

Statistical Analysis in Social Networks

Jerico Quintos

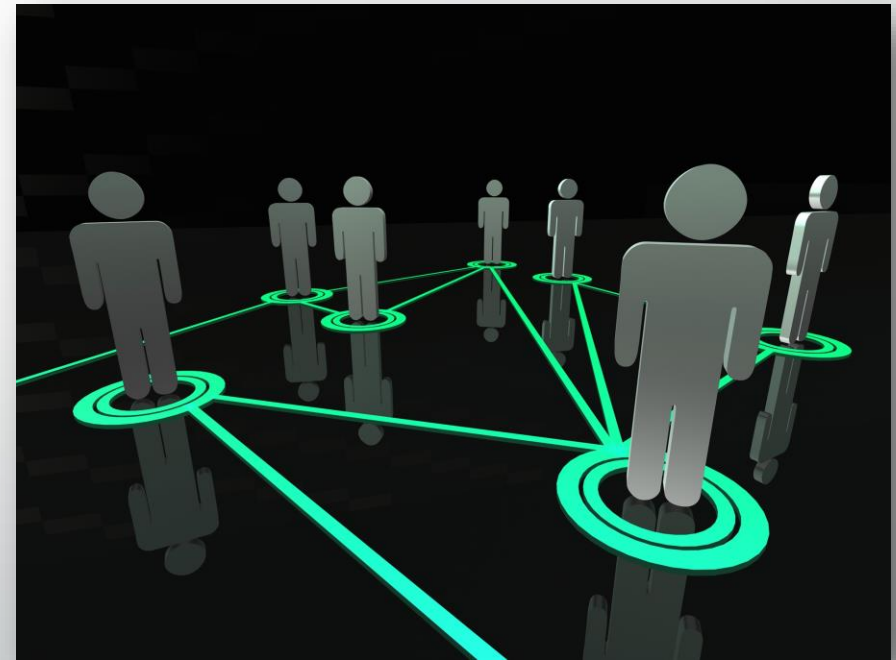


Introduction

- This dissertation is about the statistical analysis of social networks especially with the social media (e.g. Facebook, Twitter, YouTube, Instagram, etc.).
- The main objective of this dissertation is to use a statistical method in which we could determine a representation of the relationships between the individuals or actors.

What is a Social Network?

- A social network is a structure that is made up of actors such as individuals or organizations and a set of paired ties (dyad) between the actors.
- Such networks can be either directed or undirected.
- For this dissertation, we will look at the undirected network.



Dataset

- The dataset that I will be using for this dissertation is my personal Facebook data.
- By doing so, the network will now be egocentric around me i.e. an Ego Network.
- There are 346 actors in the network.



Cons of gathering data

- Difficult to keep up with the daily changes as people constantly update their information.
- Not everyone has complete profile.
- It can be very time consuming.

- An excerpt of data from the whole dataset.
- Rship = Relationship Status
 - S = Single
 - SE = Separated
 - E = Engaged
 - IR = In Relationship
 - M = Married
 - CR = Complicated Relationship
 - CU = Civil Union
 - D = Divorced
 - W = Widowed

Friend	Gender	Age	Rship	Mutual	Total
1	F	22	S	1	52
2	F	21	IR	3	845
3	M	24	IR	3	59
4	F	20	IR	3	263
5	M	18	S	3	408
6	M	18	S	8	212
7	M	16	S	7	399
8	M	17	S	7	142
9	M	17	S	4	362
10	M	22	S	3	4

Visualising Data

- For the past few months, I found out how to visualise the data with the use of these computer programs.
 - Netdraw (<https://sites.google.com/site/netdrawsoftware/>)
 - R (<http://www.r-project.org/>)
- For this dissertation, I would be using **R** as it can provides a wider variety of statistical and graphical techniques than Netdraw.

- Using the previous excerpt we can use **R** to visualise the excerpt.
- But before we do that, we need to set-up **R** and install packages specifically made for network analysis.

