FIT3152 Data analytics. Tutorial 05: Network Analysis

Pre-tutorial Activity

The dataset “UKfaculty” from the igraphdata package depicts the personal friendship network of a faculty at a UK university. The network is given as a directed, weighted graph.

1. Explore the degree distribution of nodes in the network

2. Calculate network statistics to identify key players in the network. For example which node/player has the most hub potential in the network.

3. Identify and plot community groups in the network considering edge betweenness and greedy modularity optimization. Hint: Treat the graph as undirected.

Tutorial Activities

Download and install package functions using:

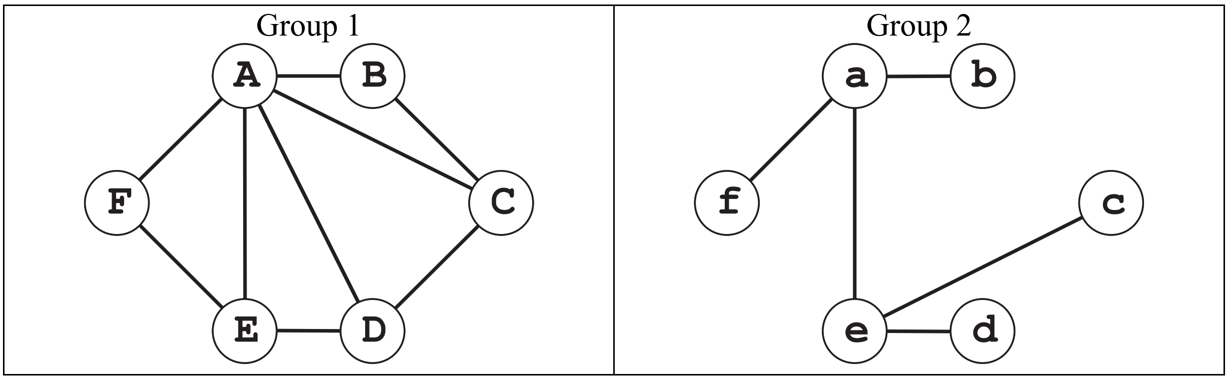
```
install.packages(c("igraph", "igraphdata"))
library(igraph)
library(igraphdata)
```

Formulas from lecture slides

Density	**Clustering/transitivity coefficient**
$$den(g) = \frac{\lvert E_g \rvert}{\lvert V_g \rvert (\lvert V_g \rvert - 1)/2}$$ where $\lvert E_g \rvert$ is number of edges, $\lvert V_g \rvert$ is number of vertices	$$clt(g) = \frac{3\tau_{\Delta}(g)}{\tau 3(g)}$$ where $3\tau_{\Delta}(g)$ is number of triangles, $\tau 3(g)$ is number of connected triples
Closeness Centrality	**Betweenness Centrality**
$$c_{Cl}(v) = \frac{1}{\sum_{u \in V} dist(u, v)}$$	$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s, t \mid v)}{\sigma(s, t)}$$ $\sigma(s, t)$ is the number of shortest paths between *s* and *t*, $\sigma(s, t \mid v)$ is the number of shortest paths between *s* and *t* passing through *v*

1 Work through the analysis on the lecture slides. In particular the Research Collaborators and the Karate networks.

2 Two student groups are based on the friendship networks below.



(a) Calculate the following graph and vertex measures by hand for each group:
<u>Degree Distribution</u>
<u>Betweenness</u>
<u>Closeness</u>
<u>Diameter</u>
Draw the distance matrix and calculate <u>Average Path Length</u>
Draw the <u>Adjacency Matrix</u>

(b) Create each graph using the igraph package in R and check your answers.

Tips

```
You might want to use this example from Lecture 5.
g <- graph.formula(J-D, J-R, J-A, A-S, A-R, A-M, S-R)
```

(c) Using any of the network measures covered in the lecture describe the difference between the two networks. Identify the most powerful individual in each of the networks with respect to their ability to control information flow. Can you describe either of the graphs in terms of the network topologies given in Slide 64 of the lecture notes?

