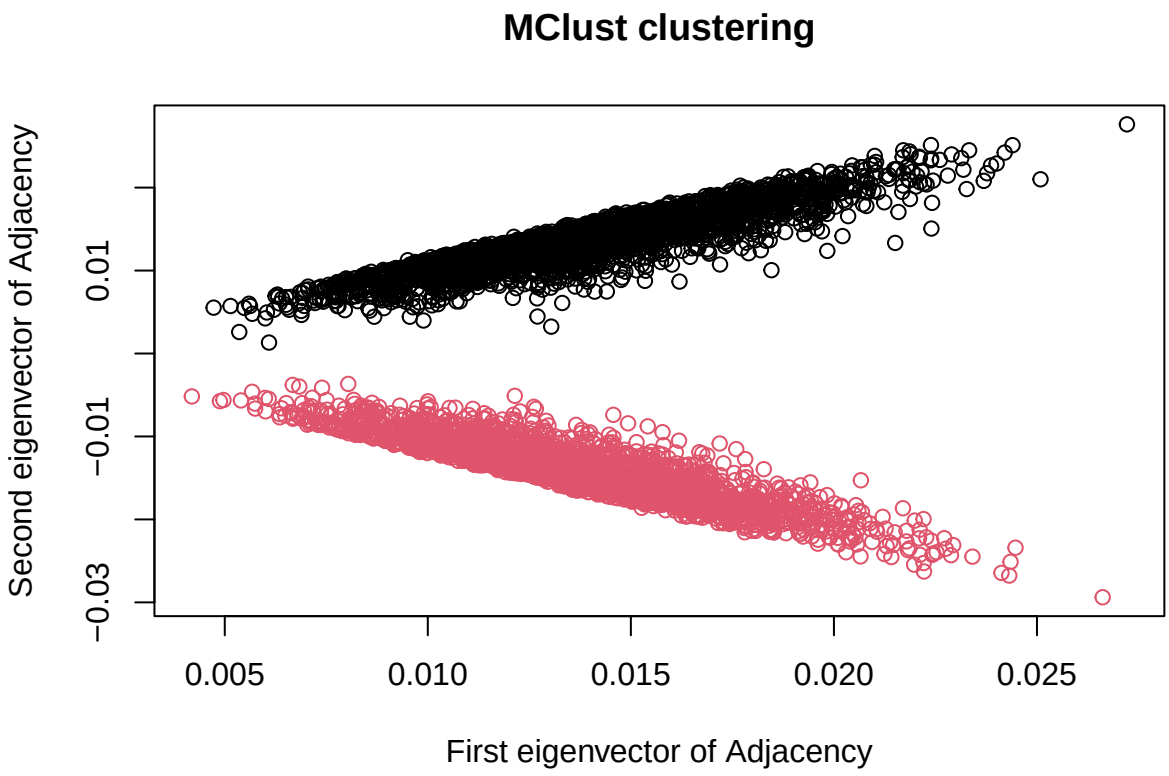
SBM model 1

```
#master function
master_sbm <- function(n, rho, aa, bb, cc, block.sizes){
  #matrix B
  pref.matrix <- rho*rbind(c(aa, bb),
                           c(bb, cc))
  set.seed(1234)
  #generate graph
  my.graph <- sample_sbm(n, pref.matrix, block.sizes, directed = FALSE, loops = TRUE)
  #generate ground truth clustering
  true_block <- rep(1:2, c(block.sizes))
  #eigen decomp of adjacency
  #adj_ed <- embed_adjacency_matrix(my.graph, 2, which="lm", scaled = F)
  adj_ed <- partial_eigen(as_adj(my.graph), n = 2)
  #k means on leading 2 eigenvectors of A
  km <- kmeans(adj_ed$vectors, centers = 2)
  #gaussian mixture on leading 2 eigenvectors of A
  gm <- Mclust(adj_ed$vectors, verbose = F, G=2)
  #ARI of kmeans vs ground truth
  ARI.km <- compare(as_membership(km$cluster), as_membership(true_block), method = "adjusted.rand")
  #ARI of gaussian mixture vs ground truth
  ARI.gm <- compare(as_membership(gm$classification), as_membership(true_block), method = "adjusted.ran
  result <- list()
  result[[1]] <- data.frame(ARI_k_means = ARI.km, ARI_mclust = ARI.gm)
  plot(adj_ed$vectors, col = true_block, xlab = "First eigenvector of Adjacency",
       ylab = "Second eigenvector of Adjacency", main = "Ground truth clustering")
  result[[2]] <- recordPlot()
  plot(adj_ed$vectors, col = km$cluster, xlab = "First eigenvector of Adjacency",
       ylab = "Second eigenvector of Adjacency", main = "Spectral clustering")
  result[[3]] <- recordPlot()
```

```
  plot(adj_ed$vectors, col = gm$classification, xlab = "First eigenvector of Adjacency",
       ylab = "Second eigenvector of Adjacency", main = "MClust clustering")
  result[[4]] <- recordPlot()
  return(result)
}

master_sbm(n, rho, aa, bb, cc, block.sizes)
```




Spectral clustering
Second eigenvector of Adjacency
0.01
−0.01
−0.03
0.005
0.010
0.015
0.020
0.025
First eigenvector of Adjacency



