Data Science Challenge

Centrality metrics

Delermando Branquinho Filho

10 de novembro de 2016

Fist Challenge

In this challenge, suppose we are looking to do social network analysis for prospective customers. We want to extract from their social network a metric called "closeness centrality".

Centrality metrics try to approximate a measure of influence of an individual within a social network. The distance between any two vertices is their shortest path. The *farness* of a given vertex v is the sum of all distances from each vertex to v. Finally, the *closeness* of a vertex v is the inverse of the *farness*.

The first part of the challenge is to rank the vertices in a given graph by their *closeness*. The graph is provided in the attached file; each line of the file consists of two vertex names separated by a single space, representing an edge between those two nodes.

Load Libraries

```
library(GGally)
library(network)
library(sna)
library(ggplot2)
```

Seting seed for reproductble researsh

```
set.seed(3)
```

Loadding Dataset

```
myData <- as.matrix(read.table("D:/Data Science/Semantix/edges.dat", quote="\"",
                    comment.char="", stringsAsFactors=FALSE))
```

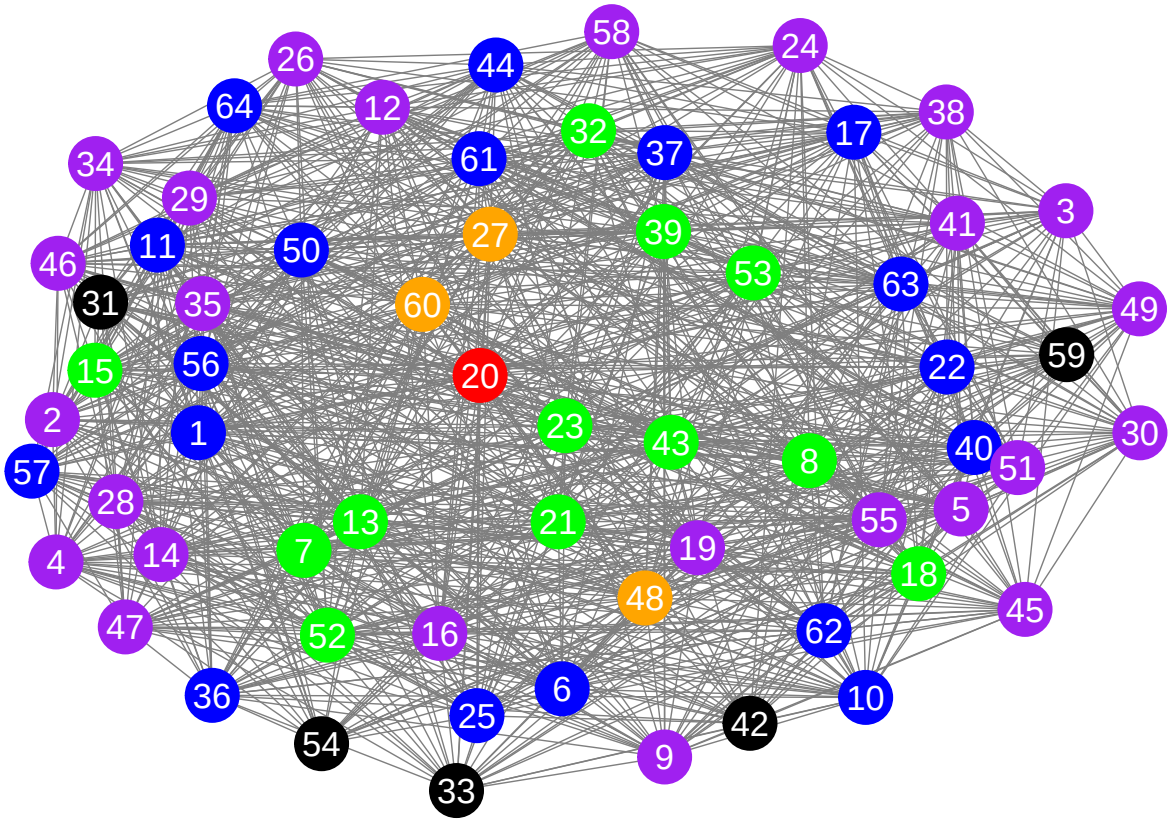
Creating Network Data

The nodes are numbered by one to number of nodes.

```
net = rgraph(myData, mode = "graph")
net = network(net, directed = FALSE)
network.vertex.names(net) = 1:dim(myData)[1]
```

Plotting

```
c.d   <- degree(net)
col<- as.integer(5*(c.d-min(c.d))/diff(range(c.d))+1)
palette <- c("black","purple","blue","green","orange","red")
g<-ggnet2(net, label = TRUE, label.color = "white",color = palette[col])
plot(g,vertex.color=palette[col],main="Degree Centrality",
     layout=layout.fruchterman.reingold)
```



Second Challenge

The second part of the challenge is to create a RESTful web server with endpoints to register edges and display the centrality of the graph.