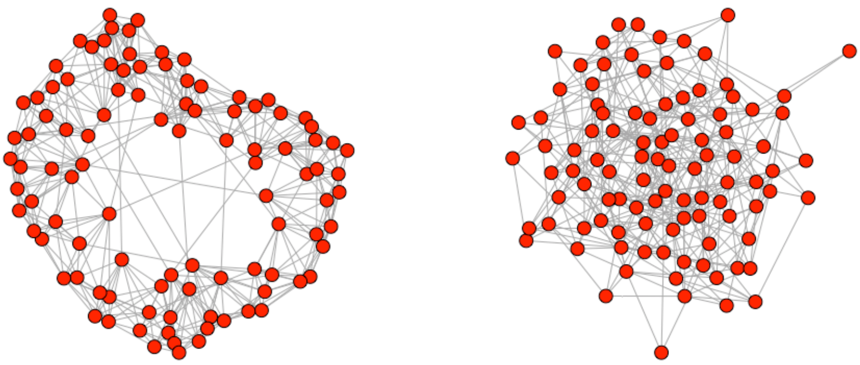
3 A group of friends have the following network (data on Moodle as Friends.csv)

	A	B	C	D	E	F	G	H	I	J
A	0	1	1	1	0	1	1	1	1	1
B	1	0	1	0	1	1	1	1	0	1
C	1	1	0	1	0	0	1	1	0	1
D	1	0	1	0	1	0	0	0	0	0
E	0	1	0	1	0	1	1	1	1	1
F	1	1	0	0	1	0	1	1	0	0
G	1	1	1	0	1	1	0	0	1	1
H	1	1	1	0	1	1	0	0	1	1
I	1	0	0	0	1	0	1	1	0	0
J	1	1	1	0	1	0	1	1	0	0

Describe the network. Who is the most dominant member of the group?

4 The two networks on Slides 38 and 39 of the lecture notes were generated with the following script. By using an appropriate choice of network statistics and vertex importance/centrality measures comment on (a) how the two networks are different, and (b) which node/nodes are the most important in each network.

(a) (b)



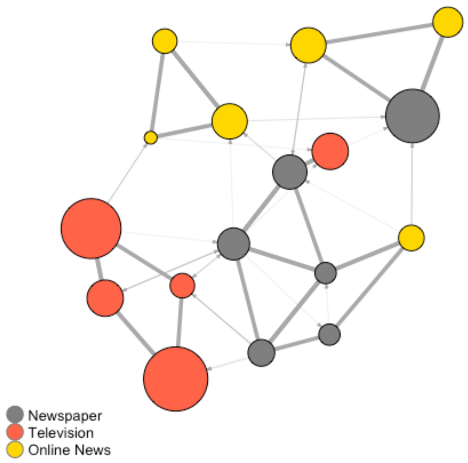
```
library(igraph)
set.seed(999)
g <- sample_smallworld(1, 100, 5, 0.03)
V(g)$label <- NA
V(g)$size <- 8
V(g)$color <- "red"
h <- erdos.renyi.game(100, 1/15)
V(h)$label <- NA
V(h)$size <- 8
V(h)$color <- "red"
par(mfrow=c(1,2))
plot(g, layout =  layout.fruchterman.reingold)
plot(h, layout =  layout.fruchterman.reingold)
```

5 Create a Complete, Ring, Tree and Star graph each of 8 vertices (using the code in the lecture notes). Calculate the network and vertex statistics and explain, with reasons, which network structure is most robust (that is a failure in any node will have least effect on information flow through the network) and by contrast, which structure is most fragile.

6 Use the following data and code to generate a basic graph.

```
library("igraph")
nodes <- read.csv("Dataset1-Media-Example-NODES.csv", header=T, as.is=T)
links <- read.csv("Dataset1-Media-Example-EDGES.csv",header=T, as.is=T)
# Create Igraph object
net <- graph_from_data_frame(d=links, vertices=nodes, directed=T)
plot(net)
```

Using the example in https://kateto.net/wp-content/uploads/2018/06/Polnet%202018%20R%20Network%20Visualization%20Workshop.pdf
try and improve the plot, along the lines of the example given.



7 Create a random graph based on the scale-free Barabási-Albert model using the code below.

```
> set.seed(9999)
> BA <- sample_pa(100)
```

Now create another graph according to Erdós-Rényi random graph model.

```
> set.seed(9999)
> ER <- sample_gnm(100, 100)
```

Now create a Small World graph according to the Watts and Strogatz model.

```
> set.seed(9999)
SW <- sample_smallworld(1, 100, 5, 0.05)
```

(a) Using any of the techniques covered in the lecture, such as degree distribution, cliques and the closeness measures comment on the main differences between the graphs.

