**Batch 10 – Data Mining 1 – Unsupervised Learning – Individual Assignment 2**

1. **Please include (copy-paste) the output ─ first fifteen rules along with header and input parameter details ─ in your submission.**

df <- read.csv("Cosmetics.csv")

df.mat <- as.matrix(df[,-1])

trs <- as(df.mat,"transactions")

transactions in sparse format with

1000 transactions (rows) and

14 items (columns)

inspect(trs)

rules = apriori(trs, parameter=list(support=0.18, confidence=0.50,target = "rules"))

rules\_high\_lift <- (head(sort(rules, by="lift"),n=15))

> inspect(head(sort(rules, by="lift"),n=15))

lhs rhs support confidence lift count

[1] {Mascara} => {Eye.shadow} 0.321 0.8991597 2.359999 321

[2] {Eye.shadow} => {Mascara} 0.321 0.8425197 2.359999 321

[3] {Concealer} => {Eyeliner} 0.297 0.6719457 1.470341 297

[4] {Eyeliner} => {Concealer} 0.297 0.6498906 1.470341 297

[5] {Blush} => {Mascara} 0.184 0.5068871 1.419852 184

[6] {Mascara} => {Blush} 0.184 0.5154062 1.419852 184

[7] {Blush} => {Concealer} 0.220 0.6060606 1.371178 220

[8] {Lip.Gloss} => {Foundation} 0.356 0.7265306 1.355468 356

[9] {Foundation} => {Lip.Gloss} 0.356 0.6641791 1.355468 356

[10] {Blush} => {Eye.shadow} 0.182 0.5013774 1.315951 182

[11] {Mascara} => {Concealer} 0.204 0.5714286 1.292825 204

[12] {Eye.shadow} => {Concealer} 0.201 0.5275591 1.193573 201

[13] {Eye.shadow} => {Lip.Gloss} 0.201 0.5275591 1.076651 201

[14] {Mascara} => {Lip.Gloss} 0.181 0.5070028 1.034700 181

[15] {Eye.shadow} => {Foundation} 0.211 0.5538058 1.033220 211

> apriori(trs, parameter=list(support=0.18, confidence=0.50,target = "rules"))

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target

0.5 0.1 1 none FALSE TRUE 5 0.18 1 10 rules

ext

FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 180

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[14 item(s), 1000 transaction(s)] done [0.00s].

sorting and recoding items ... [11 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 done [0.10s].

writing ... [20 rule(s)] done [0.00s].

creating S4 object ... done [0.00s].

set of 20 rules

**b) What is the support of the first rule? Explain how it has been calculated for this rule.**

The support is 0.321. It is calculated by this formula = Probability(Mascara and Eyeshadow)

It means that the number of itemsets containing Mascara and Eyeshadow are 321 out of total 1000 transactions. -> 321/1000 = 0.321

**c) What is the confidence of the first rule? Explain how it has been calculated for this rule.**

The confidence is 89.9% . It is calculated using formula = P(Mascara and Eyeshadow)/ P(Mascara). It means that if Mascara is purchased with 89% confidence , Eyeshadow will also be purchased. It is a conditional probability that a transaction selected randomly will include all items in consequent given that transaction includes all items in antecedent.

**d) What is the lift ratio of the first rule? Explain how it has been calculated**.

Lift ratio is 2.35 of first rule. Strength of this rule is high. Level of association is higher than would be expected if they were independent. Larger the Lift ratio Greater is the strength of association

Lift ratio = confidence/ benchmark confidence

**e) Reviewing the first fifteen rules, comment on their *redundancy* (read as constructed from same item set/tuple). How many distinct rules did you find from the first 15 rules?**

There are 11 distinct rules. Redundancy among the rules is because of the same item sets in transactions.

**f) Interpret the first three distinct (i.e., excluding the redundant ones, if any, as defined above) rules, in the output, in words.**

Eyeshadow - a coloured cosmetic applied to the eyelids or to the skin around the eyes to accentuate them.