Degree

The node degree is the number of relations (edges) of the nodes.

```
degree(net)
```

```
##  [1] 68 60 58 60 60 66 70 70 62 66 68 62 74 60 72 60 64 72 60 84 72 68 74
## [24] 62 64 62 80 56 60 58 54 72 54 62 62 64 66 56 70 64 56 54 74 64 58 62
## [47] 58 78 58 68 62 70 70 50 60 64 64 56 48 80 66 64 68 68
```

Betweenness centrality is even more important statistical property of a network. This is applied in a lot of real-world problems, such as finding influential people in a social network, finding crucial hubs in a computer network, finding border crossing points which have a largest traffic or trade flow.

```
degree(net,cmode="indegree")
```

```
##  [1] 34 30 29 30 30 33 35 35 31 33 34 31 37 30 36 30 32 36 30 42 36 34 37
## [24] 31 32 31 40 28 30 29 27 36 27 31 31 32 33 28 35 32 28 27 37 32 29 31
## [47] 29 39 29 34 31 35 35 25 30 32 32 28 24 40 33 32 34 34
```

Closeness (farness) centrality indicates how long it will take for information from a given node to reach other nodes in the network. The smaller the value, the more central role the node plays in the network.

```
degree(net,cmode="outdegree")
```

```
##  [1] 34 30 29 30 30 33 35 35 31 33 34 31 37 30 36 30 32 36 30 42 36 34 37
## [24] 31 32 31 40 28 30 29 27 36 27 31 31 32 33 28 35 32 28 27 37 32 29 31
## [47] 29 39 29 34 31 35 35 25 30 32 32 28 24 40 33 32 34 34
```

Density

The density of a graph is the number of existing edges divided by the number of possible ones (assuming no duplicates or loops)

```
gden(net,mode="graph")
```

```
## [1] 0.5104167
```

Connectedness takes one or more graphs and returns the Krackhardt connectedness scores

```
connectedness(net)
```

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

Centrality

Centralization returns the centralization GLI (graph-level index) for a given graph in dat, given a (node) centrality measure FUN. Centralization follows Freeman's (1979) generalized definition of network centralization, and can be used with any properly defined centrality measure.

```
centralization(net,degree,mode="graph")
```

```
## [1] 0.1612903
```

Betweenness Centrality

A measure of the degree to which a given node lies on the shortest paths (geodesics) between other nodes in the graph The centroid point of the map is:

```
betweenness(net)
```

```
##  [1] 30.05947 25.19998 20.75964 26.58483 27.17971 30.68609 34.43803
##  [8] 41.01433 28.26190 29.03885 32.85947 27.63613 41.97873 25.10197
## [15] 35.29942 27.49838 29.34142 37.08752 26.96945 55.08536 43.59232
## [22] 35.95958 44.04161 27.42004 30.85394 27.27176 50.24379 22.72735
## [29] 29.66883 25.44973 20.48037 36.48415 20.24311 25.92500 27.39319
## [36] 30.08765 31.07328 22.39590 34.71839 32.45347 23.00109 19.97051
## [43] 49.53453 29.43356 23.96949 24.80950 21.48905 46.42623 20.58471
## [50] 36.50695 29.81747 36.61419 38.73116 18.18571 29.64120 29.18628
## [57] 29.37676 21.83202 15.18298 49.24428 31.09992 31.82963 30.82478
## [64] 36.14385
```

```
centralization(net,betweenness,mode="graph")
```

```
## [1] 0.006304762
```

Closeness Centrality

Closeness Centrality (CLC) is a category of measures that rate the centrality of a node by its closeness (distance) to other nodes. Closeness Centrality decreases if either the number of nodes reachable from the node in question decreases, or the distances between the nodes increases.

```
closeness(net)
```

```
##  [1] 0.6847826 0.6562500 0.6494845 0.6562500 0.6562500 0.6774194 0.6923077
##  [8] 0.6923077 0.6631579 0.6774194 0.6847826 0.6631579 0.7078652 0.6562500
## [15] 0.7000000 0.6562500 0.6702128 0.7000000 0.6562500 0.7500000 0.7000000
## [22] 0.6847826 0.7078652 0.6631579 0.6702128 0.6631579 0.7325581 0.6428571
## [29] 0.6562500 0.6494845 0.6363636 0.7000000 0.6363636 0.6631579 0.6631579
## [36] 0.6702128 0.6774194 0.6428571 0.6923077 0.6702128 0.6428571 0.6363636
## [43] 0.7078652 0.6702128 0.6494845 0.6631579 0.6494845 0.7241379 0.6494845
## [50] 0.6847826 0.6631579 0.6923077 0.6923077 0.6237624 0.6562500 0.6702128
## [57] 0.6702128 0.6428571 0.6176471 0.7325581 0.6774194 0.6702128 0.6847826
## [64] 0.6847826
```

```
centralization(net,closeness,mode="graph")
```

```
## [1] 0.1590797
```

Egocentricity The egocentric network (or ego net) of vertex v in graph G is defined as the subgraph of G induced by v and its neighbors.

```
newEgo <- ego.extract(net,6)
head(closeness(newEgo))
```

```
##              6
## [1,] 1.0000000
## [2,] 0.6600000
## [3,] 0.7173913
## [4,] 0.7021277
## [5,] 0.7173913
## [6,] 0.7021277
```

REFERENCES

Alex, B. 1950. Communication patterns in task-oriented groups. J. Acoust. Soc. Am. 22 (6): 725-730.
Sabidussi, G. 1966. The centrality index of a graph. Psychometrika. 31: 581-603.