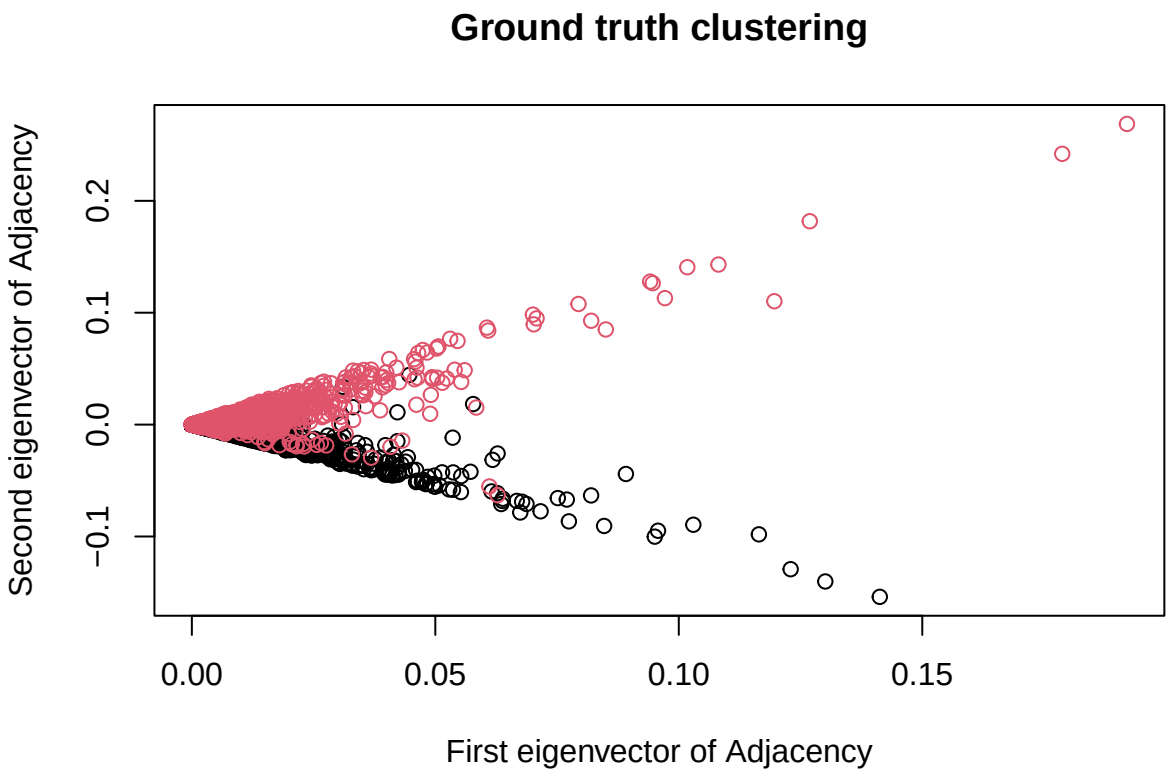

```
## [[1]]
##   ARI_k_means ARI_mclust
## 1           1          1
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

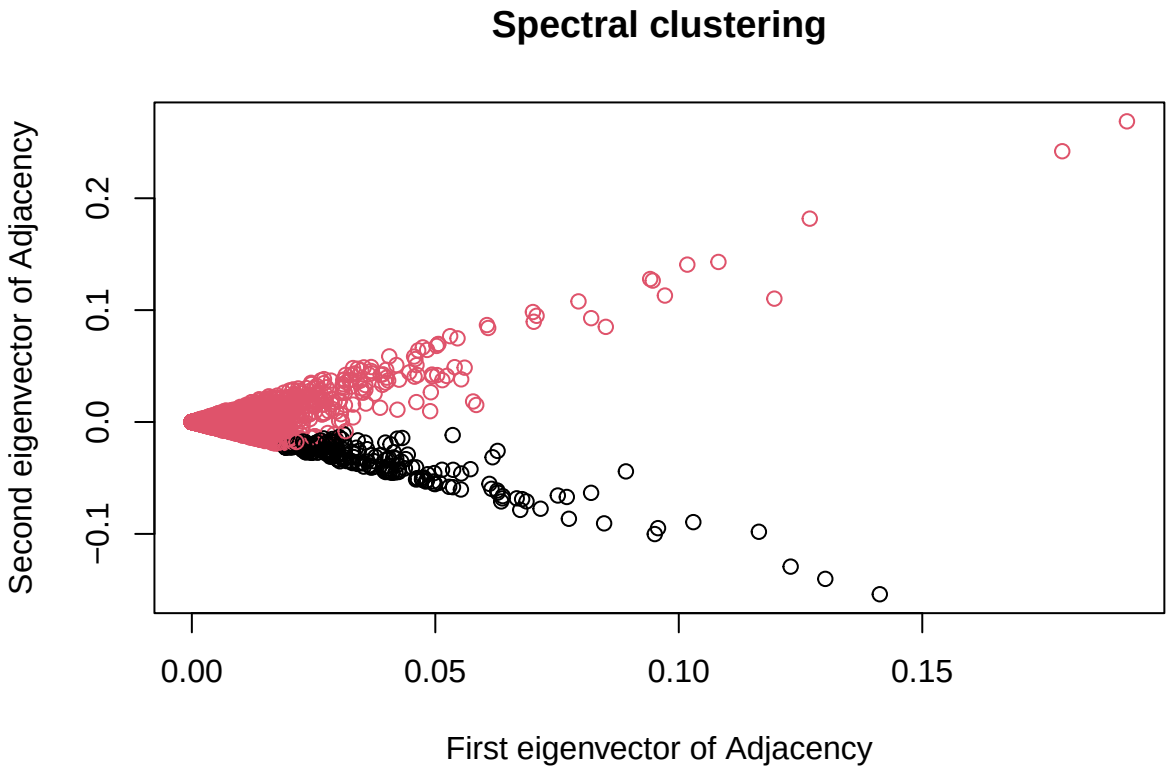
```
# adjustedRandIndex(c(rep(1,2500), rep(2,2500)),m1.k$cluster)
# adjustedRandIndex(c(rep(1,2500), rep(2,2500)),m1.g$classification)
# adjustedRandIndex(m1.k$cluster, m1.g$classification)

#test
# library(randnet)
# test <- reg.SSP(as_adj(my.graph), K=2)
# plot(my.graph, vertex.color=test$cluster)
```

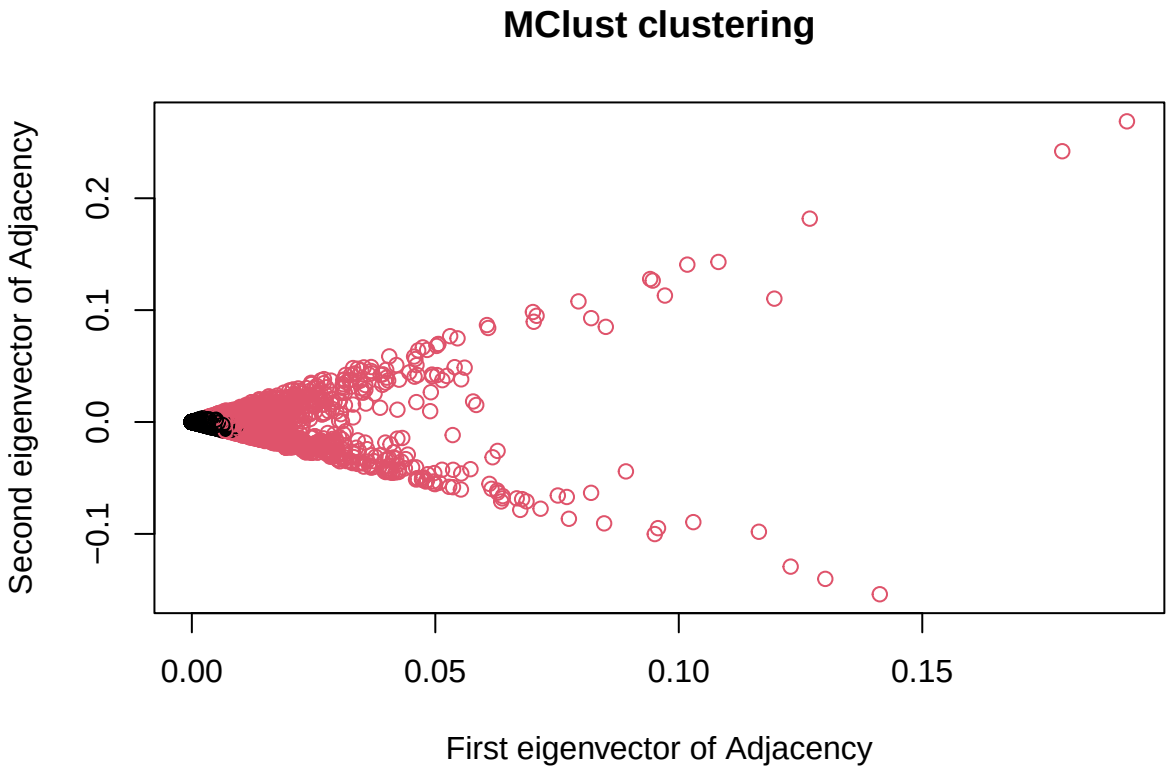
SBM model 2

```
# SBM model 2
n <- 5000
```

```
rho <- 1/n
aa <- 4.5
bb <- 0.25
cc <- 4.5
block.sizes = c(1/2, 1/2)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```