Mascara - a cosmetic for darkening and thickening the eyelashes.

Blush - **cosmetic** typically used to redden the cheeks so as to provide a more youthful appearance, and to emphasize the cheekbones.

Concealer - a flesh-toned cosmetic stick used to cover spots, blemishes, and dark under-eye circles.

**g) Based on the distinct rules that you identified in Part (f), suggest some action that’ll benefit the business owner.**

The products are placed together because of frequent pattern analysis. It is selling strategies to sell more things together. The products like mascara and eyeshadow can be bought from same manufacturer and can be placed together for frequent buying.

**Question 2**

**Step 2:** Create a directed network graph of airline routes using *routes.dat*

**Step 3: Create community-based clusters. If you are using R, choose leading.eigenvector approach. If you are using any other software, use the equivalent algorithm provided with the software.**

**Step 4: Answer the following questions.**

1. **What would you call a *community* in a social-media network? Intuitive, qualitative answers are expected/acceptable.**

Community is a group of people having similar interests like same geographical location, same lifestyle. In social media network, a community can be of particular political party or football team, ipl team. People in a community might not have any direct relationship.

1. **Extend the definition of community (as you suggested above in Part a) to the community of airports.**

It can be that group of airports within which the connections are dense but connections between different communities is sparser. Airports which are international hubs, aiports which are less busy, airports which are at higher altitudes.

1. **How many distinct airports are there in the dataset? How many *communities* of airports got identified? List the number of airports in each cluster/community in a table.**

3425 are the distinct number of airports. There are 25 communities identified. The number of airports are given below:

> sizes(A)

Community sizes

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

137 2 4 2 4 2 4 10 635 340 776 334 274 1 6 1 2 21 198 656 1 1

23 24 25

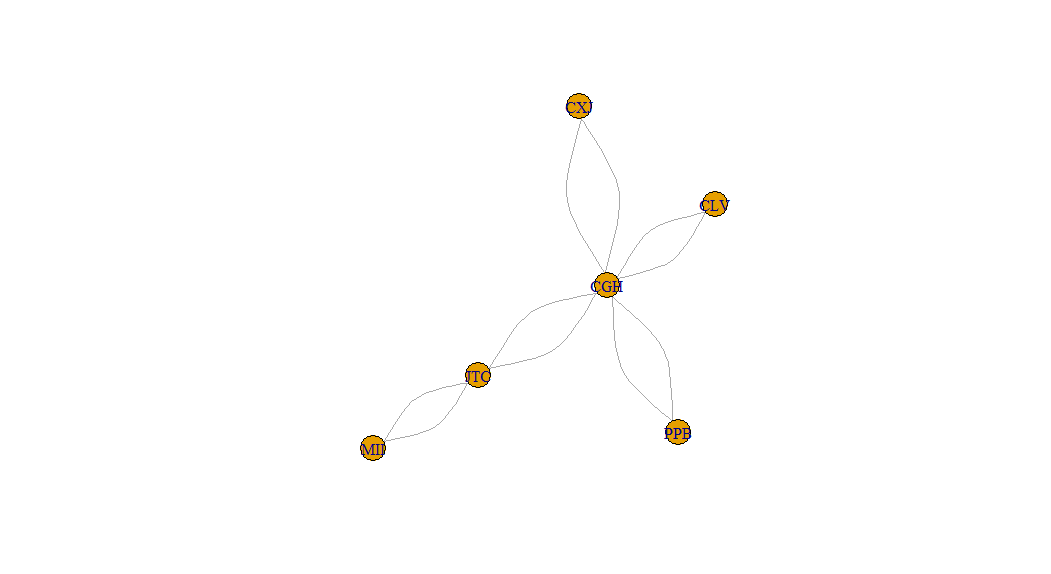
1 1 12

1. **Interpret/characterize, at least, five of the communities you identified above.**

First Community

sub1 <- subgraph(AirlineNW, which(membership(A) == 15))

