# Topic: Network Analytics

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Anandakrishnan k v Batch ID:** 19042021

**Topic: Network Analytics**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**

There are two datasets consisting of information for the connecting routes and flight halt. Create network analytics models on both the datasets separately and measure degree centrality, degree of closeness centrality, and degree of in-between centrality.

* 1. **What is the business objective?**

Create network analytics models on datasets and measure degree centrality, degree of closeness centrality, and degree of in-between centrality

* 1. **Are there any constraints?**

**Maximize:** Accuracy of the result

**Minimize:** the process time

**R code:-**

### Airport #######

#install.packages("igraph")

library("igraph")

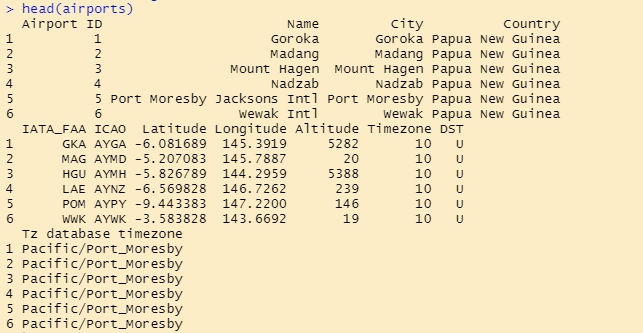
##### Example: Airlines Network using Edge List #####

## Data source: http://openflights.org/data.html

airports <- read.csv("G:/network/Datasets\_Network Analytics/flight\_hault.csv", header = FALSE) # header = false ==> Not cosidering first row as name of the column

colnames(airports) <- c("Airport ID","Name","City","Country","IATA\_FAA","ICAO","Latitude","Longitude","Altitude","Timezone","DST","Tz database timezone")

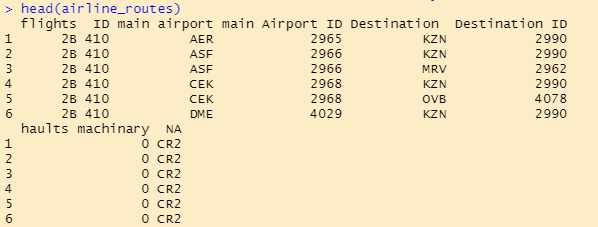
head(airports)



airline\_routes <- read.csv("G:/network/Datasets\_Network Analytics/connecting\_routes.csv", header=FALSE)

colnames(airline\_routes) <- c("flights", "ID", "main airport","main Airport ID","Destination ","Destination ID","haults","machinary")

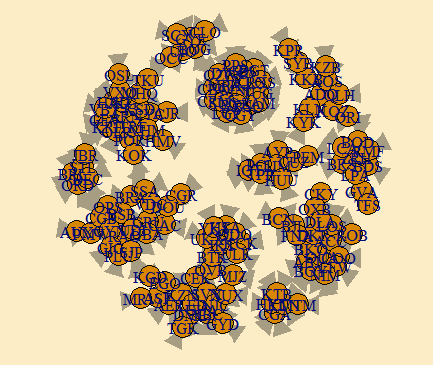
head(airline\_routes)



# taken only first 300 from the list

AirlineNW <- graph.edgelist(as.matrix(airline\_routes[1:300 , c(3, 5)]), directed = TRUE) ## considering only first 300 rows and edges taken as directed : nodes ==> source airport & destination airport column datas

plot(AirlineNW)



## How many airports are there in the network?

?vcount

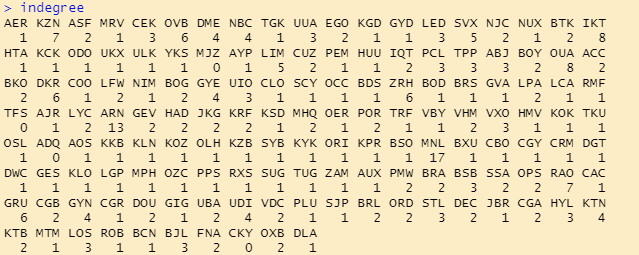
vcount(AirlineNW) #143

# Which airport has most flights coming in, and how many?

?degree

indegree <- degree(AirlineNW, mode = "in")