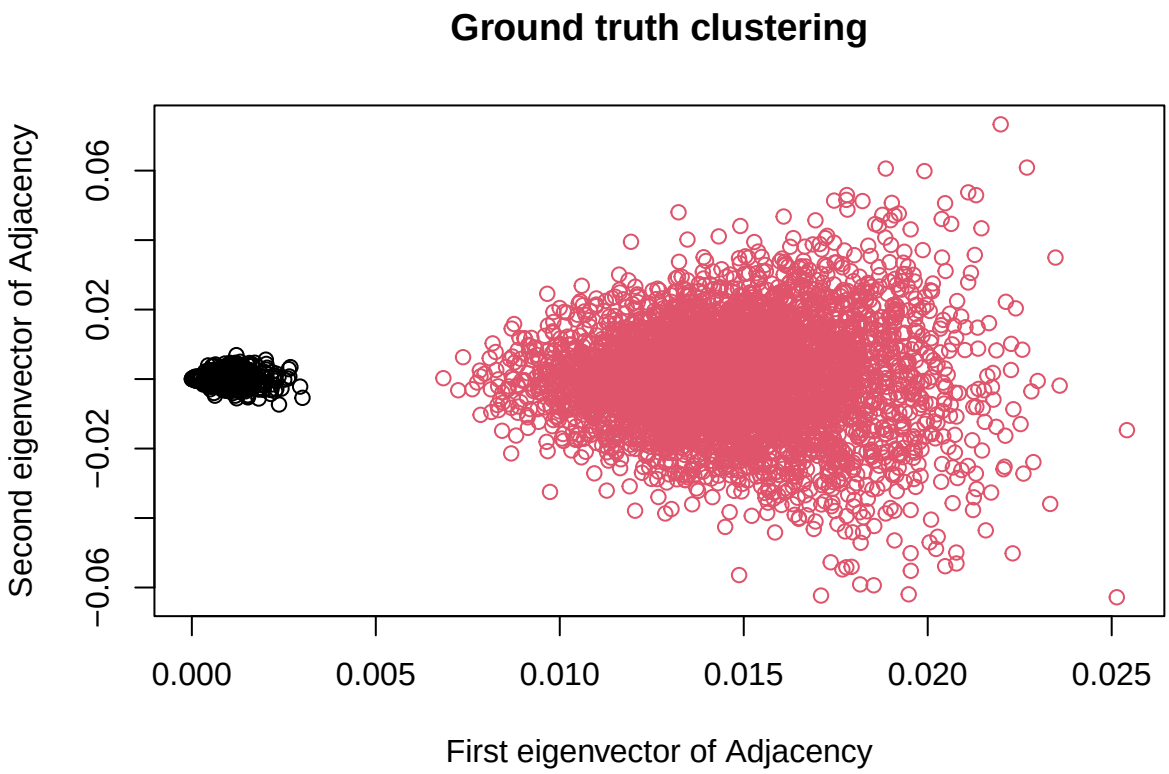

Spectral clustering
Second eigenvector of Adjacency
−0.1
0.0
0.1
0.2
0.00
0.05
0.10
0.15
First eigenvector of Adjacency

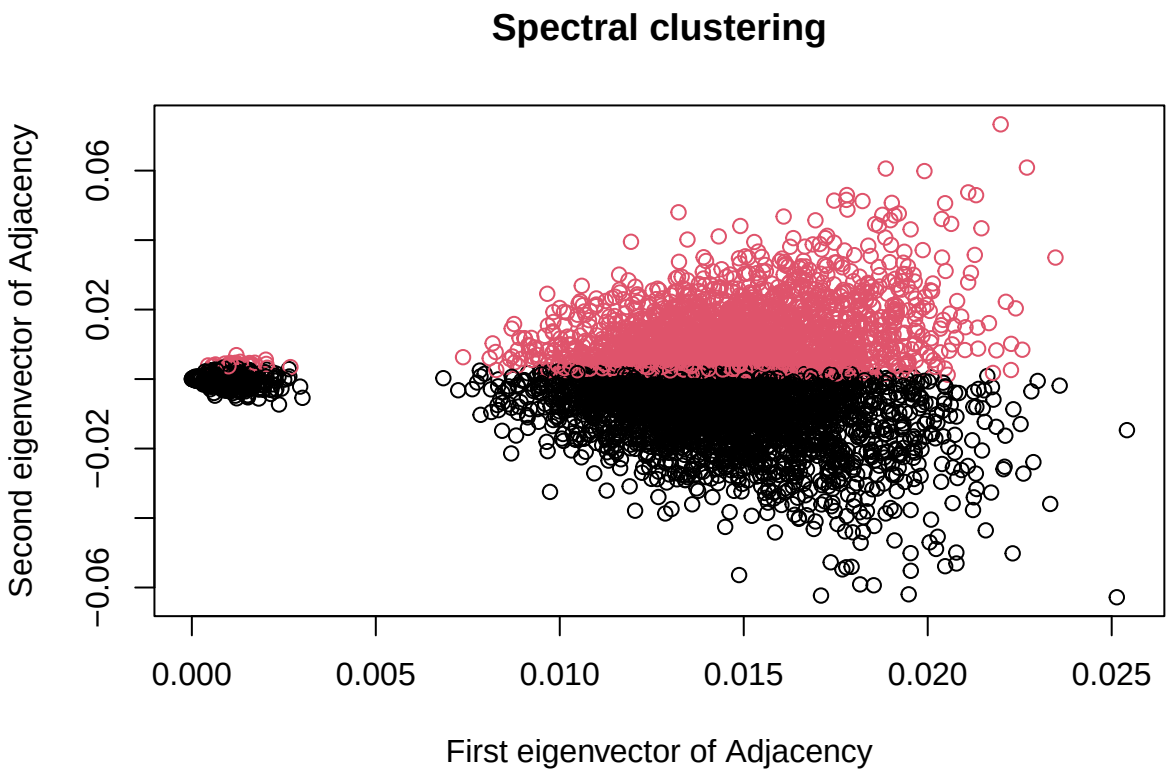


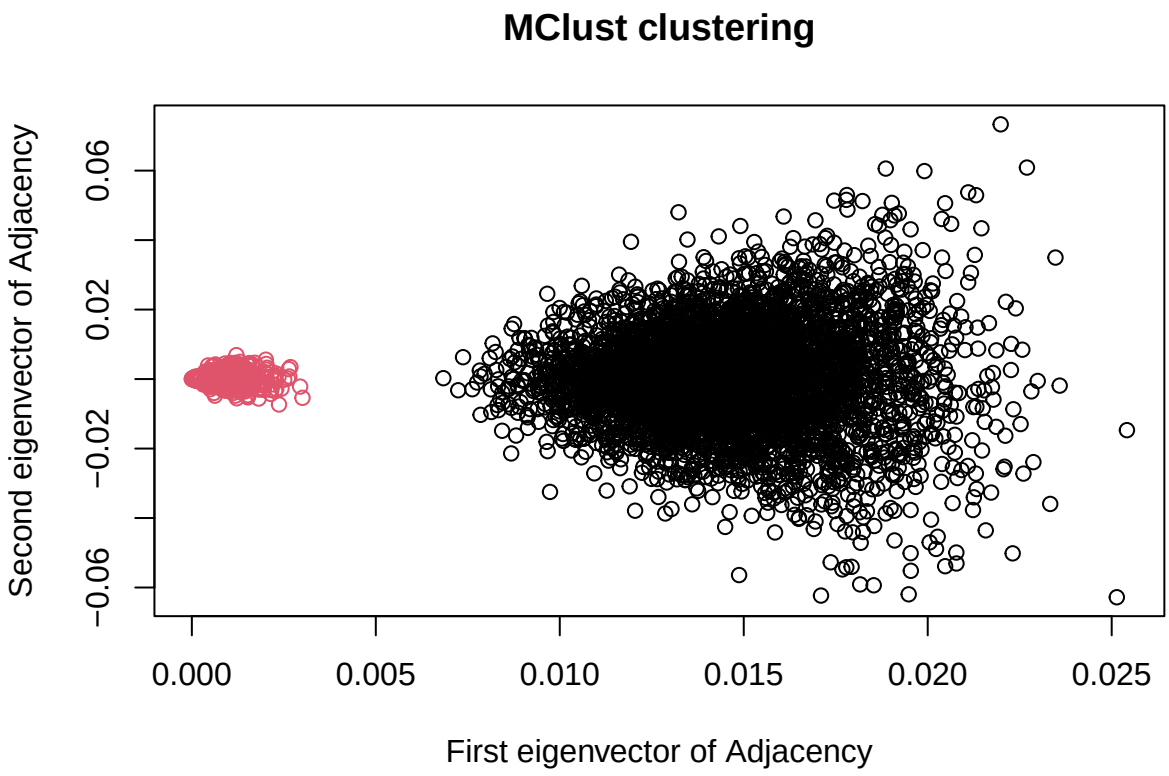

```
## [[1]]
##   ARI_k_means  ARI_mclust
## 1 0.005622734 0.007148986
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

SBM model 3