Friend	Gender	Age	Rship	Mutual	Total
1	F	22	S	1	52
2	F	21	IR	3	845
3	M	24	IR	3	59
4	F	20	IR	3	263
5	M	18	S	3	408
6	M	18	S	8	212
7	M	16	S	7	399
8	M	17	S	7	142
9	M	17	S	4	362
10	M	22	S	3	4

Setting up R

- The Comprehensive **R** Archive Network (CRAN) (<http://cran.r-project.org/>)
 - An archive of packages for a variety of statistical areas.
- This can be done by writing the code `install.packages()` in **R** e.g. `install.packages("network")` and by loading the packages using `library()`.
- For this dissertation, we will use the packages "network", "statnet" and "igraph".

- Need to tell **R** how the actors are connected through an edge list.
- Number of connections is called *Degree*.

Friend	Connected to
1	7
2	6, 7, 8
3	6, 7, 8
4	5, 6, 7
5	4, 6, 8
6	2, 3, 4, 5, 7, 8, 9, 10
7	1, 2, 3, 4, 6, 8, 9
8	2, 3, 5, 6, 7, 9, 10
9	6, 7, 8, 10
10	6, 8, 9

```

library("igraph")

edgelist <- read.table("friendedge.txt")
attribute <- read.csv("friend.csv")

colnames(edgelist) <- c('ego', 'alter', 'tie')
edge_full <- graph.data.frame(edgelist)
edge_full_sym <- as.undirected(edge_full, mode='collapse')

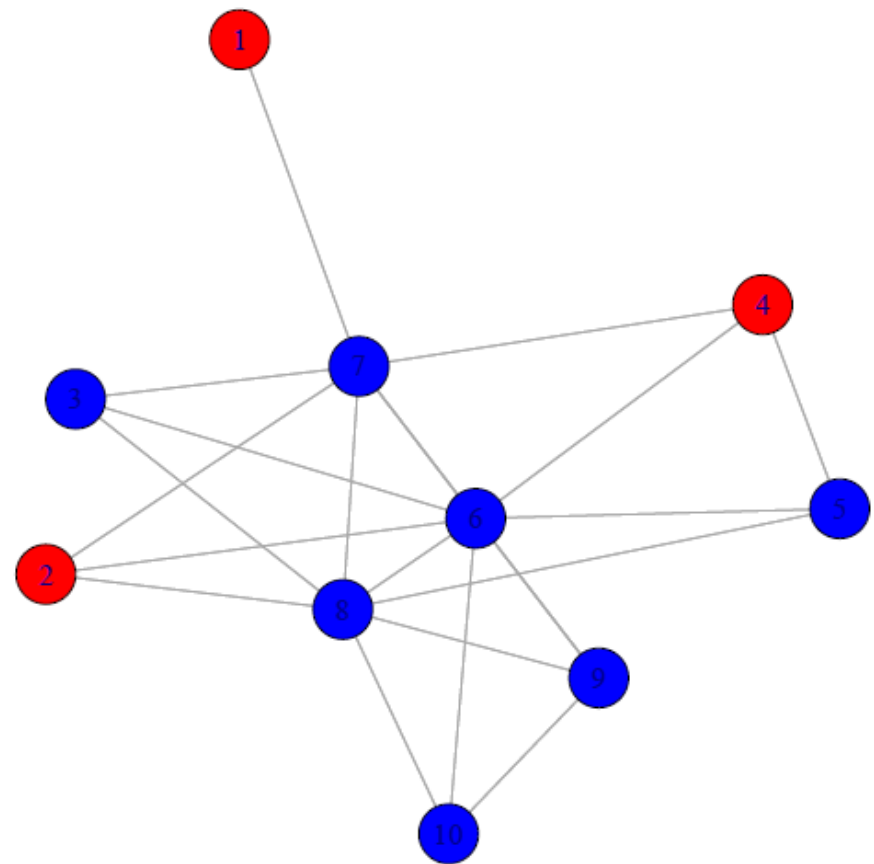
attribute = cbind(1:length(attribute[,1]), attribute)
edge_full <- graph.data.frame(d = edgelist, vertices = attribute)

plot(edge_full)

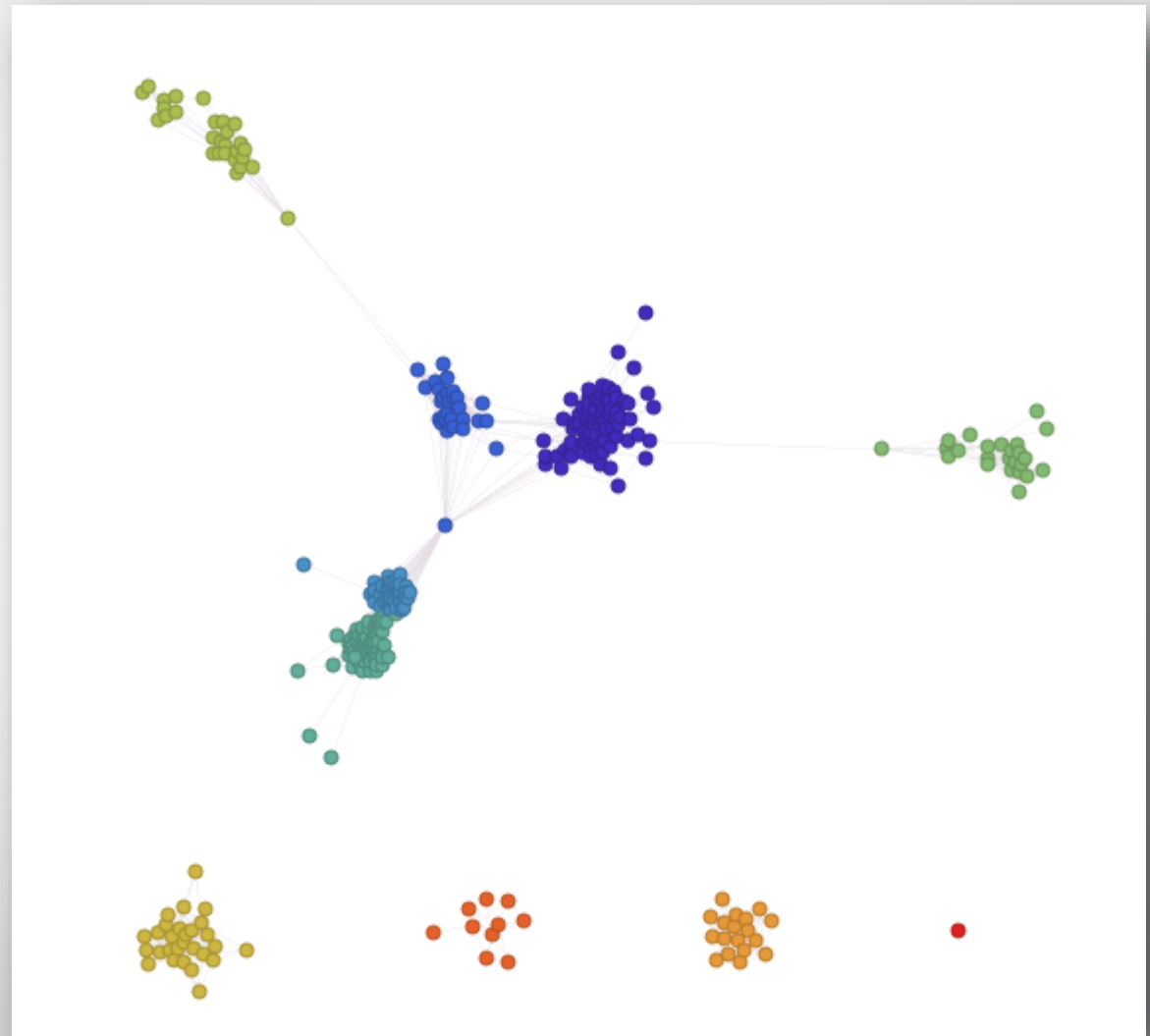
edge_layout <- layout.fruchterman.reingold(edge_full)
plot(edge_full, layout = edge_layout)

edge_colors = get.vertex.attribute(edge_full, "gender")
colors = c('Red', 'Blue')
edge_colors[edge_colors == "F"] = colors[1]
edge_colors[edge_colors == "M"] = colors[2]
plot(edge_full, layout = edge_layout, vertex.color = edge_colors, edge.arrow.size=0)