(b) Which type of network has more powerful individuals with respect to their ability to control information flow in the network?

8 The file rfid contains encounter network data for staff in a hospital. Looking at the simplified network, describe the network and calculate summary statistics. Using the results of your analysis, identify the most important people in the network. Use the following code to create the data

```
> library(igraph)
> library(igraphdata)
> data(rfid)
> nrfid <- rfid
> nrfid <- simplify(nrfid, remove.multiple = TRUE, remove.loops = TRUE,
+                   edge.attr.comb = getIgraphOpt("edge.attr.comb"))
> plot(nrfid)
```

9 The following data records the attendance 18 women at social functions over a 9 month period during the 1930s. The data was recorded by Davis, and reported in Davis, A., Gardner, B. B. and M. R. Gardner (1941) *Deep South, Chicago*: The University of Chicago Press.

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	X	X	X	X	X	X		X	X					
2. Miss Laura Mandeville	X	X	X		X	X	X	X						
3. Miss Theresa Anderson		X	X	X	X	X	X	X	X					
4. Miss Brenda Rogers	X		X	X	X	X	X	X						
5. Miss Charlotte McDowd			X	X	X		X							
6. Miss Frances Anderson			X		X	X		X						
7. Miss Eleanor Nye					X	X	X	X						
8. Miss Pearl Oglethorpe						X		X	X					
9. Miss Ruth DeSand					X		X	X	X					
10. Miss Verne Sanderson							X	X	X			X		
11. Miss Myra Liddell								X	X	X		X		
12. Miss Katherine Rogers								X	X	X		X	X	X
13. Mrs. Sylvia Avondale							X	X	X	X		X	X	X
14. Mrs. Nora Fayette						X	X		X	X	X	X	X	X
15. Mrs. Helen Lloyd							X	X		X	X	X		
16. Mrs. Dorothy Murchison								X	X					
17. Mrs. Olivia Carleton									X		X			
18. Mrs. Flora Price									X		X			

In the table above, an X indicates the attendance by a woman at an event. By treating each pair of women who attended the same event as being 'connected' and aggregating this over the 14 meetings construct a social network for the women using the following methods:

(a) Treat the graph as unweighted. That is, the number of meetings each pair of women attended is not counted (1 if the pair were present at any meeting and 0 otherwise), and

(b) Treat the graph as weighted. That is, each pair of women attending each meeting is counted. For example, the first pair of women (Jefferson and Mandeville) would have a weight of 6 since they both attended 6 meetings where the other was in attendance.

You may want to create a csv file for this problem.

Analyse the data and describe the social network formed using both methods. Are there differences in your analysis for part (b) that become apparent using weighted edges.