```
# SBM model 3
n <- 5000
rho <- log(n)/n
aa <- 4.5
bb <- 0.25
cc <- 4.5
block.sizes = c(1/10, 9/10)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```


Ground truth clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
0.000
0.005
0.010
0.015
0.020
0.025
First eigenvector of Adjacency

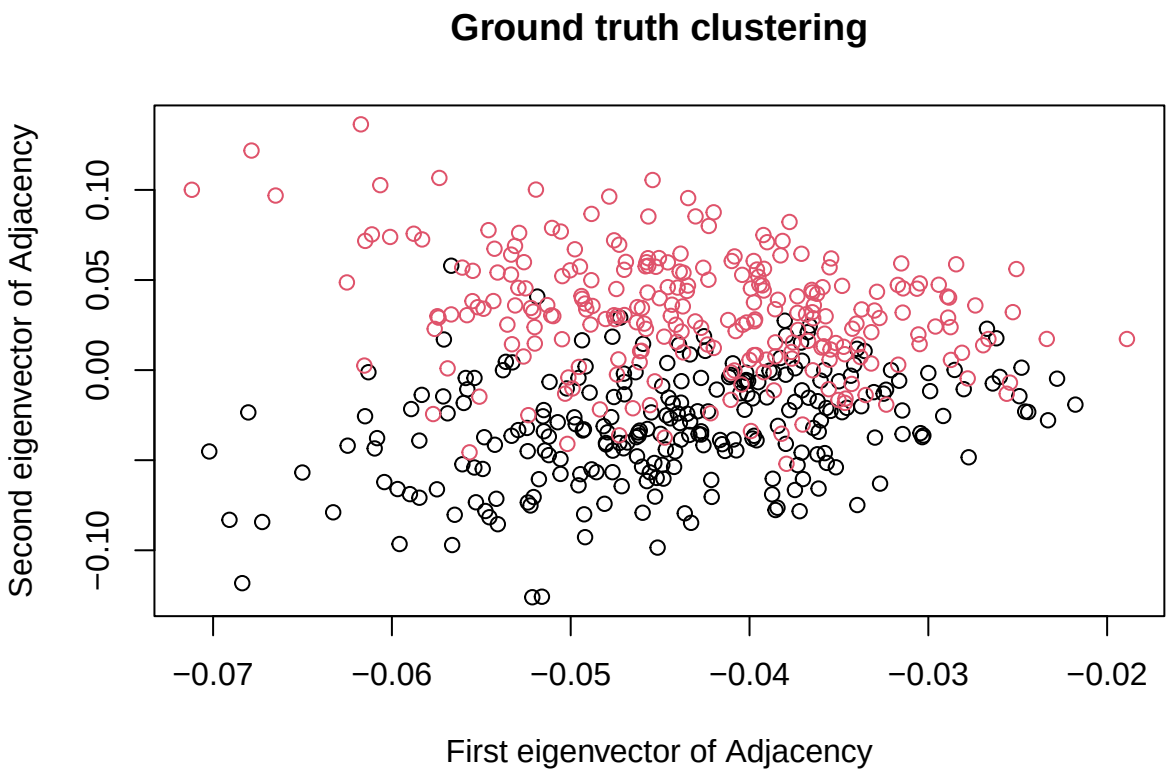

Spectral clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
0.000
0.005
0.010
0.015
0.020
0.025
First eigenvector of Adjacency

