Topic: Network Analytics

1) Use the flights_hault and connecting routes datasets to find betweenness, closeness centrality, page rank, directed and undirected graphs.

Dataset: Flights_hault

column names:

ID","Name","City","Country","IATA_FAA","ICAO","Latitude","Longitude","Altitude","Time","DST","Tz database time"

Dataset: Connecting routes

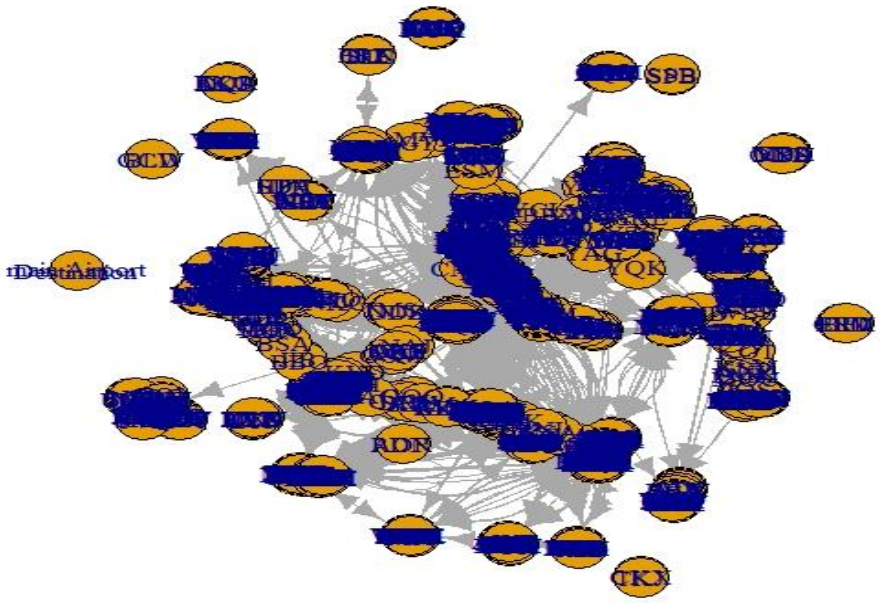
column names:

"flights", " ID", "main Airport","main Airport ID", "Destination ", "Destination ID", "haults", "machinary"

Solution

1) Loading the dataset of flights_haul
2) With the help of column names given, new columns are created for the dataset.
3) Loading the dataset of connecting_routes
4) With the help of column names given, new columns are created for the dataset.

Graphical Visualization:



1) Using the function **vcount** we can see how many airports are there in the network, so we got **3425** airports are available in the network using the dataset provided.
2) With the help of **ecount** function, we came to know that there are **67663** connections in the network.
3) With the help of **indegree** function, we came to know that **Atlanta airport** has most flights coming in, with the count of **911** flights and the given airport id is **3682**.
4) With the help of **Outdegree** function, we can clearly say that **Atlanta airport** has most flights going out, with the count of **915** flights and the given airport id is **3682.**

5) With the help of **closeness** function, we can find out which of the airport is close to most of the airport, here **FRA (Frankfurt, Germany)** is the closet airport with closeness value of 0.018 with the airport code **338**.
6) Using **betweenness** function, we found that **LAX (Los Angeles, U.S.A)** is the most common routes and hence it is the international hub for all the points in center with a betweenness value of 0.088 with the airport code **3484**.

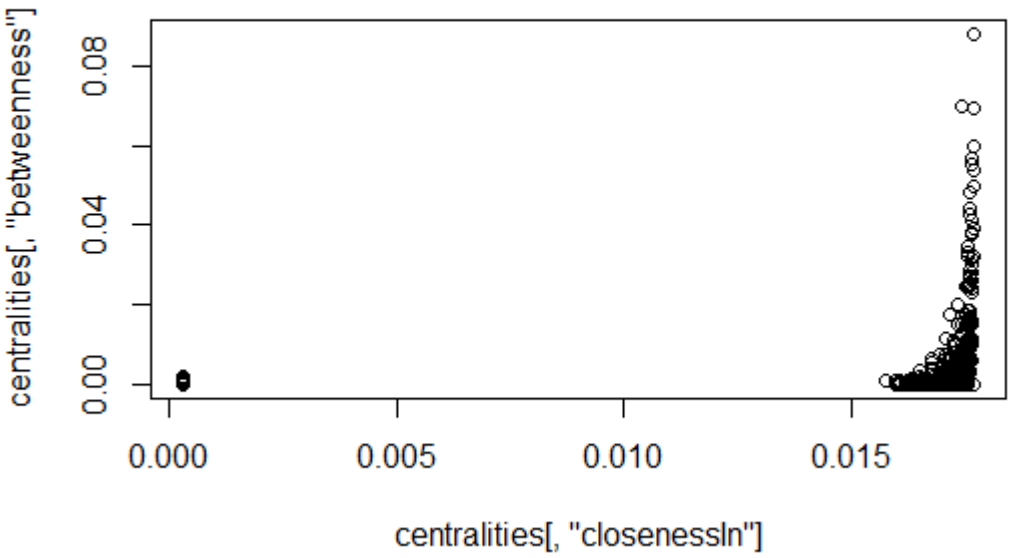
7) So, now taking the degree, closeness and betweenness centralities together below are the top 6 airports shown.

