CS595 Intro to Web Science, Assignment #6

Valentina Neblitt-Jones

October 31, 2013

Question 1

We know the result of the Karate Club (Zachary, 1977) split. Prove or disprove that the result of split could have been predicted by the weighted graph of social interactions. How well does the mathematical model represent reality?

Generously support your answer with all supporting equations, code, graphs, arguments, etc.

Useful sources include:

- Original paper
  - `http://aris.ss.uci.edu/~lin/76.pdf`
- Slides
  - `http://www-personal.umich.edu/~ladamic/courses/networks/si614w06/ppt/lecture18.ppt`
  - `http://clair.si.umich.edu/si767/papers/Week03/Community/CommunityDetection.pptx`
- Code and data
  - `http://networkx.github.io/documentation/latest/examples/graph/karate_club.html`
  - `http://nbviewer.ipython.org/url/courses.cit.cornell.edu/info6010/resources/11notes.ipynb`
  - `http://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989`
  - `http://stackoverflow.com/questions/5822265/are-there-implementations-of-algorithms-for-community-detection-in-graphs`
  - `http://konect.uni-koblenz.de/networks/ucidata-zachary`
  - `http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#zachary`

Answer to Question 1

It was difficult to decide on approach here since we were provided information both on R and Python. Initially I thought Python followed by R for the graphs which appear to be expected due to the mention of graphs in the question. I started trying to follow the example and it was clear my computer was missing some important software. However, attempt to rectify that failed in more areas than one. After Corren McCoy raised the issue of getting the data set into R, it seemed that R was a possibility for answering this question. Scott Ainsworth's earlier lecture on R did reveal much more capability than we had employed on the previous assignment which had Python doing most of the heavy lifting and a smaller amount of work using R to create the graphs. Although Shawn Jones and I have had different approaches to previous problems, we were in the same boat this time as far as doubts about tools to use for this problem. We both experimented

with edge.betweenness.community, fastgreedy.community, walktrap.community, spinglass.community, leading.eigenvector.community, label.propagation.community and I even tried infomap. We were not finding what we expected. Then Corren provided her steps in using R to the class email list.

First I imported the karate club data. Then I assigned it to "k" and plotted the graph. Then I calculated the edge betweenness of k. Next I ordered the result in decreasing order. The next thing to figure out was how to reference an item in order to remove it. I did a search for this and shared a Stack Overflow page with Shawn. The first answer was not quite right, but he found something of value lower on the answer page. Negative indices means do not include the element. So we used -1 to pop the element off the front. Next "get.edge" expects the graph name (k) and the id (a) and returns the end points of the edge with the edge id supplied and "delete.edges" removes the specified edge from the graph and preserves the vertices. P is used to select edges based on their end points. Then I plotted the graph again in order to track the graph changes. This repeats until the number of clusters is equal to two. (Listing 1)

Listing 1: karateClub.R

```
data(karate)
k <- karate
plot.igraph(k) #Original graph

repeat
        {
        kedge <- edge.betweenness(k)
        korder <- order(kedge, decreasing=TRUE)
        a <- korder[-1]
        b <- get.edge(k,a)
        k <- delete.edges(k, E(k,P=b))
        plot.igraph(k)
        if (clusters(k)$no == 2) break()
        }
```

Values	
a	integer[60]
b	numeric[2]
k	igraph[9]
karate	igraph[9]
kedge	numeric[61]
korder	integer[61]