indegree



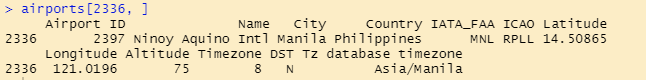
max(indegree)

index <- which(indegree == max(indegree))

indegree[index]

which(airports$IATA\_FAA == "MNL") # 2336

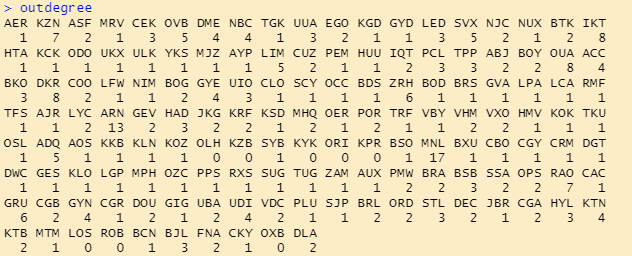
airports[2336, ]



# Which airport has most flights going out of, and how many?

outdegree <- degree(AirlineNW, mode = "out")

outdegree



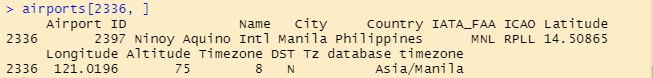
max(outdegree)

index <- which(outdegree == max(outdegree))

outdegree[index]

which(airports$IATA\_FAA == "MNL") # 2336

airports[2336, ]

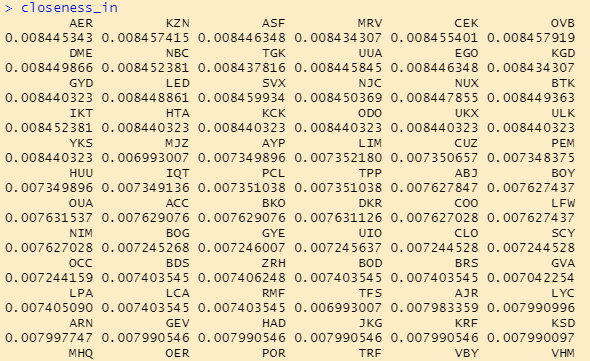


# Which airport is close to most of the airports (in terms of number of flights in)

?closeness

closeness\_in <- closeness(AirlineNW, mode = "in", normalized = TRUE)

closeness\_in



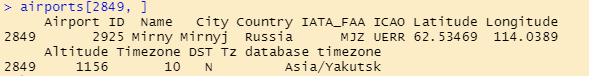
max(closeness\_in)

index <- which(closeness\_in == max(closeness\_in))

closeness\_in[index]

which(airports$IATA\_FAA == "SVX") # 2849

airports[2896, ]

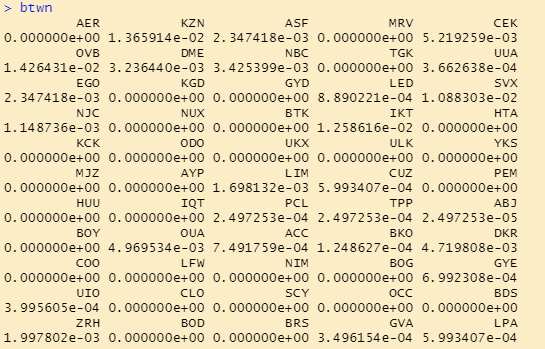


# Which airport comes in between most of the routes and hence is an important international hub?

?betweenness

btwn <- betweenness(AirlineNW, normalized = TRUE)

btwn



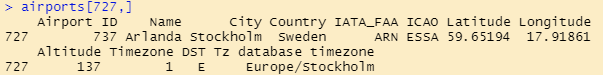
max(btwn)

index <- which(btwn == max(btwn))

btwn[index]

which(airports$IATA\_FAA == "ARN") # 727

airports[727,]