```



- We are hoping to achieve this diagram by the end of the project.
- This was obtained through the Facebook option from Wolfram Alpha.
(<http://www.wolframalpha.com/>)





Statistical Methods

Applying to **R** using the previous excerpt

Node-Level Statistic

- Computes the in-degree and out-degree for each node (actors) to calculate reachability of each actors.
- Calculates geodesics (shortest path) between each actor pair.
- Since the network is undirected, this is can be skipped.

Network-level Statistic

- Calculate the mean and standard deviation of the degree.
- Calculate the graph density, reciprocity, and transitivity.

```
deg_full_in <- degree(edge_full, mode="in")
mean(deg_full_in)
sd(deg_full_in)

graph.density(edge_full)
reciprocity(edge_full)
transitivity(edge_full)
```

Heterogeneity

- Homophily – property of nodes that has similar attributes. Heterogeneity is the opposite.
- Using a statistic called IQV or I Index of Qualitative Variation.
- Only works with categorical variables that have been numerically coded to integer values that ascend sequentially from 0.

```
get_iqvs <- function(graph, attribute) {  
  mat <- get.adjacency(graph)  
  attr_levels = get.vertex.attribute(graph,  
                                     attribute,  
                                     V(graph))  
  
  num_levels = length(unique(attr_levels))  
  iqvs = rep(0, nrow(mat))  
  for (ego in 1:nrow(mat)) {  
  
    alter_attr_counts = rep(0, num_levels)  
    num_alters_this_ego = 0  
    sq_fraction_sum = 0  
    for (alter in 1:ncol(mat)) {  
      if (mat[ego, alter] == 1) {  
  
        num_alters_this_ego = num_alters_this_ego + 1  
        alter_attr = get.vertex.attribute(graph,  
                                           attribute, (alter - 1))  
        alter_attr_counts[alter_attr + 1] =  
          alter_attr_counts[alter_attr + 1] + 1  
      }  
    }  
    for (i in 1:num_levels) {  
      attr_fraction = alter_attr_counts[i] /  
        num_alters_this_ego  
      sq_fraction_sum = sq_fraction_sum + attr_fraction ^ 2  
    }  
    blau_index = 1 - sq_fraction_sum  
    iqvs[ego] = blau_index / (1 - (1 / num_levels))  
  }  
  return(iqvs)  
}  
edge_iqvs <- get_iqvs(edge_full, 'gender')
```


What's next?

- We will research more into several topics as follows:
 - Centralities, Structural equivalences, Block modelling, QAP regression, Exponential-Family random graph Models.

References

Books

KOLACZYK, Eric D. *Statistical Analysis of Network Data: Methods and Models*. 2009 Edition. New York: Springer, 2009.

SCOTT, John and Peter J. CARRINGTON, ed. *The SAGE Handbook of Social Network Analysis*. London: Sage, 2011.

Websites

Social Network Analysis in R.

MCFARLAND, Daniel A. et al., 2010 [viewed 15 Jan 2014]. Available from:

<http://sna.stanford.edu/rlabs.php>.

Introduction to social network methods.

HANNEMAN, Robert A. and Mark RIDDLE, 2005. [viewed 13 Nov 2013]. Available from:

<http://faculty.ucr.edu/~hanneman/nettext/>.