hw2 - Network Analysis

Qian LI

4/23/2018

Intro

Two graph models will be generated for iris data to represent inner relationship inside the data. Specifically, nodes represent the 150 flower observations. I will use two graph models: association graph and gaussion graphical model.

Association model

First is the association model. We will first load the iris data, and then construct an unweighted graph where an edge represents whether or not the correlation between two observation is statistically significant at a 0.01 level.

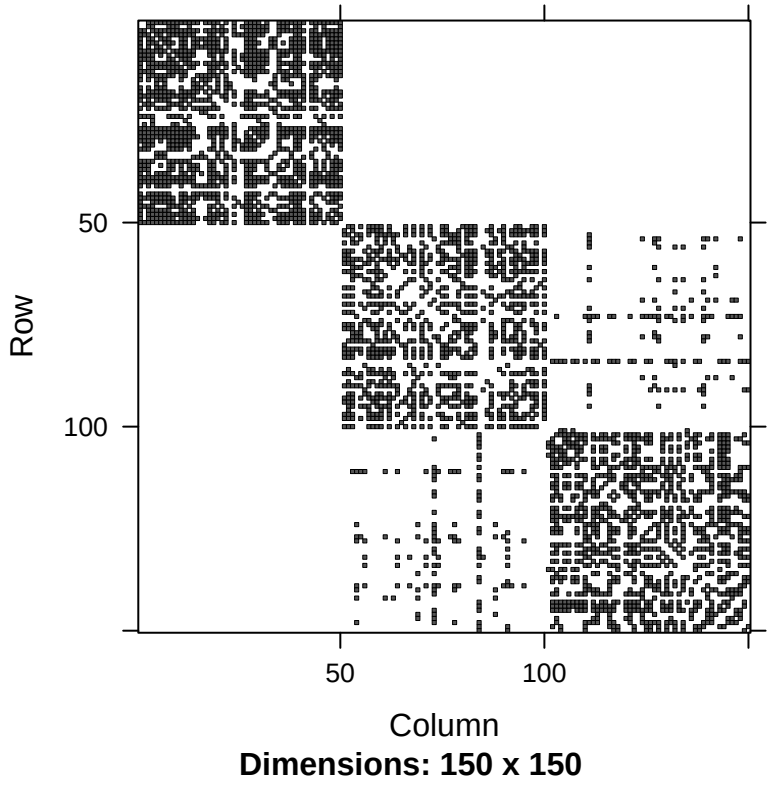
(a)

```
iris.t <- t(iris[, -5])

m <- dim(iris.t)[1]  # 4 features
n <- dim(iris.t)[2] # 150 observations

# correlation matrix
iris.cor <- cor(iris.t) # 150 * 150 matrix

# Fisher transform
z <- 0.5 * log((1 + iris.cor) / (1 - iris.cor))

# calculate p-values from Normal Distribution
z.vec <- z[upper.tri(z)]
corr.pvals <- 2 * pnorm(abs(z.vec), 0, sqrt(1 / (m-3)), lower.tail = FALSE)

# adjust p-values using Benjamin-Hochberg multiple
corr.pvals.adjusted <- p.adjust(corr.pvals, "BH")

# how many values are significant at 0.01 level
length(corr.pvals.adjusted[corr.pvals.adjusted < 0.01])
```

```
## [1] 1846
```

```
# create a network out of this
iris.adjacency <- matrix(0, n, n)
# first fill in the upper triangle
iris.adjacency[upper.tri(iris.adjacency)] <- corr.pvals.adjusted < 0.01
# now fill the lower triangle with the transpose of the upper
iris.adjacency <- iris.adjacency + t(iris.adjacency)
```

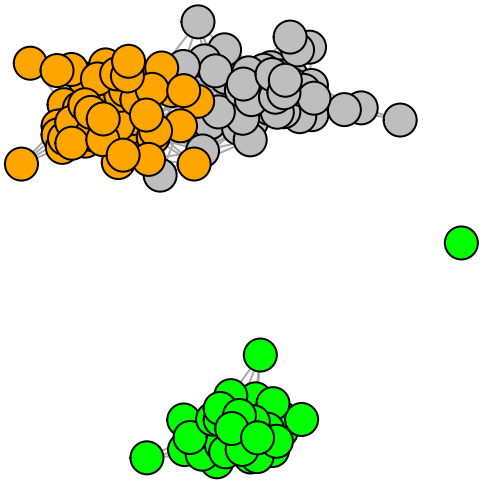
```
# image of the adjacency matrix
image(Matrix(iris.adjacency))
```



There are 1846 edges that are statistically significant at a 0.01 level, which is far less than 150*150. Rest of the edges will be removed. From the adjacency matrix image, we can see that the matrix is very sparse, as expected.