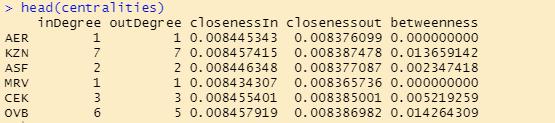


# Degree, closeness-in, closeness\_out, and betweenness centralities together

centralities <- cbind(indegree, outdegree, closeness\_in,closeness\_out, btwn)

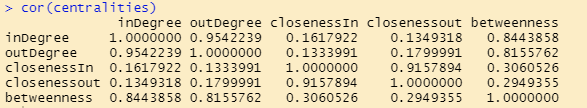
colnames(centralities) <- c("inDegree","outDegree","closenessIn","closenessout","betweenness")

head(centralities)



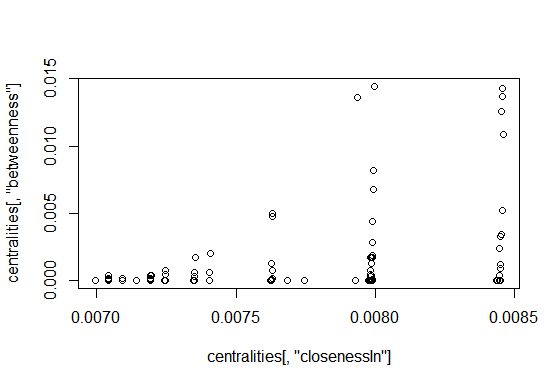
# correlations of the centralities

cor(centralities)



# Any pair with low correlation?

plot(centralities[, "closenessIn"], centralities[, "betweenness"])



**Summary:-**

From the correlation result between the columns of centralities data

* Indegree & outdegree have the high correlation 🡺 whatever the flight get in the airport is going out.
* ClosenessIn & closenessout have high correlation🡺 which airport is close to the all other aiport have almost same indegree & outdegree
* The correlation of closeness with indegree,outdegree& betweeness is very low🡺 which airport is closest to the other airport might not be a connecting airport for the flight to those. So its betweeness become also low value. The between closeness & betweeness showing the same
* High correlation for betweeness with indegree and outdegree 🡺 which airport have provide the shortest path between all other airport should have more number of connecting flights

**Python Code:-**

Simply pasted the Python code

import pandas as pd

import numpy as np

import networkx as nx

import matplotlib.pyplot as plt

# Degree Centrality

G = pd.read\_csv("G://network//Datasets\_Network Analytics//connecting\_routes.csv",sep=',')

E = pd.read\_csv("G://network//Datasets\_Network Analytics//flight\_hault.csv",sep=',')

G = G.iloc[:, 0:8]

G.columns= ["flights", "ID", "main airport","main Airport ID","Destination","Destination ID","haults","machinary"]

G.head(10)

E.columns= ["Airport ID","Name","City","Country","IATA\_FAA","ICAO","Latitude","Longitude","Altitude","Timezone","DST","Tz database timezone"]

E.head(10)

g = nx.Graph()

# only considering first 40 airport in list

g = nx.from\_pandas\_edgelist(G.iloc[0:40,:], source = 'main airport', target = 'Destination')

print(nx.info(g))

# Degree Centrality

b = nx.degree\_centrality(g)

print(b)

pos = nx.spring\_layout(g, k = 0.15)

nx.draw\_networkx(g, pos, node\_size = 25, node\_color = 'blue')

# closeness centrality

closeness = nx.closeness\_centrality(g)

print(closeness)

## Betweeness Centrality

b = nx.betweenness\_centrality(g) # Betweeness\_Centrality

print(b)

## Eigen-Vector Centrality

evg = nx.eigenvector\_centrality(g) # Eigen vector centrality

print(evg)

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:-**

* Closeness helps the client to choose the location for local retail stores
* Betweenness helps the client to find the location or nodes that’s provide a shortest connecting path betweeness between different groups.