	inDegree	outDegree	closenessIn	betweenness
AER	26	26	0.01801138	2.057750e-05
KZN	28	28	0.01801868	2.211094e-05
ASF	8	8	0.01796951	3.970297e-07
MRV	22	22	0.01805165	2.315226e-05
CEK	20	20	0.01797724	2.473746e-06
OVB	90	87	0.01813761	1.385724e-03

8) Now we did the correlation for the centralities using --> cor(centralities)

```
> cor(centralities)
              inDegree  outDegree closenessIn betweenness
inDegree    1.00000000 0.99977112  0.08950814  0.73365075
outDegree   0.99977112 1.00000000  0.08900154  0.73374679
closenessIn 0.08950814 0.08900154  1.00000000  0.04768086
betweenness 0.73365075 0.73374679  0.04768086  1.00000000
>
```

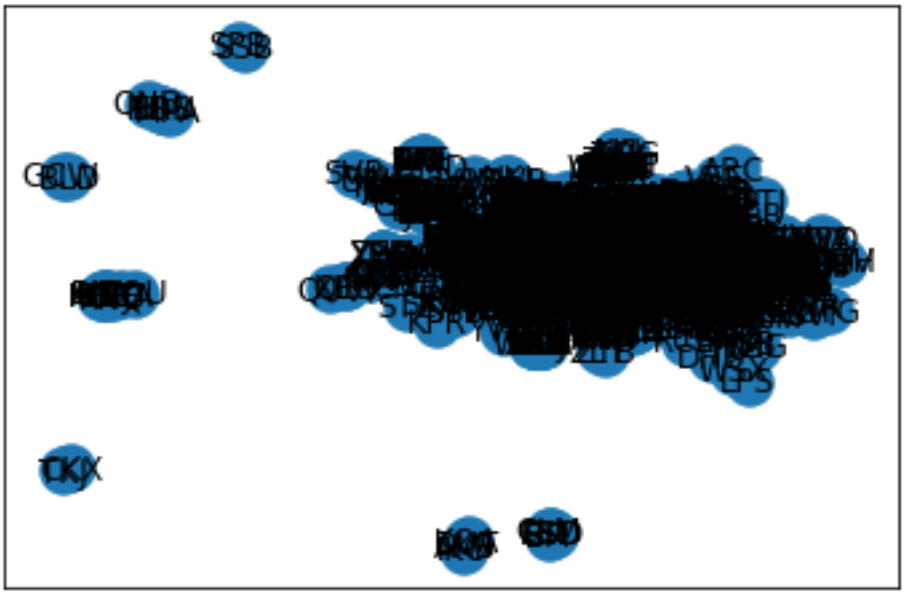
Pair Plot for closeness and betweenness:



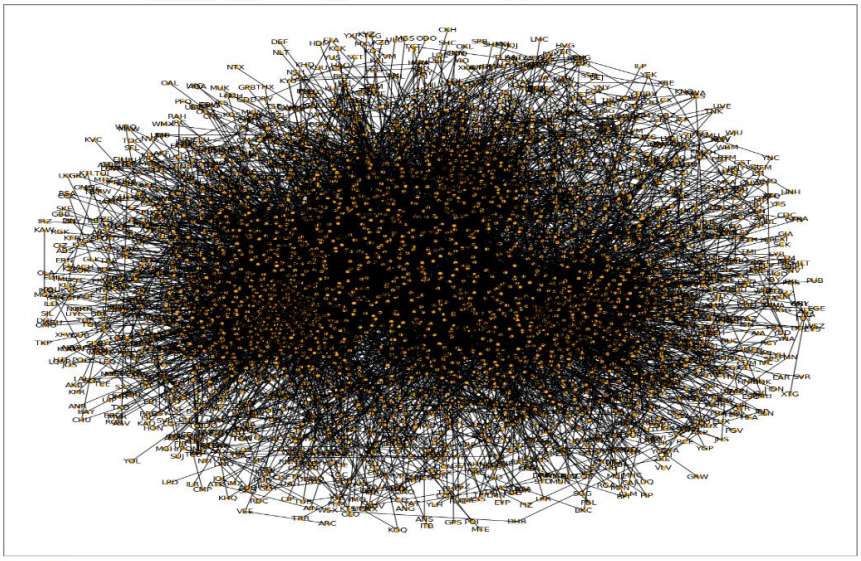
9) Using **eigen_centrality** function, we say that Atlanta is the most important airport with **eigen value 0.27.**

10) Using page rank(pg_rank) function, we say that **Atlanta Airport(U.S.A)** is ranked as the best airport in google page rank system.

Eigen-Vector Centrality:



Betweenness Centrality Graph:

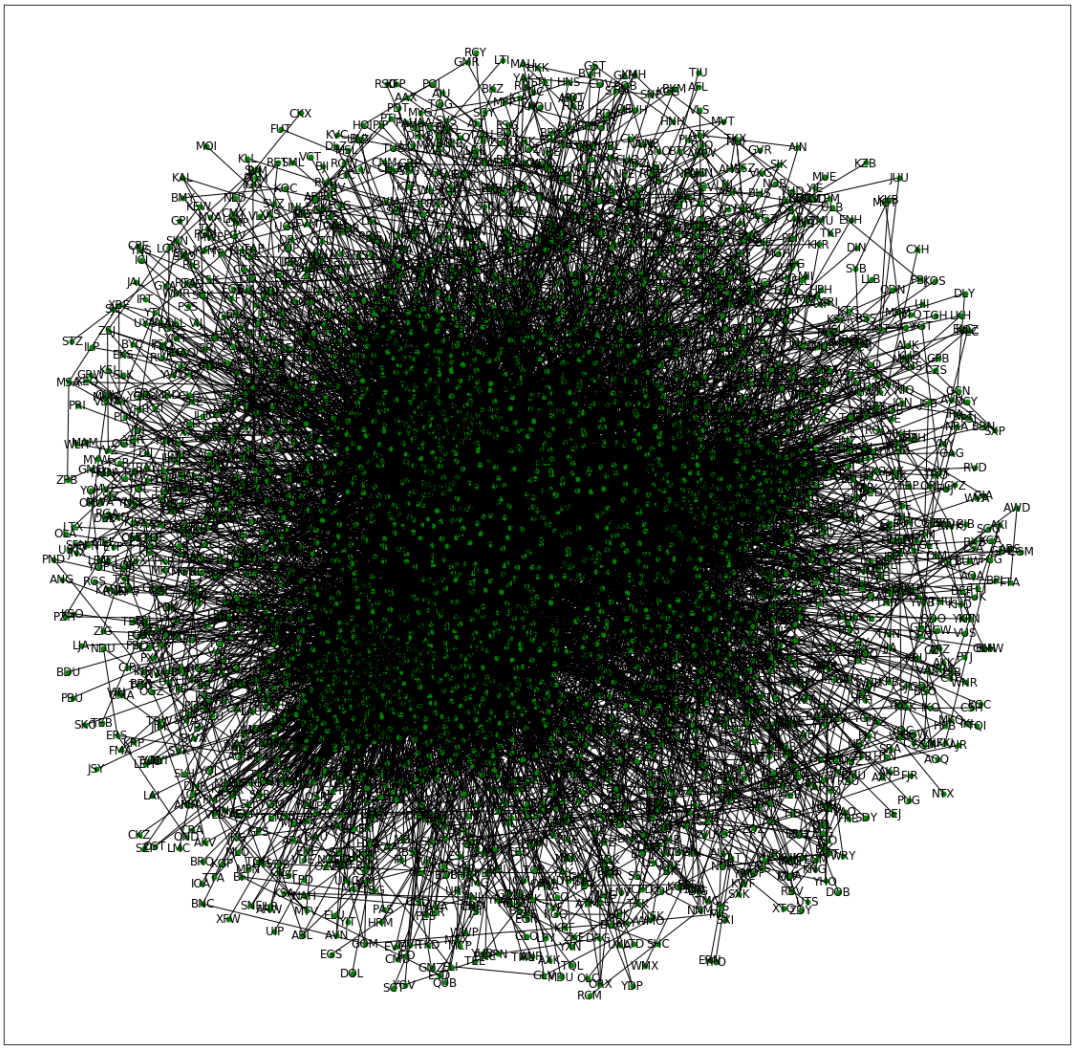


Betweenness Centrality:

- Number of nodes: 3425
- Number of edges: 19257
- Average degree: 11.2450

Closeness Centrality:

- Number of nodes: 3425
- Number of edges: 19257
- Average degree: 11.2450



Average Clustering value is 0.48

Cluster Co-efficient is 3425.

2) Business Problem: Three data sets of social media networks are attached.

- plot circular graph

- Star Graph for the three data sets

- Check for its nodes and directed and undirected graphs

Draw your insights on the data.

Solution:

A Graph is a non-linear data structure consisting of nodes and edges. The nodes are sometimes also referred to as vertices and the edges are lines or arcs that connect any two nodes in the graph.

Graphical Representation of Facebook

Node Graphs

Diagram 1 (g1)

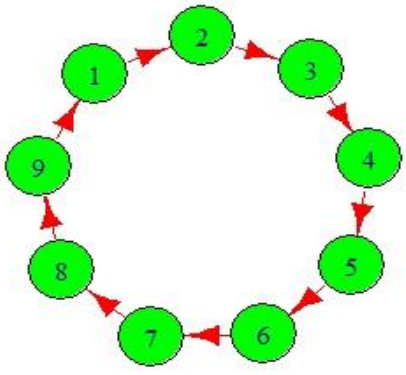
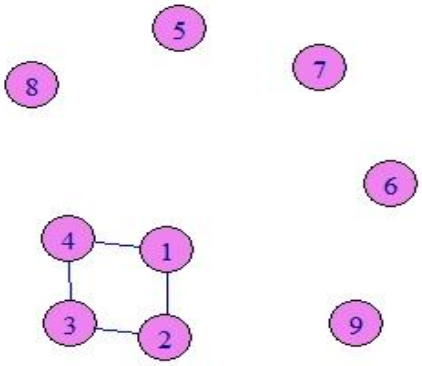


Diagram 2 (g2)



- ✓ In the above both the graphs, the set of vertices “V” = {1,2,3,4,5,6,7,8,9} and the set of edges “E” = { 12, 23, 34, 41} in diagram 2 which are

undirected but whereas, in diagram1 we can see they are all directed to each other.

- ✓ Here we plotted the graph, using vertex.color & Size and also selected color for Edges.
- ✓ Using V(g2) & E (g2) we get

  V(g2)

  → 9/9 vertices, from 0f 14867:

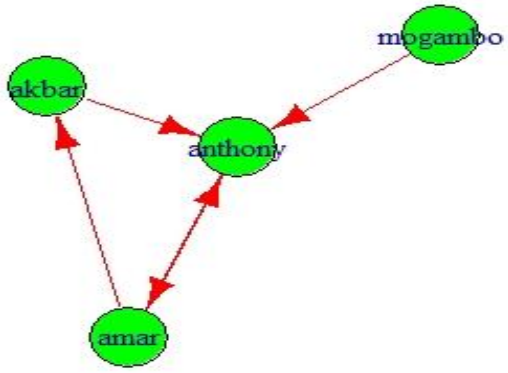
  [1] 1 2 3 4 5 6 7 8 9

  > E(g2)

  → 4/4 edges from 0f 14867:

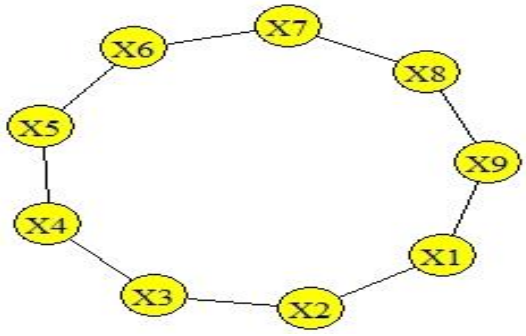
  [1] 1--2 2--3 3--4 1--4

Diagram 3 (g3)