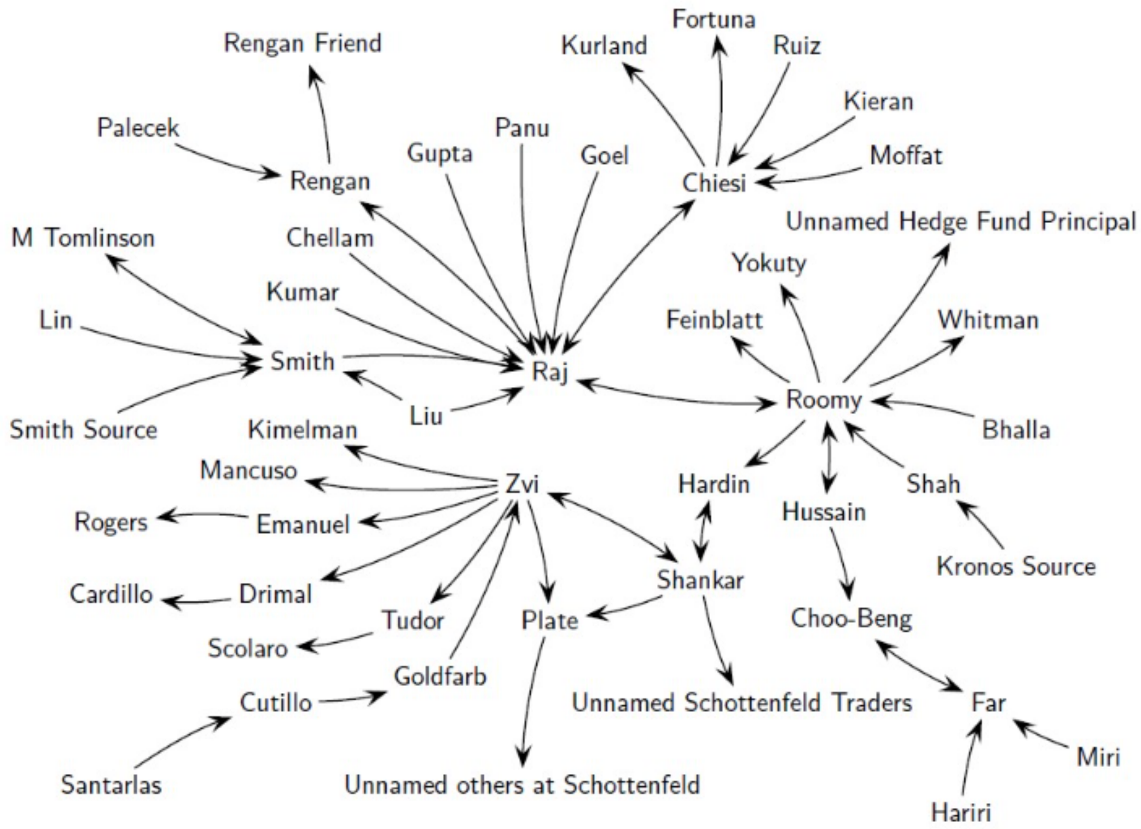
For more information on the data see:
http://svitsrv25.epfl.ch/R-doc/library/latentnet/html/davis.html

Tips

```
You might want to use these examples from Lecture 5 as a way of creating
the social network:

# To make a complete graph from a set of vertices
# from https://igraph.org/r/doc/graph_from_literal.html
gg = graph_from_literal(A:B:C:D:E:F:G -- A:B:C:D:E:F:G)
# Make a second graph with some shared players
hh = graph_from_literal(A:B:H:I:J -- A:B:H:I:J)
# now make a union
# from https://igraph.org/r/doc/union.igraph.html
ii = (gg %u% hh)

# make a data frame of people and club membership
Person = as.data.frame((c("A", "B", "C", "D", "E", "F", "G", "H", "I",
"J", "A", "B")))
Club = as.data.frame((c("X", "X", "X", "X", "X", "X", "X", "Y", "Y", "Y",
"Y", "Y")))
ClubData = cbind(Person, Club)
colnames(ClubData) = c("Person", "Club")
UniquePerson = unique(Person)
colnames(UniquePerson) = "Person"

g <- make_empty_graph(directed = FALSE)

# add vertices
for (i in 1 : nrow(UniquePerson)) {
  g <- add_vertices(g, 1, name = as.character(UniquePerson$Person[i]))
}

# make complete graph for each club and add to g
for (k in unique(ClubData$Club)){
temp = ClubData[(ClubData$Club == k),]

# combine each pair of agents to make an edge list
Edgelist = as.data.frame(t(combn(temp$Person,2)))
colnames(Edgelist) = c("P1", "P2")

# add edges
for (i in 1 : nrow(Edgelist)) {
g <- add_edges(g,
c(as.character(Edgelist$P1[i]),as.character(Edgelist$P2[i])))
}

# following line just to check groups are correct, not needed for graph
print(temp)
}

plot(g)

g = simplify(g)

plot(g)
```

10 The following graph shows an alleged insider trading network. Construct the directed graph in igraph and using your analysis identify who, in addition to Raj Rajaratnam, was an important player in the network. (You might want to enter your data into R as a directed graph formula (Slides 64, 65), using a two letter abbreviation for each person)

For more information on the data see:
https://www.valuewalk.com/2015/03/information-networks-evidence-from-illegal-insider-trading-tips/



Tips

```
You might want to use this example from Lecture 5 as a way of creating
the social network:

g <- graph.formula(J-+D, J++R, J+-A, A-+S, A++R, A-+M, S-+R)
```