Eg:- connecting airport for all other airports, location of central retail store for all other local retail stores

* Eighen values helps to find the influence or strength of a perticular person, service or node

**Q2)**

**Problem statement**

There are three datasets given (Facebook, Instagram, and LinkedIn). Construct and visualize the following networks:

* circular network for Facebook
* star network for Instagram
* star network for LinkedIn

Create a network using an adjacency matrix (undirected only). The snapshots of those datasets are given below:

**R code:-**

################### Part I: Network construction and visulaization ################################################

#install.packages("igraph")

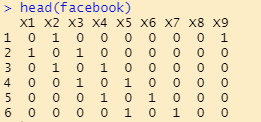
library("igraph")

?igraph

###### Example 1: facebook Circular Network using Adjacency Matrix ########

facebook <- read.csv("G://network//Datasets\_Network Analytics//facebook.csv", header=TRUE)

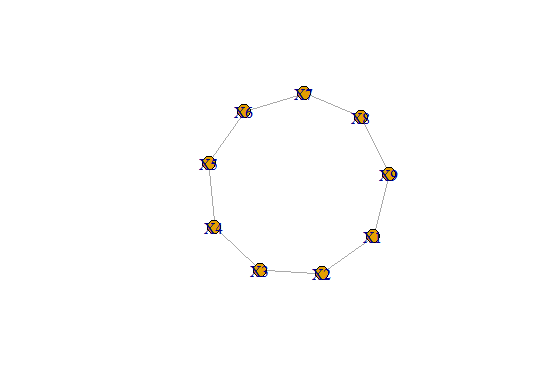
head(facebook)



# create a newtwork using adjacency matrix

facebk\_circle <- graph.adjacency(as.matrix(facebook), mode="undirected", weighted=TRUE)

plot(facebk\_circle)

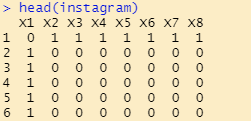


###### Example 2: instagram Star Network using Adjacency Matrix from file ##################

# Load the adjacency matrix from the csv file

instagram <- read.csv("G://network//Datasets\_Network Analytics//instagram.csv", header = TRUE)

head(instagram)



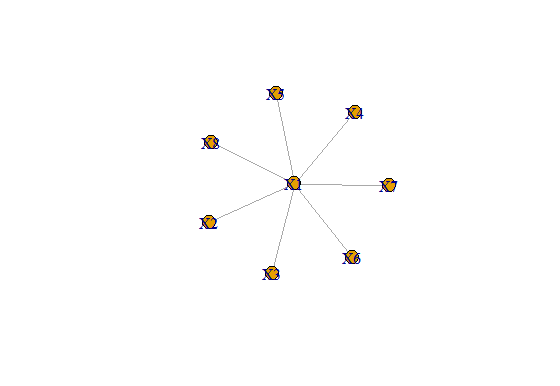
# create a newtwork using adjacency matrix

?graph.adjacency

# help file for the api graph.adjacency

instagram\_star <- graph.adjacency(as.matrix(instagram), mode = "undirected")

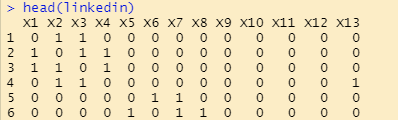
plot(instagram\_star)



###### Example 3: linkedin Star using Adjacency Matrix #########

linkedin <- read.csv("G://network//Datasets\_Network Analytics//linkedin.csv", header = TRUE)

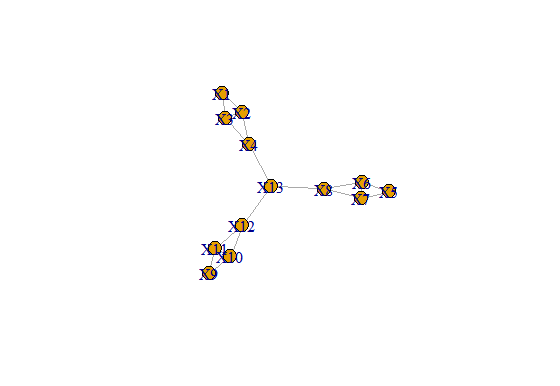
head(linkedin)



# create a newtwork using adjacency matrix

linke\_star <- graph.adjacency(as.matrix(linkedin), mode = "undirected", weighted = TRUE)

plot(linke\_star)



**Summary:-**

* facebook network have same closeness, degree centrality, betweenes for all nodes
* instgram network except X1 all other nodes have same and lower values for closeness, degree centrality, betweenes with that of X1
* linkedin network X13 have higher betweeness value. That provides the only shortest path between nodes of sub networks