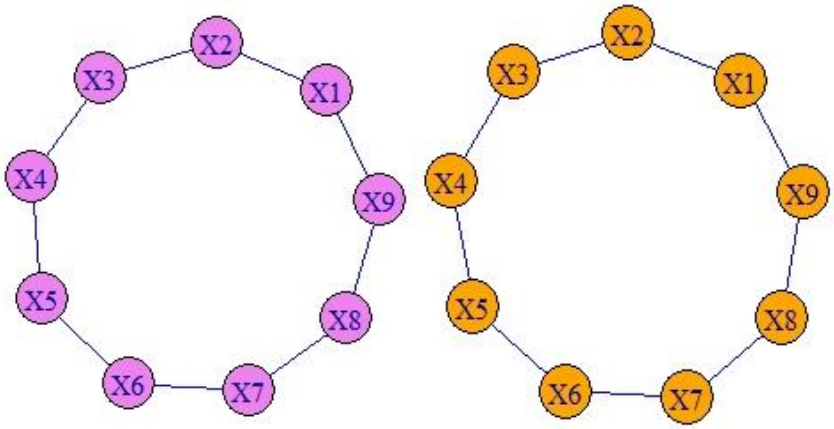
- ✓ In the above graph, amar,akbar & Anthony are connected to each other with a direction so it is called as Directed Graph and mogambo is the only one who is connected to rest all but not redirected by anyone. However if you see Anthony has direction from amar & akbar so it is called as Parallel.

Circular Graph:



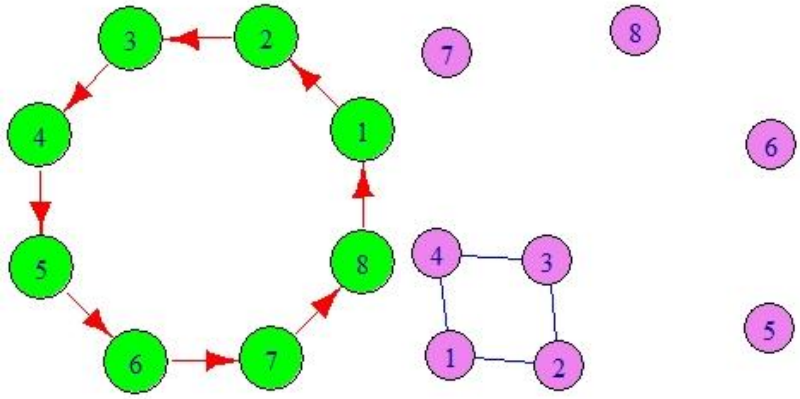
- ✓ In the above graph, we have 9 nodes which are undirected graph i.e., the nodes are connected but they do not have any direction.

Star Graphs Adjacency Matrix:



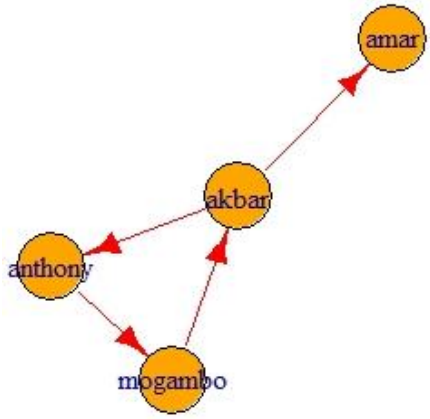
Graphical Representation of Instagram:

Node Graphs:



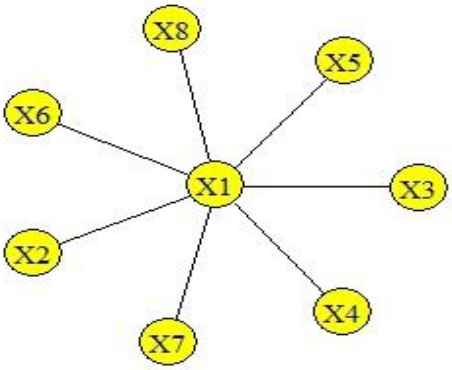
- ✓ In the above both the graphs, the set of vertices "V" = {1,2,3,4,5,6,7,8,9} and the set of edges "E" = { 12, 23, 34, 41} in diagram 2 which are undirected but whereas, in diagram1 we can see they are all directed to each other.

- ✓ Here we plotted the graph, using vertex.color & Size and also selected color for Edges.