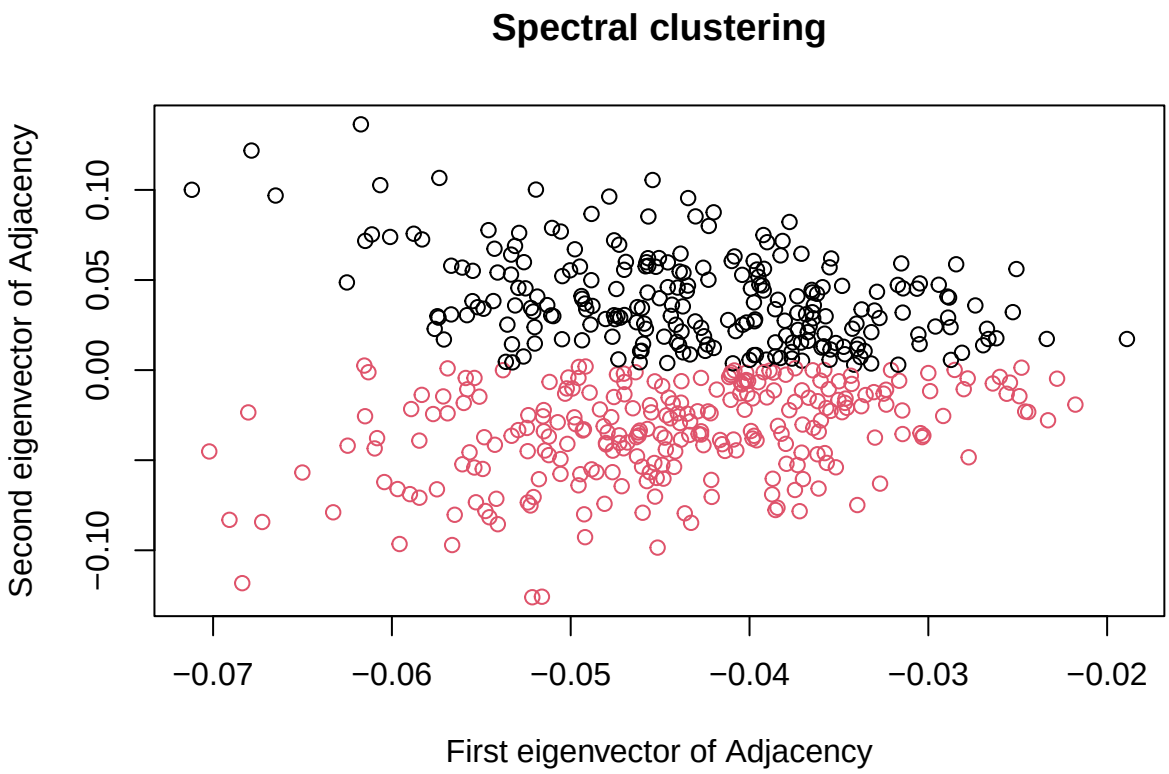


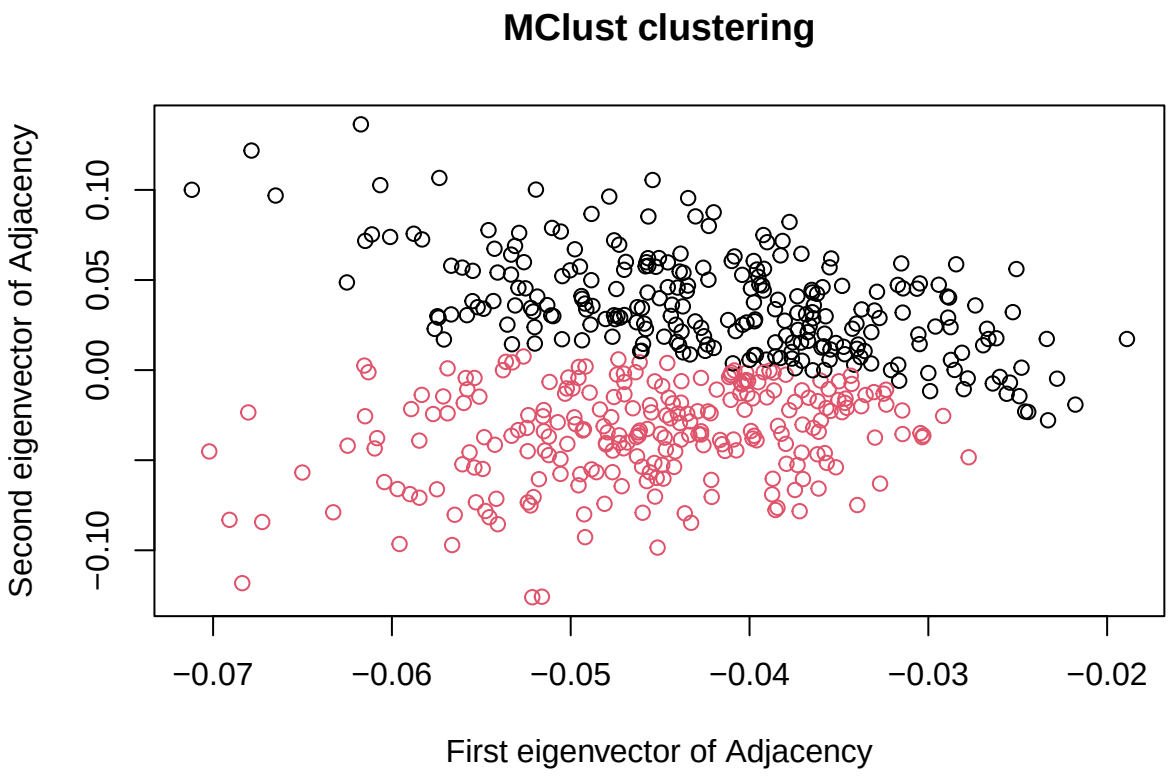

```
## [[1]]
##   ARI_k_means ARI_mclust
## 1 -0.02187893          1
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

SBM model 4

```
# SBM model 4
n <- 500
rho <- log(n)/n
aa <- 4.5
bb <- 2.5
cc <- 4.5
block.sizes = c(1/2, 1/2)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```


Ground truth clustering
Second eigenvector of Adjacency
0.10
0.05
0.00
−0.10
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
First eigenvector of Adjacency

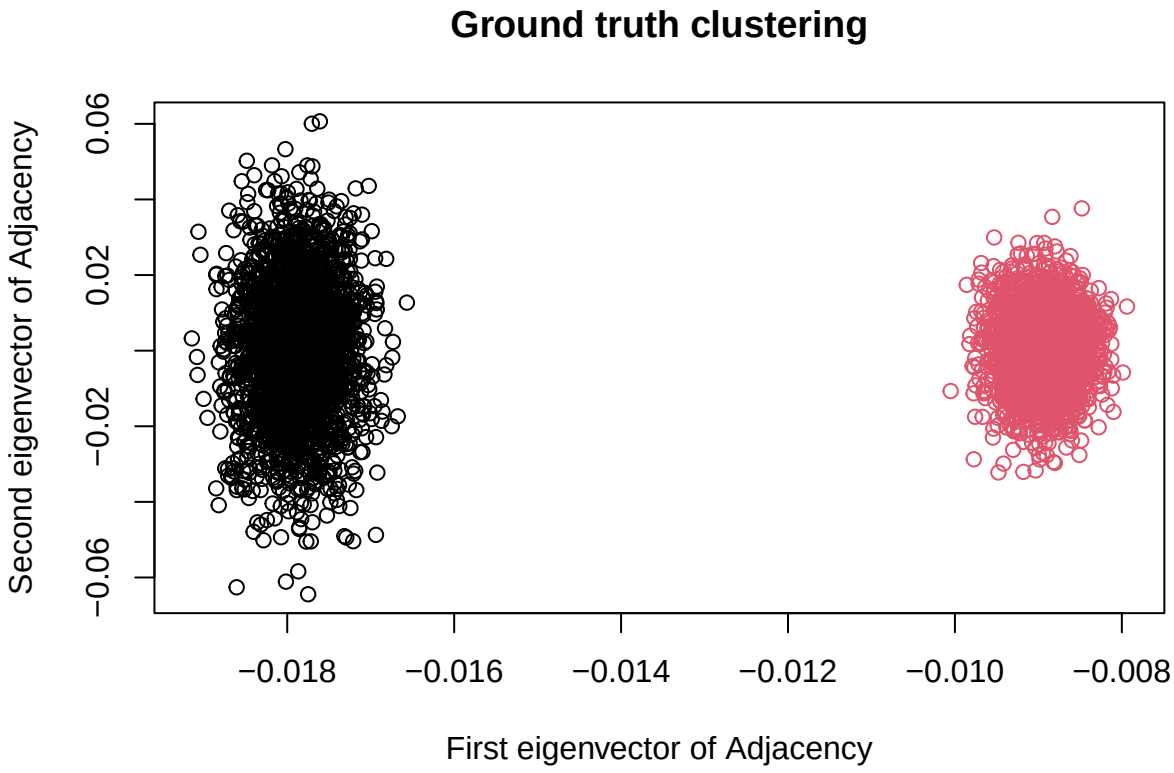

Spectral clustering
Second eigenvector of Adjacency
0.10
0.05
0.00
−0.10
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
First eigenvector of Adjacency