**Problem Statement: -**

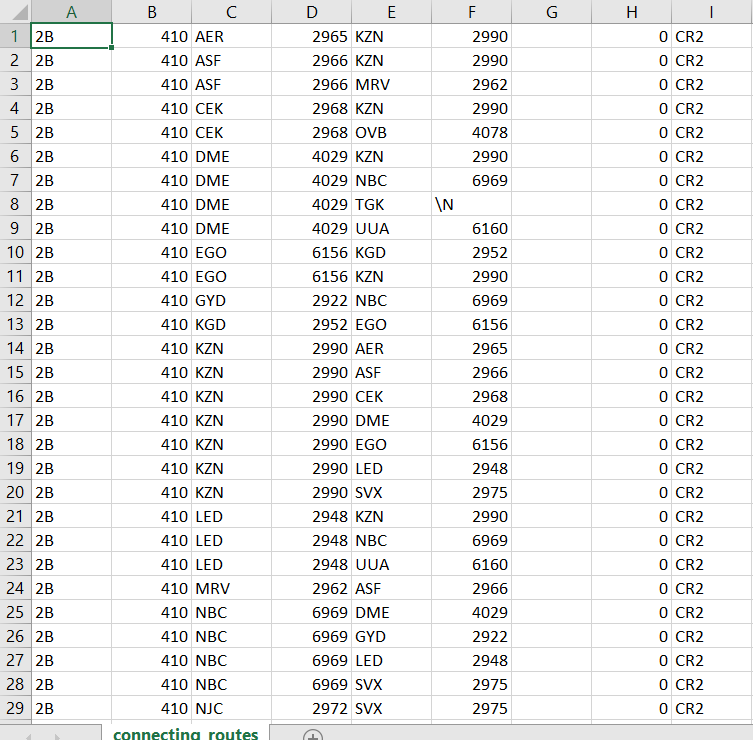
There are two datasets consisting of information for the connecting routes and flight halt. Create network analytics models on both the datasets separately and measure degree centrality, degree of closeness centrality, and degree of in-between centrality.

* Create a network using edge list matrix (directed only**).**
* Columns to be used ***in R*:**

Flight\_halt=c("ID","Name","City","Country","IATA\_FAA","ICAO","Latitude","Longitude","Altitude","Time","DST","Tz database time")

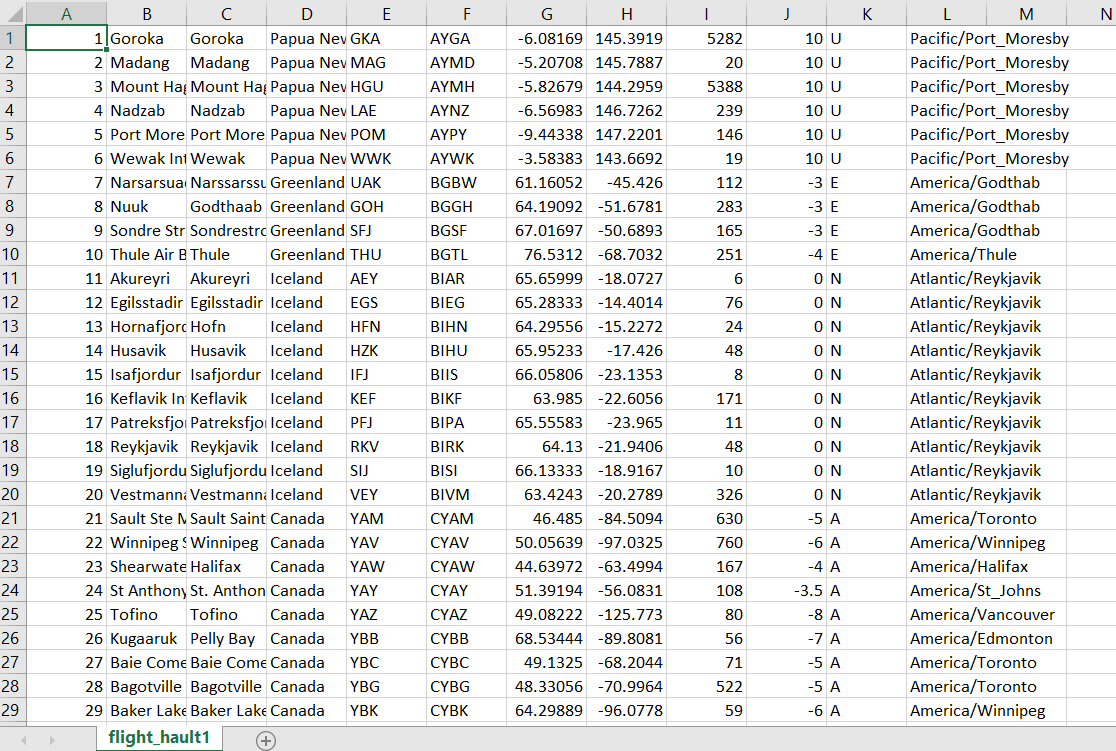
connecting routes=c("flights", " ID", "main Airport”, “main Airport ID", "Destination ","Destination ID","haults","machinary")