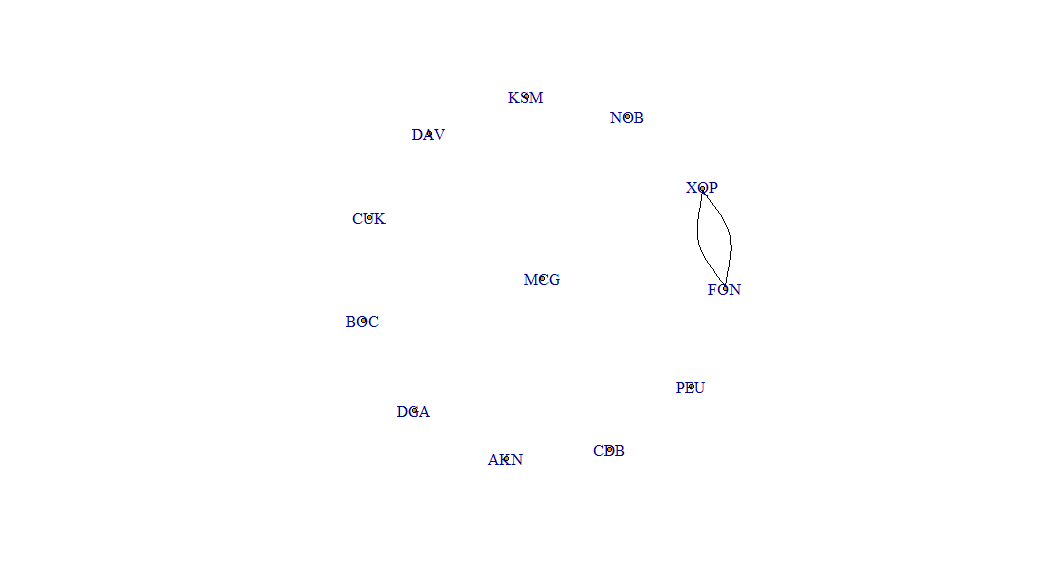
Interpret – CGH airport is most connected airport and has flights in both directions.



Second Community

sub2 <- subgraph(AirlineNW, which(membership(A) == 25))

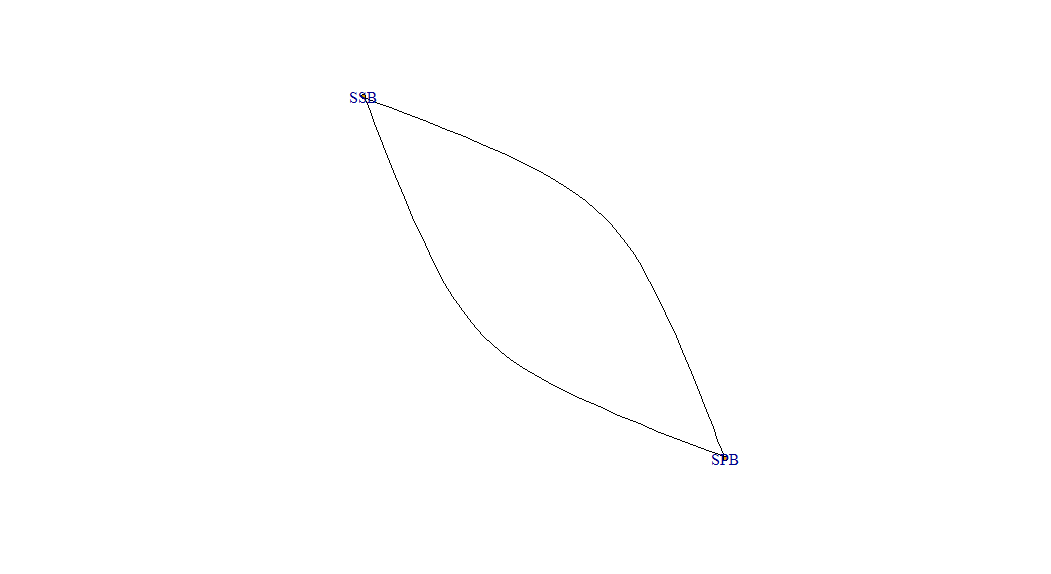
Interpret – The airports are not connected with each other and this community will be of those airports which are distant and least populated.