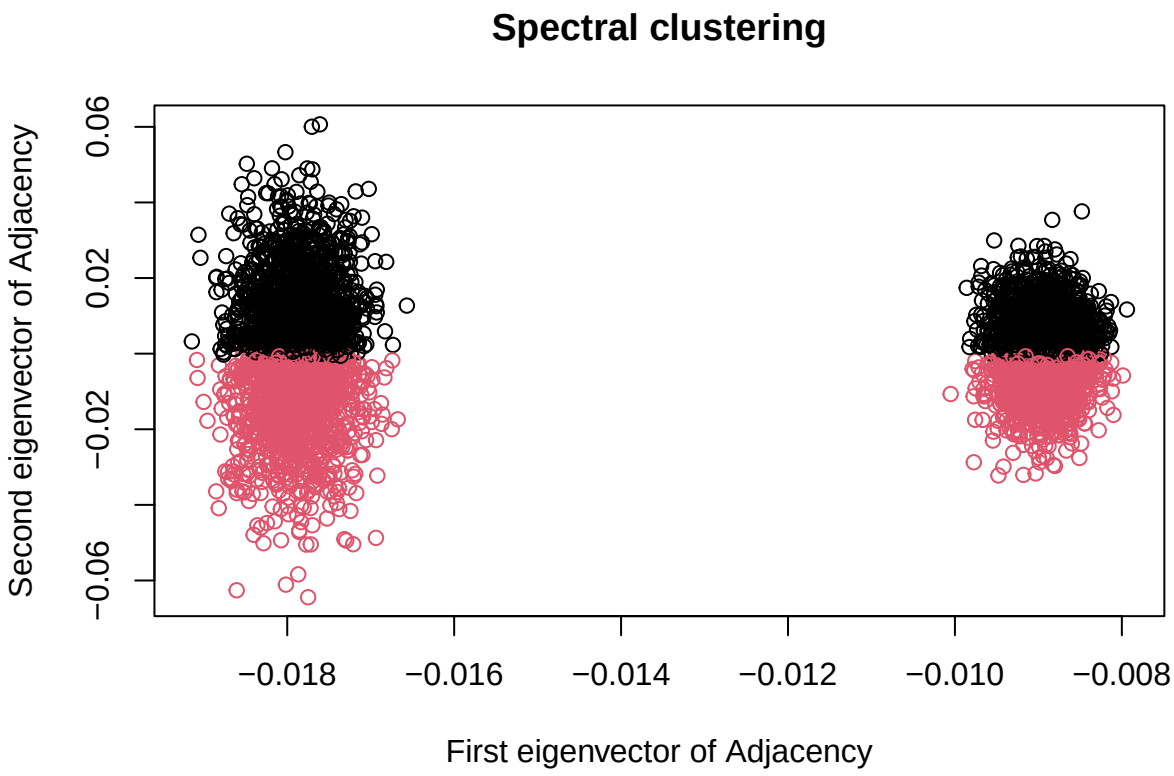


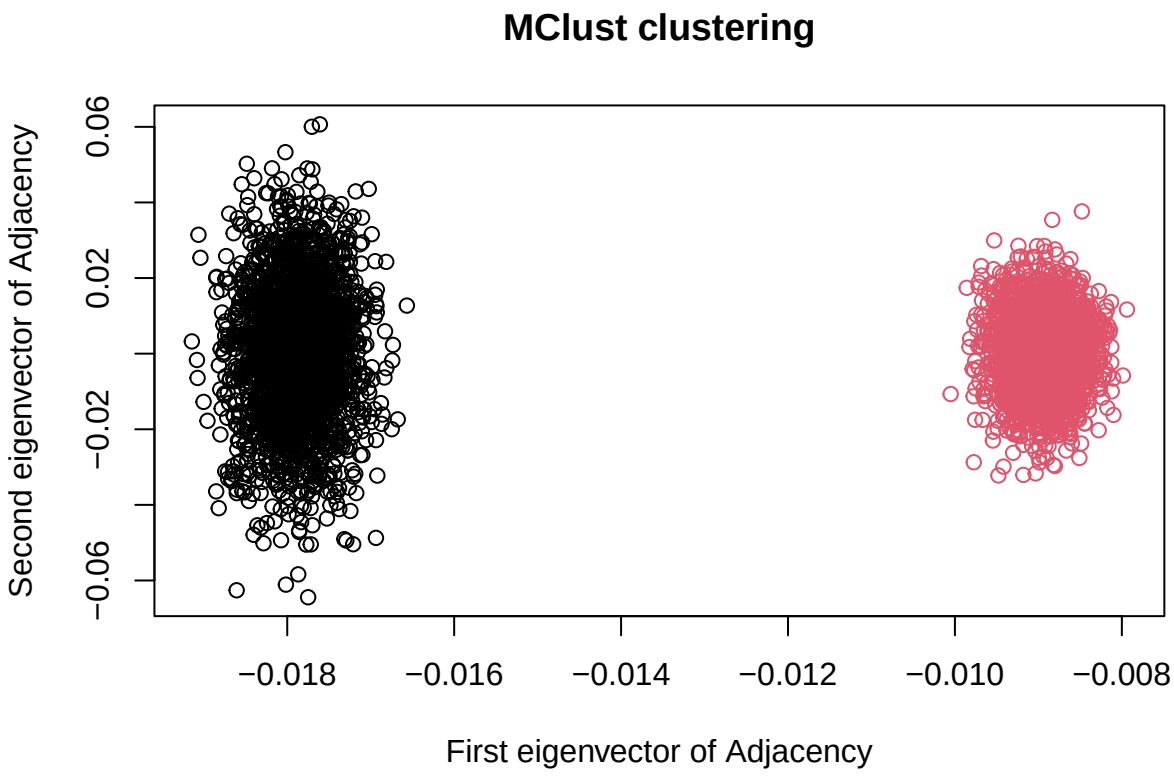

```
## [[1]]
##   ARI_k_means ARI_mclust
## 1   0.5348944  0.4613206
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

SBM model 5.A

```
# SBM model 5.A
n <- 5000
rho <- 1/10
aa <- 4
bb <- 2
cc <- 1
block.sizes = c(1/2, 1/2)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```


Ground truth clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
−0.018
−0.016
−0.014
−0.012
−0.010
−0.008
First eigenvector of Adjacency

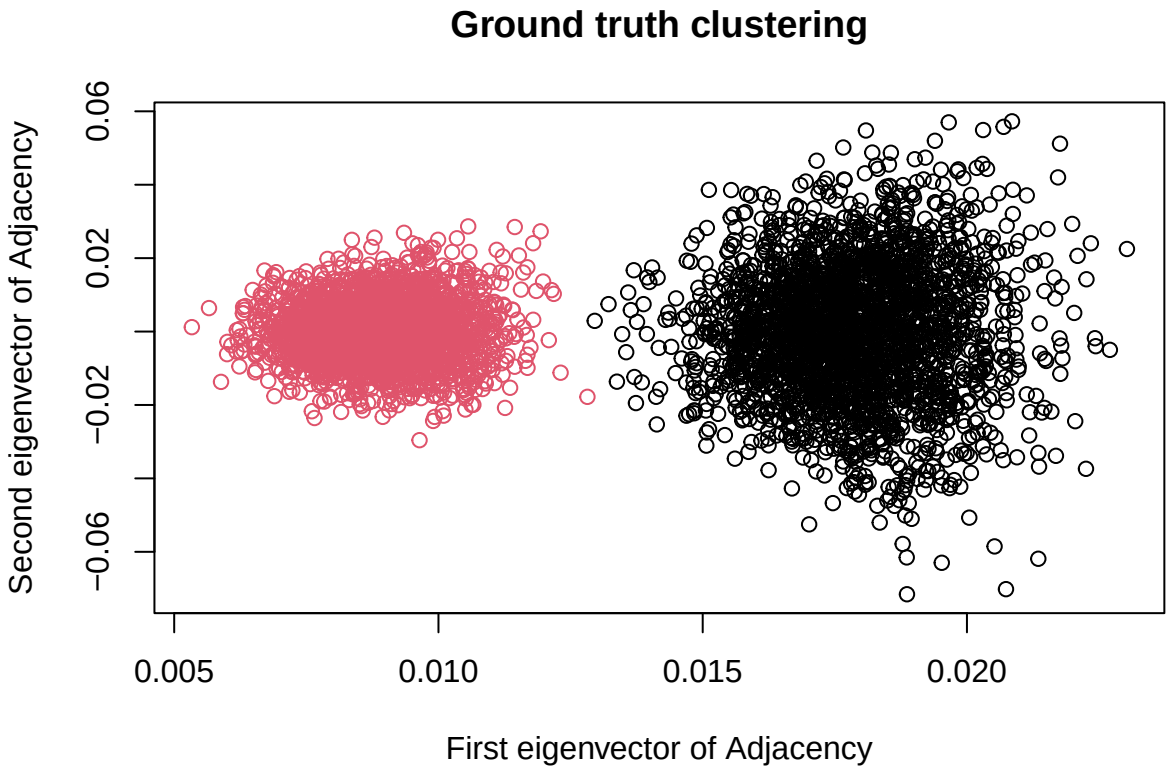

Spectral clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
−0.018
−0.016
−0.014
−0.012
−0.010
−0.008
First eigenvector of Adjacency