**connecting routes**





**Flight\_hault1**



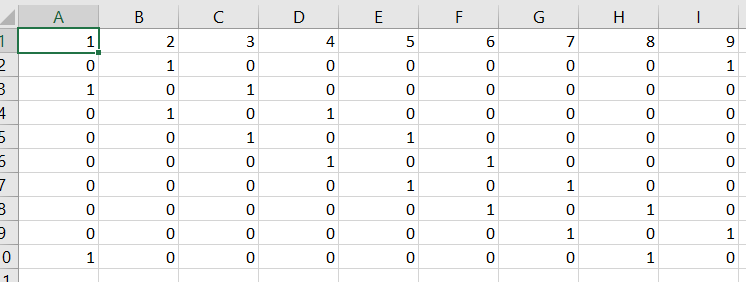
**Problem statement**

There are three datasets given (Facebook, Instagram, and LinkedIn). Construct and visualize the following networks:

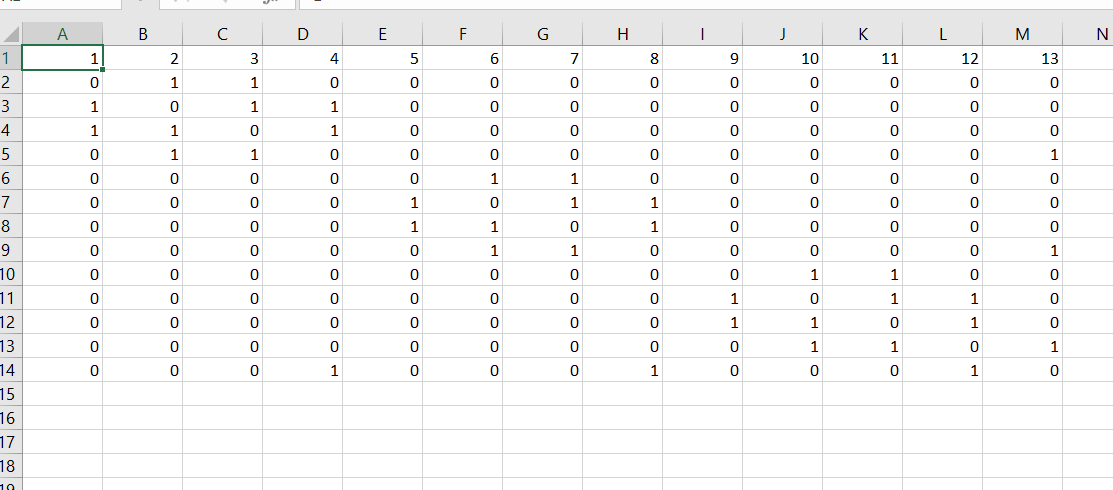
* circular network for Facebook
* star network for Instagram
* star network for LinkedIn

Create a network using an adjacency matrix (undirected only). The snapshots of those datasets are given below:

**Facebook**

****

**Instagram**

****

**LinkedIn**