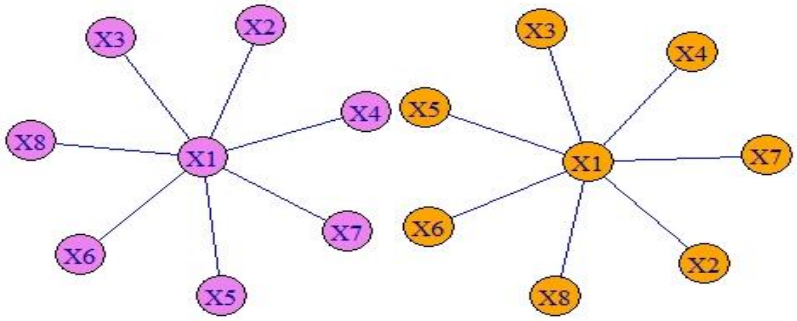
- ✓ In the above graph, Anthony, mogambo & akbar are connected to each other with a direction so it is called as Simple Directed Graph and Amar is the only one who is connected only by Akbar but not by rest of all.

Circular Graph:



- ✓ Here , there are total of 8 Nodes / Vertices and Node X1 is connected by all nodes and they are not having the direction,hence it is called as Undirected Graph.

Star Graphs Adjacency Matrix:

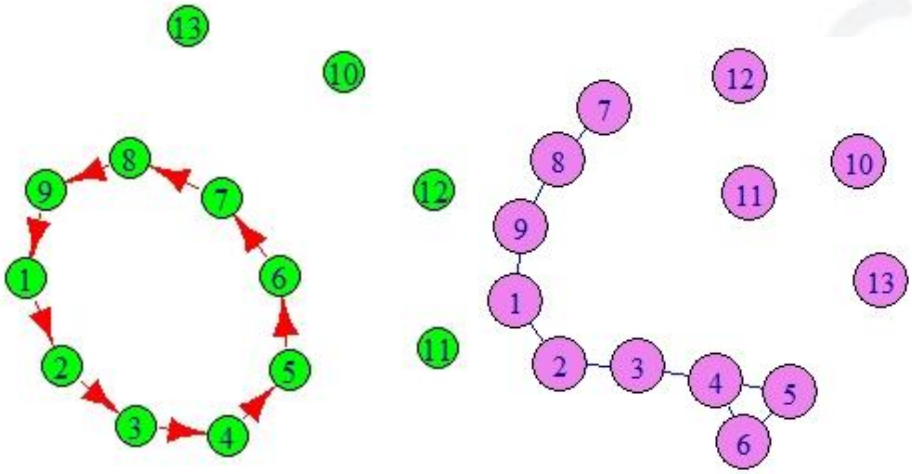


- ✓ Here , there are total of 8 Nodes / Vertices and Node X1 is connected by all nodes and they are not having the direction,hence it is called as Undirected Graph.

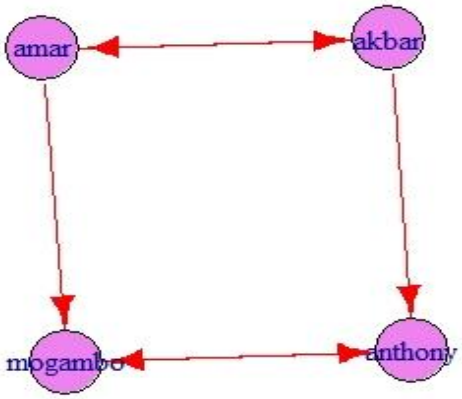
Graphical Representation of LinkedIn:

Node Graphs:

Diagram 1 **Diagram 2**

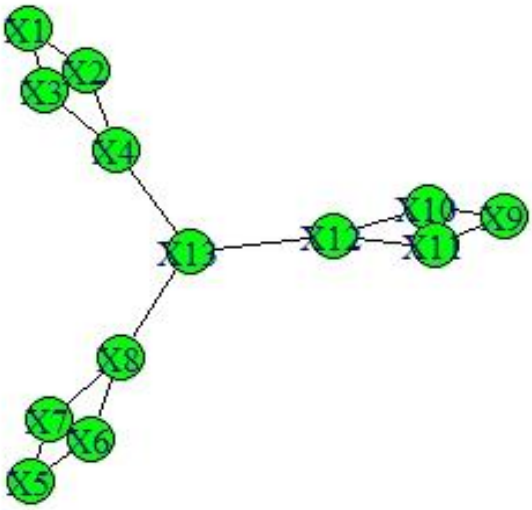


- ✓ In the above both the graphs, there are 13 set of vertices "V" = {1,2,3,4,5,6,7,8,9,10,11,12,13} and the set of edges "E" = {12,23,34,45,56,67,78,89,91} in diagram 1 and "E" = {12, 23, 34, 45,56,64,78,89,91} in diagram 2 and both graph are Directed Graph as they have directions and the left over nodes / vertices are called as isolated.