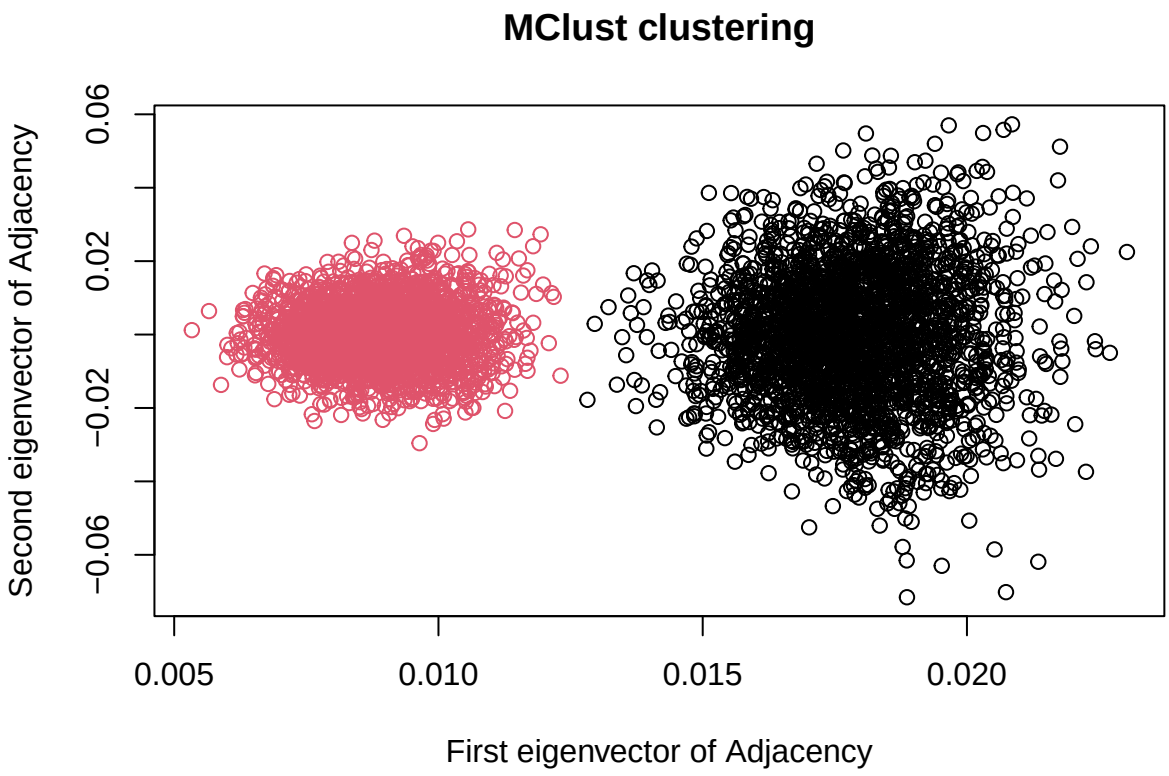

```
## [[1]]
##       ARI_k_means ARI_mclust
## 1 -0.000003556493          1
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

SBM model 5.B

```
# SBM model 5.B
n <- 5000
rho <- 1/100
aa <- 4
bb <- 2
cc <- 1
block.sizes = c(1/2, 1/2)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```


Ground truth clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
0.005
0.010
0.015
0.020
First eigenvector of Adjacency

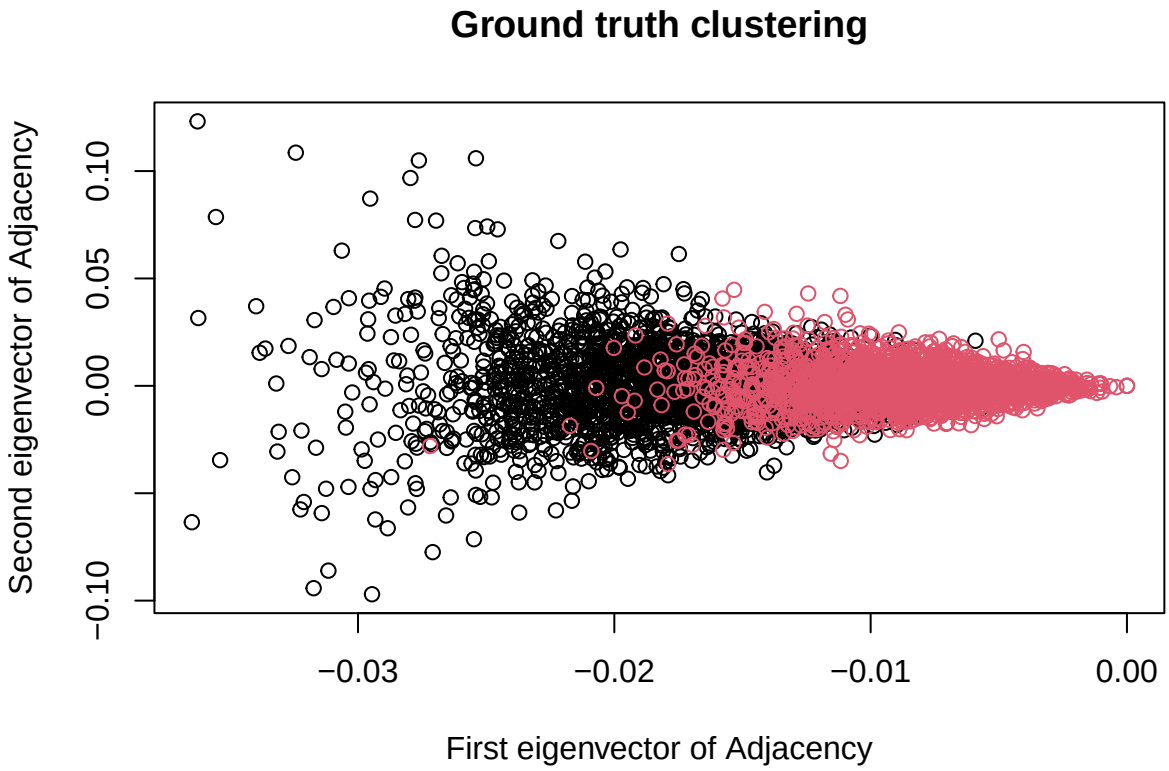

Spectral clustering
Second eigenvector of Adjacency
0.06
0.02
−0.02
−0.06
0.005
0.010
0.015
0.020
First eigenvector of Adjacency