Third community

sub2 <- subgraph(AirlineNW, which(membership(A) == 2))

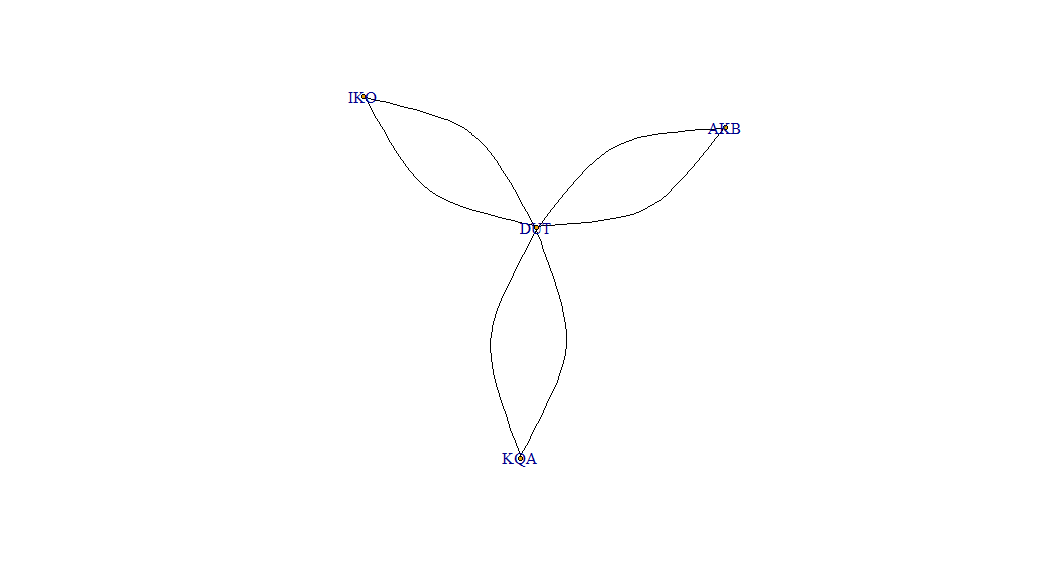
These two airports have routes from and to and these airports might have many flights in between.



Fourth Community

sub2 <- subgraph(AirlineNW, which(membership(A) == 3))

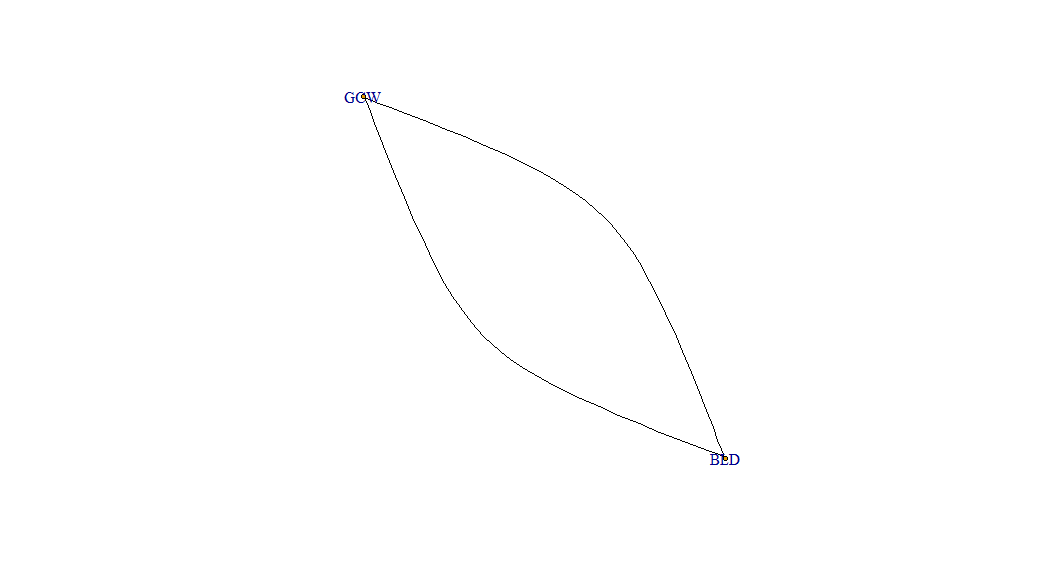
Interpret - DUT airport is central hub to other airports in community.