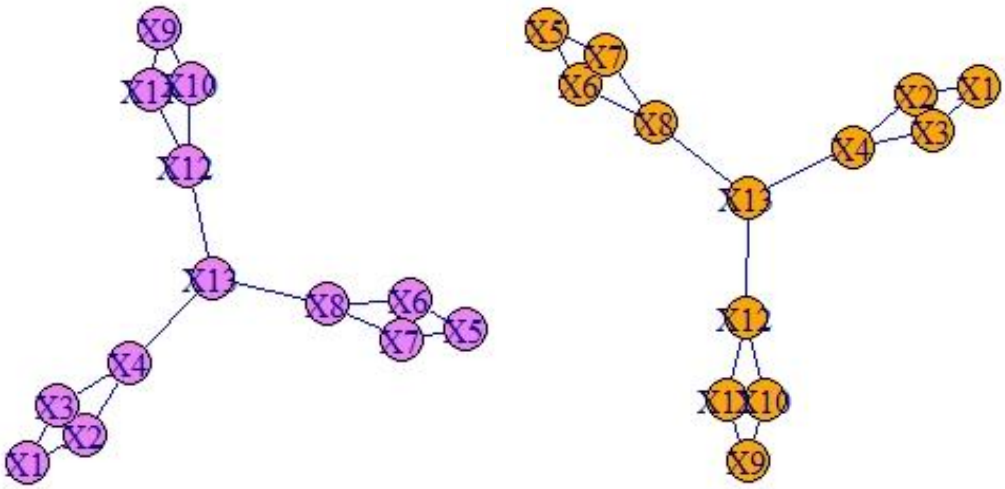
- ✓ In the above graph, all the nodes are directed so it is called as Directed Graph, However, Amar & Akbar have direction and also follow each other similarly mogambo & Anthony also have direction and also follows each other.

Circular Graphs:



- ✓ **In the above graph, Node 13 is connected to node 12 , node 4 & node 8 and a group of 4 nodes are connected (1, 2, 3, 4) , (5, 6, 7, 8) , (9, 10, 11, 12) and these group are not having directions respectively. However, following each other is seen in few patterns like (X2 & X3 followed by X1 & X2) , (X6 & X7 followed by X5 & X8) , (X10 & X11 followed by X9 & X12).**

Star Graphs Adjacency Matrix:



Inferences

- ✓ **Social data can help us create better content like Facebook we will able to see whether images, videos or links do better or in visual centric networks like Instagram we can check what type of images perform best.**
- ✓ **They also help us understand what are our best social networks are like not just Facebook or Instagram who has over 2 billion users or 800 million users respectively but also small social networks like Flickr / Pinterest can also help us execute strategy better.**
- ✓ **These kinds of analytics also help us to understand our competitors who also collects data and get the unique data. So, using these analytics we will be able to figure out what is trending and what is failing. It will also help us to avoid any mistakes our competitors are making and focus only on techniques that will fetch us good results.**
- ✓ **Finally, it will us understand our audience and their likes and dislikes so that we can act accordingly. Lastly timing is the most important key in social media marketing considering that fact if we post anything when they are online / available, our posts will drive more engagement, traffic and sales.**

Instructions:

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

Using R and Python codes perform:

2. Data Pre-processing

2.1 Data cleaning, Feature Engineering etc.

3. Exploratory Data Analysis (EDA):

3.1. Summary

3.2. Univariate analysis

3.3. Bivariate analysis

4. Model Building

4.1 Build the Network model on the given data sets.

4.2 Perform the Network Analytics.

4.3 Model(s) Improvement (Try with different no. of clusters)

5. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

6. problem statement information and data dictionary is given in next page.

Note:

The assignment should be submitted in the following format:

- **R code**
- **Python code**
- **Code Modularization should be maintained**
- **Documentation of the modules (elaborating on steps mentioned above).**