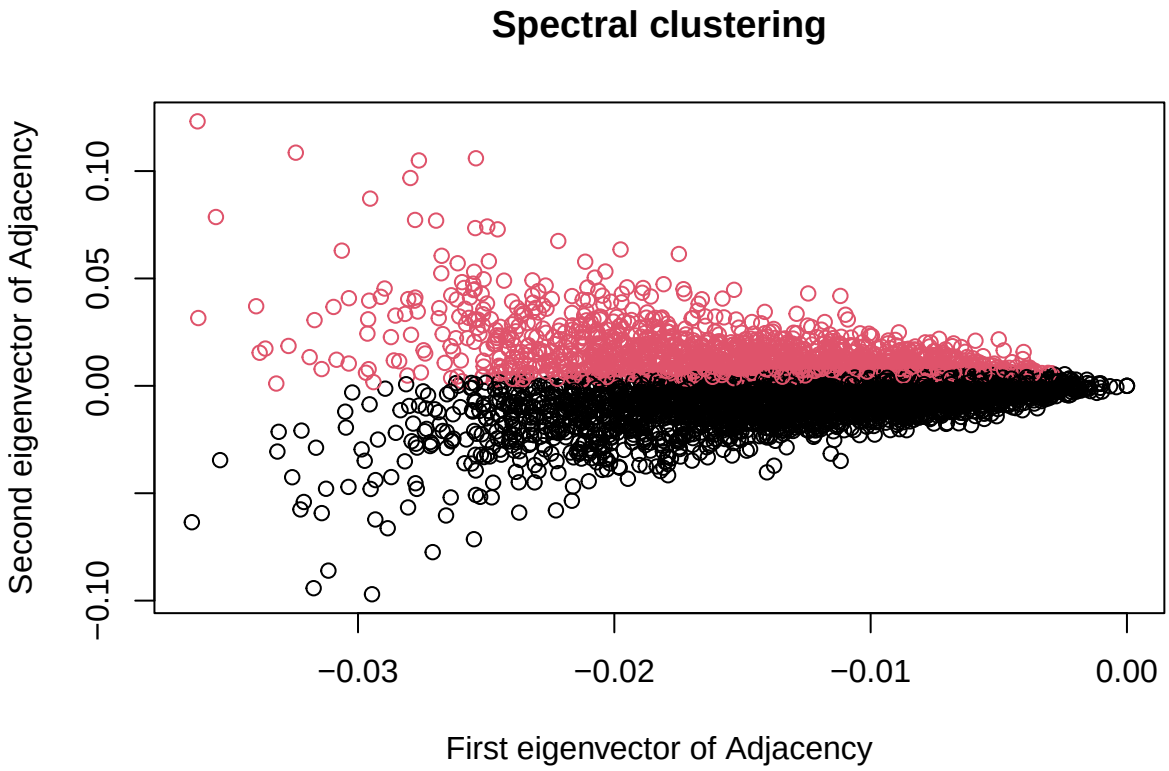


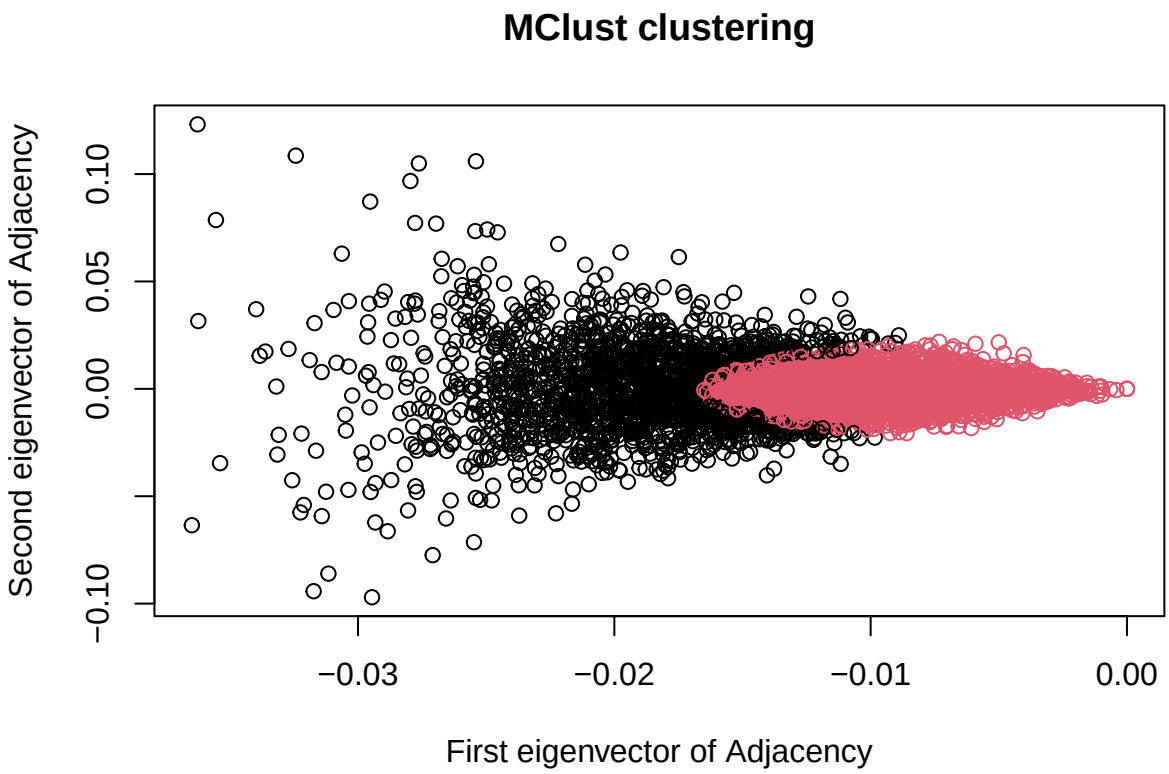

```
## [[1]]
##      ARI_k_means ARI_mclust
## 1 -0.00009183641     0.9992
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

SBM model 5.C

```
# SBM model 5.C
n <- 5000
rho <- 1/1000
aa <- 4
bb <- 2
cc <- 1
block.sizes = c(1/2, 1/2)*n

master_sbm(n, rho, aa, bb, cc, block.sizes)
```


Ground truth clustering
Second eigenvector of Adjacency
0.10
0.05
0.00
−0.10
−0.03
−0.02
−0.01
0.00
First eigenvector of Adjacency


Spectral clustering
Second eigenvector of Adjacency
0.10
0.05
0.00
−0.10
−0.03
−0.02
−0.01
0.00
First eigenvector of Adjacency




```
## [[1]]
##   ARI_k_means ARI_mclust
## 1  0.01988353   0.341871
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

Part II (spectral relaxations)

Section 5.4 of von Luxburg (2007) mentions that in general, "there is no guarantee whatsoever on the quality of the solution of the relaxed [min cut optimization] problem compared to the exact solution" [p.403]. Here we investigate further.

Consider the so-called cockroach graphs from Guattery and Miller (1998) as presented in von Luxburg (2007).

(1) Using R, numerically confirm the discussion of cut properties and eigenstructure as detailed on pages 403–404. In other words, generate and analyze large cockroach graphs.

(2) How do your findings change in the presence of perturbations (i.e., stochastic/deterministic modifications to edges and/or vertices)? You are free to specify different choices of perturbation mechanisms.

Bonus: Design *adversarial perturbations* that degrade the quality of cuts and spectral clusterings for cockroach graphs. Alternatively, demonstrate that cockroach graphs possess robustness to particular types of graph (matrix) perturbations.

Bonus: Investigate and explain the properties of spectral clustering in stochastic blockmodels with more than two communities.