(b)

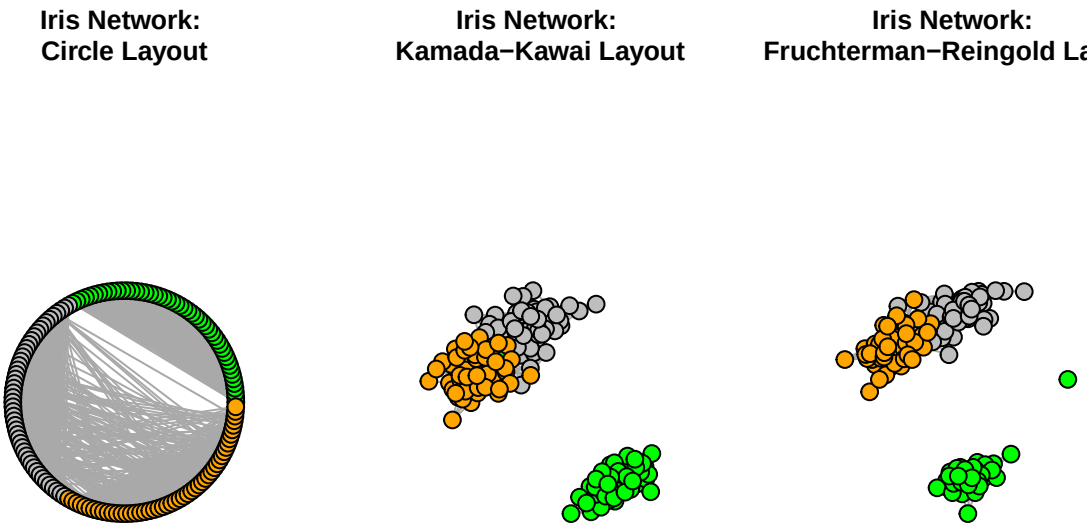
```
# convert the adjacency matrix to an igraph object
iris.g <- graph.adjacency(iris.adjacency, mode = "undirected")

# color the vertices based on species name
V(iris.g)[iris$Species == "setosa"]$color <- "green"
V(iris.g)[iris$Species == "versicolor"]$color <- "grey"
V(iris.g)[iris$Species == "virginica"]$color <- "orange"
igraph_options(vertex.label=NA)

plot(iris.g)
title("Iris Network - Association")
```



The plot of the Iris network is displayed, with vertices colored based on different species names. It's clear that different species are mostly grouped together.

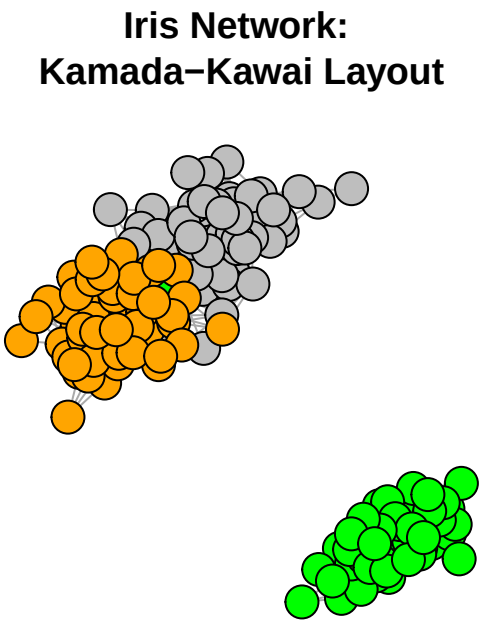
(c)

```
par(mfrow = c(1, 3))
igraph_options(vertex.label=NA)
plot(iris.g, layout=layout.circle)
title("Iris Network: \nCircle Layout")
plot(iris.g, layout=layout.kamada.kawai)
title("Iris Network: \nKamada-Kawai Layout")
plot(iris.g, layout=layout.fruchterman.reingold)
title("Iris Network: \nFruchterman-Reingold Layout")
```



Experimenting with different layouts, from personally opinion, circle does not convey inherent community structure well even though vertices of same color are together. Between Kamada-Kawai and Fruchterman-Reingold, there is one single vertice in Fruchterman-Reingold layout. So I prefer Kamada-Kawai.

```
par(mfrow = c(1, 1))
igraph_options(vertex.label=NA)
plot(iris.g, layout=layout.kamada.kawai)
title("Iris Network: \nKamada-Kawai Layout")
```



Here is network plot using Kamada-Kawai layout, which best represents the inherent community structure.