Fifth Community

sub2 <- subgraph(AirlineNW, which(membership(A) == 4))

These two airports have routes from and to and these airports might have many flights in between.



**Step 5: Compute the centralities (in-degree, out-degree, in-closeness, eigenvector, betweenness) of each airport. Now, run k-Means clustering to group the airports based on their centralities alone. Take k equal to the number of communities you obtained in Part c, above.**

**(e) Do you observe the groups obtained in Step 4 to be similar to or different from what you obtained in Step 5? Why?**

Interpret – Groups in clusters are different because of the number of airports in each cluster. It is because the algorithm difference in K-means and leading eigen vector approach.

> head(centralities)

inDegree outDegree closenessIn betweenness

AER 52 52 0.03172838 2.087755e-05

KZN 56 56 0.03175163 2.209804e-05

ASF 16 16 0.03159721 3.983581e-07

MRV 44 44 0.03186213 2.295775e-05

CEK 40 40 0.03162055 2.393366e-06

OVB 177 177 0.03212098 1.189939e-03

eigenv <- eigen\_centrality(AirlineNW, directed = TRUE, scale = FALSE, weights = NULL)

> eigenv

$vector

AER KZN ASF MRV CEK OVB

2.587475e-03 2.603360e-03 1.008290e-03 2.982193e-03 1.856655e-03 9.060106e-03

DME NBC TGK UUA EGO KGD

3.736878e-02 1.198811e-03 2.115196e-04 3.909785e-04 9.496344e-04 2.424661e-03

fit <- kmeans(normalized\_data, centers=25, iter.max=10, nstart=4)

Warning messages:

1: did not converge in 10 iterations

2: did not converge in 10 iterations

3: did not converge in 10 iterations

> ## centers: either the number of clusters, or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in x is chosen as the initial centres.

> ## iter.max: the maxi1mum number of iterations allowed.

> ## nstart: if centers is a number, how many random sets should be chosen.

> fit

K-means clustering with 25 clusters of sizes 37, 17, 236, 4, 717, 99, 61, 7, 378, 24, 28, 41, 119, 823, 17, 56, 15, 179, 80, 14, 45, 24, 25, 362, 17

**Step 6:** Now, run k-Means clustering again on the airports based on their centralities. **Go with a value of k as you find appropriate.**

1. Interpret the clustering outcome.

**Interpret** – I find the value of K=2 as appropriate because the centralities in first cluster are positive and in the second cluster are negative.

> fit <- kmeans(normalized\_data, centers=2, iter.max=10, nstart=4)

> ## centers: either the number of clusters, or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in x is chosen as the initial centres.

> ## iter.max: the maxi1mum number of iterations allowed.

> ## nstart: if centers is a number, how many random sets should be chosen.

> fit

K-means clustering with 2 clusters of sizes 97, 3328

Cluster means:

inDegree outDegree closenessIn betweenness eigen

1 4.7834106 4.7834106 0.50066237 3.9212441 4.5751376

2 -0.1394203 -0.1394203 -0.01459262 -0.1142911 -0.1333499

**Step 6:** Carefully observe the centralities of the airports in the dataset.

(g) **If your organization is planning on launching a new flight service on a couple of new routes, what will that be (based on the information you have in this data alone)? Explain your answer. What other information would have helped you to make a better decision?**

Interpretation 🡪 The new flight services can be from