Gaussian graphical model

Next, we'll fit the Gaussian graphical model to the iris data.

(a)

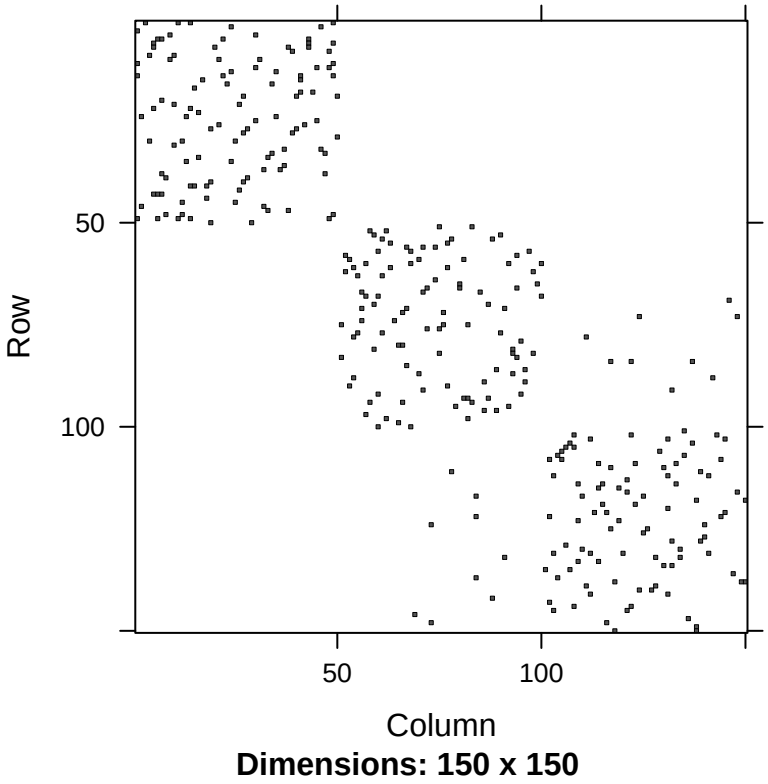
```
# create a huge object for analysis from the observed data matrix
# this is fitting the Gaussian graphical model to this data
huge.out <- huge(scale(iris.t))
```

```
## Conducting Meinshausen & Buhlmann graph estimation (mb)....done
```

```
# Now, we need to select which partial correlation values are
# statistical significant. There are two primary ways of doing this
# The first is known to be prone to under-selection. This is seen
# in the empty graph formed using the "ric" criterion below
huge.opt1 <- huge.select(huge.out, criterion = "ric")
```

```
## Conducting rotation information criterion (ric) selection....done
## Computing the optimal graph....done
```

```
# Plot matrix
image(huge.opt1$refit)
```



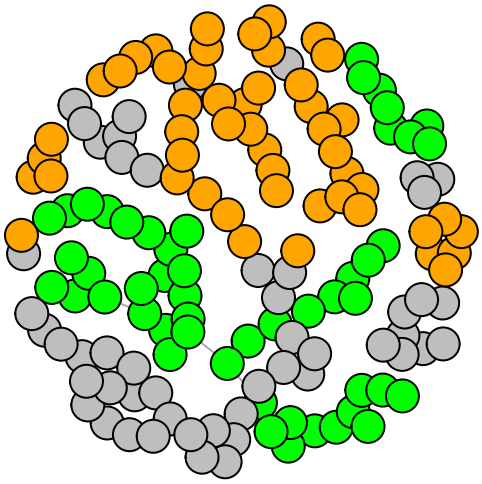
The matrix image seems more sparse than the previous one.

(b)

```
#Create a graph object and plot the result
igraph <- graph.adjacency(huge.opt1$refit, mode = "undirected")
summary(igraph) #150 nodes and 162 edges
```

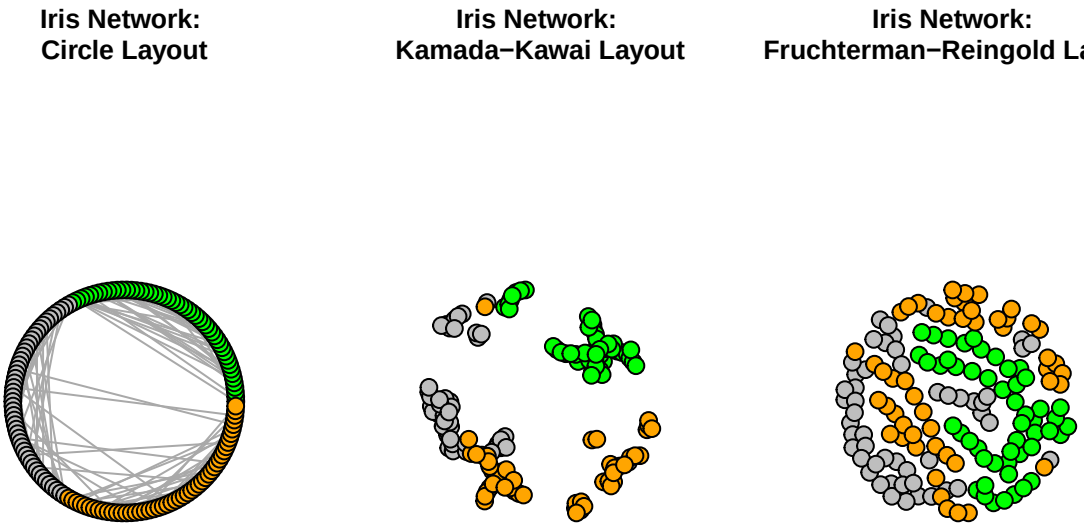
```
## IGRAPH 9f978f7 U--- 150 162 --
```

```
# color the vertices based on species name
V(igraph)[iris$Species == "setosa"]$color <- "green"
V(igraph)[iris$Species == "versicolor"]$color <- "grey"
V(igraph)[iris$Species == "virginica"]$color <- "orange"

plot(igraph)
title("Iris Network - Gaussion")
```



(c)

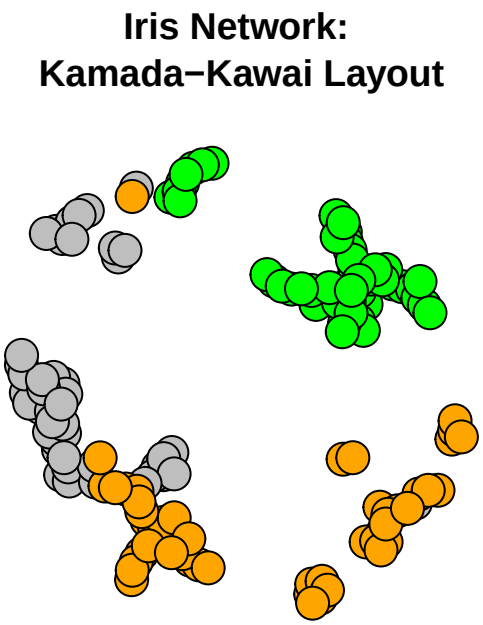
```
par(mfrow = c(1, 3))
igraph.options(vertex.label=NA)
plot(igraph, layout = layout.circle)
title("Iris Network: \nCircle Layout")
plot(igraph, layout = layout.kamada.kawai)
title("Iris Network: \nKamada-Kawai Layout")
plot(igraph, layout = layout.fruchterman.reingold)
title("Iris Network: \nFruchterman-Reingold Layout")
```



Experimenting with different layouts, I still don't like the circle layout. Kamada-Kawai layout seperates vertices of different color better. So I think Kamada-Kawai layout will best represents the inherent community structure.

```
par(mfrow = c(1, 1))
igraph.options(vertex.label=NA)
```

```r
plot(igraph, layout = layout.kamada.kawai)
title("Iris Network: \nKamada-Kawai Layout")
```



(d) Discussion of the two models

From the image of matrices generated by the two models,

- Similarities: Both matrics are symmetric.
- Differences: Gaussion graphical model generates a more sparse model than Association graph.

From the plot of iris network (Kamada-Kawai layout),

- Similarities: Vertices of different color are mostly separate.
- Differences: Association graph groups vertices of same color better. In addition, the two models display the inherent structure somewhat different. In more details, Association graph separates green vertices from the rest while orange vertices and grey vertices touch together on several vertices. Gaussion graphical model, however, they are all not clearly separated from the rest.