Figure 1: Results of Assignments

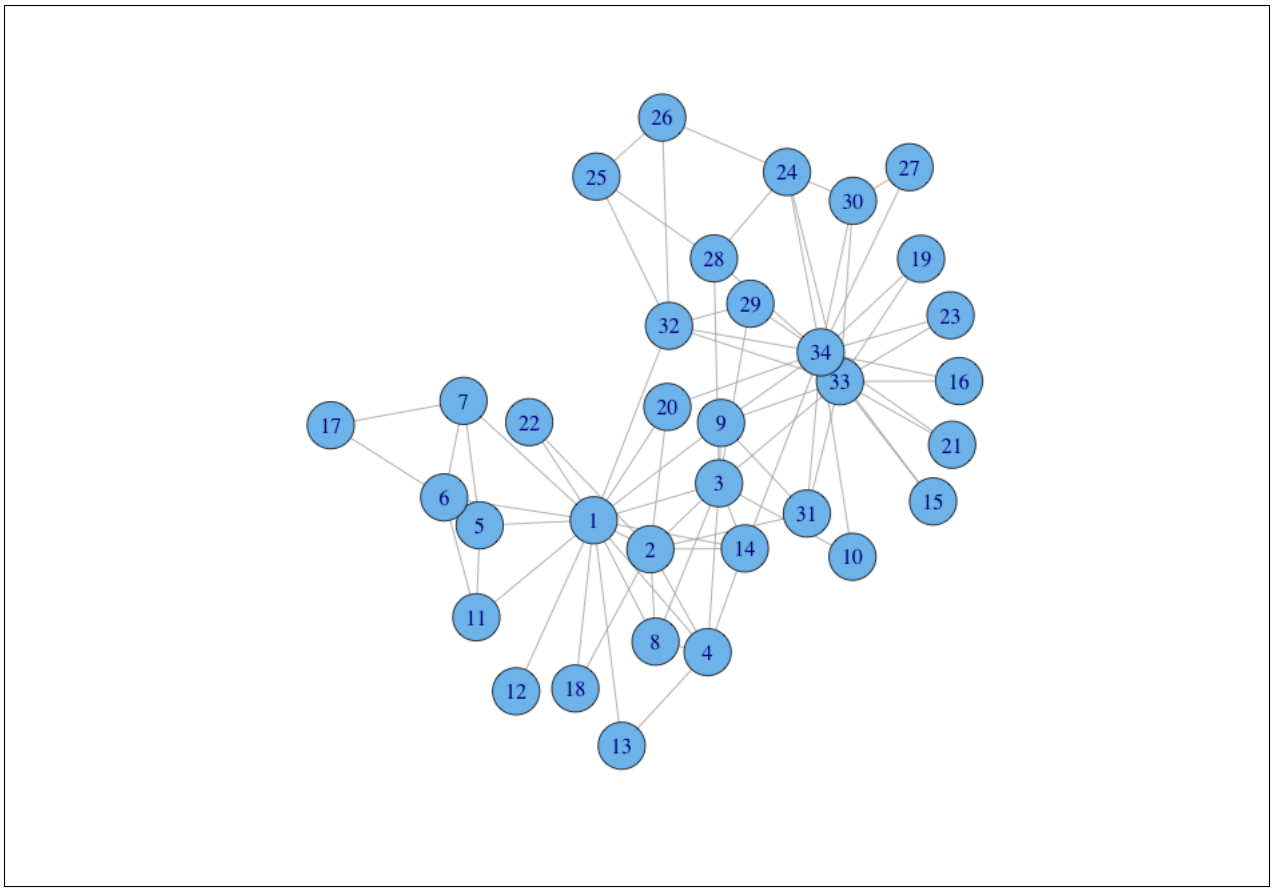

Figure 2: Before Community Split

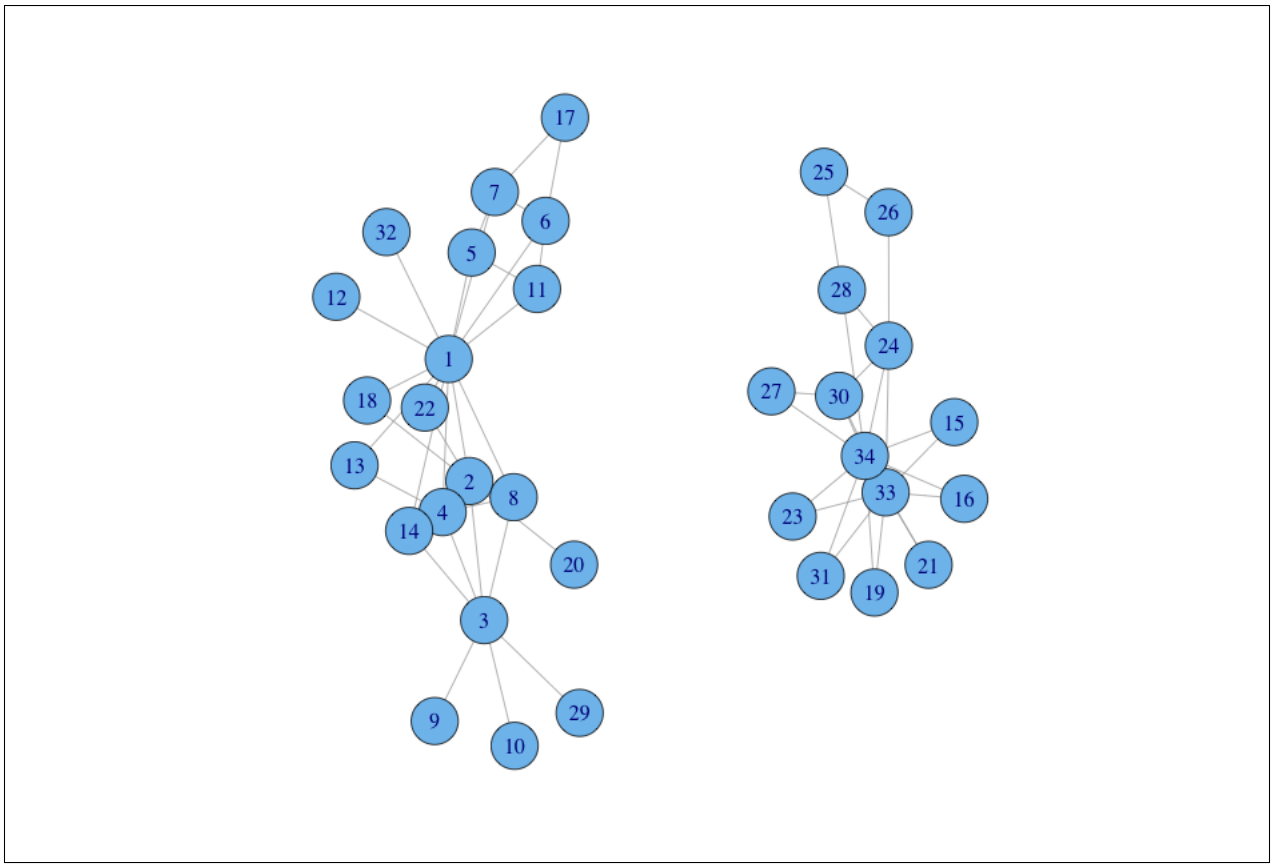

Figure 3: After Community Split

The resulting graph (Figure 3) supplies the information for the Model column in Table 1 and reveals that there are 32 hits, 2 misses or 94% hits, 6% misses. This is close to the paper's results - 33 hits, 1 miss or 97% hits, 3% misses. So the model is close to representing reality.

Identifier	Model	Actual	Hit/Miss
1	Mr. Hi	Mr. Hi	Hit
2	Mr. Hi	Mr. Hi	Hit
3	Mr. Hi	Mr. Hi	Hit
4	Mr. Hi	Mr. Hi	Hit
5	Mr. Hi	Mr. Hi	Hit
6	Mr. Hi	Mr. Hi	Hit
7	Mr. Hi	Mr. Hi	Hit
8	Mr. Hi	Mr. Hi	Hit
9	Mr. Hi	Mr. Hi	Hit
10	Mr. Hi	John	Miss
11	Mr. Hi	Mr. Hi	Hit
12	Mr. Hi	Mr. Hi	Hit
13	Mr. Hi	Mr. Hi	Hit
14	Mr. Hi	Mr. Hi	Hit
15	John	John	Hit
16	John	John	Hit
17	Mr. Hi	Mr. Hi	Hit
18	Mr. Hi	Mr. Hi	Hit
19	John	John	Hit
20	Mr. Hi	Mr. Hi	Hit
21	John	John	Hit
22	Mr. Hi	Mr. Hi	Hit
23	John	John	Hit
24	John	John	Hit
25	John	John	Hit
26	John	John	Hit
27	John	John	Hit
28	John	John	Hit
29	John	John	Hit
30	John	John	Hit
31	John	John	Hit
32	Mr. Hi	John	Miss
33	John	John	Hit
34	John	John	Hit

Table 1: Results of Model vs. Actual

Extra Credit, 3 Points

We know the group split into two different groups. Suppose the disagreements in the group were more nuanced – what would the clubs look like if they split into groups of 3, 4, and 5?

Answer to Extra Credit

Working with the assumption that the solution to the regular question is correct, I just changed the cluster number in Listing 1 to match 3, 4 and 5 separately. With more factions the larger clusters still belong to Mr. Hi and John. Clusters not including them are quite small. In all three cases, factions without Mr. Hi or John are of size four. In all three cases, factions containing Mr. Hi are bigger than those containing John. Table 2 shows the counts for each clusters. Figure 4 has one size 4 cluster. Figure 5 has two size 4 clusters. Figure 6 has three size 4 clusters.

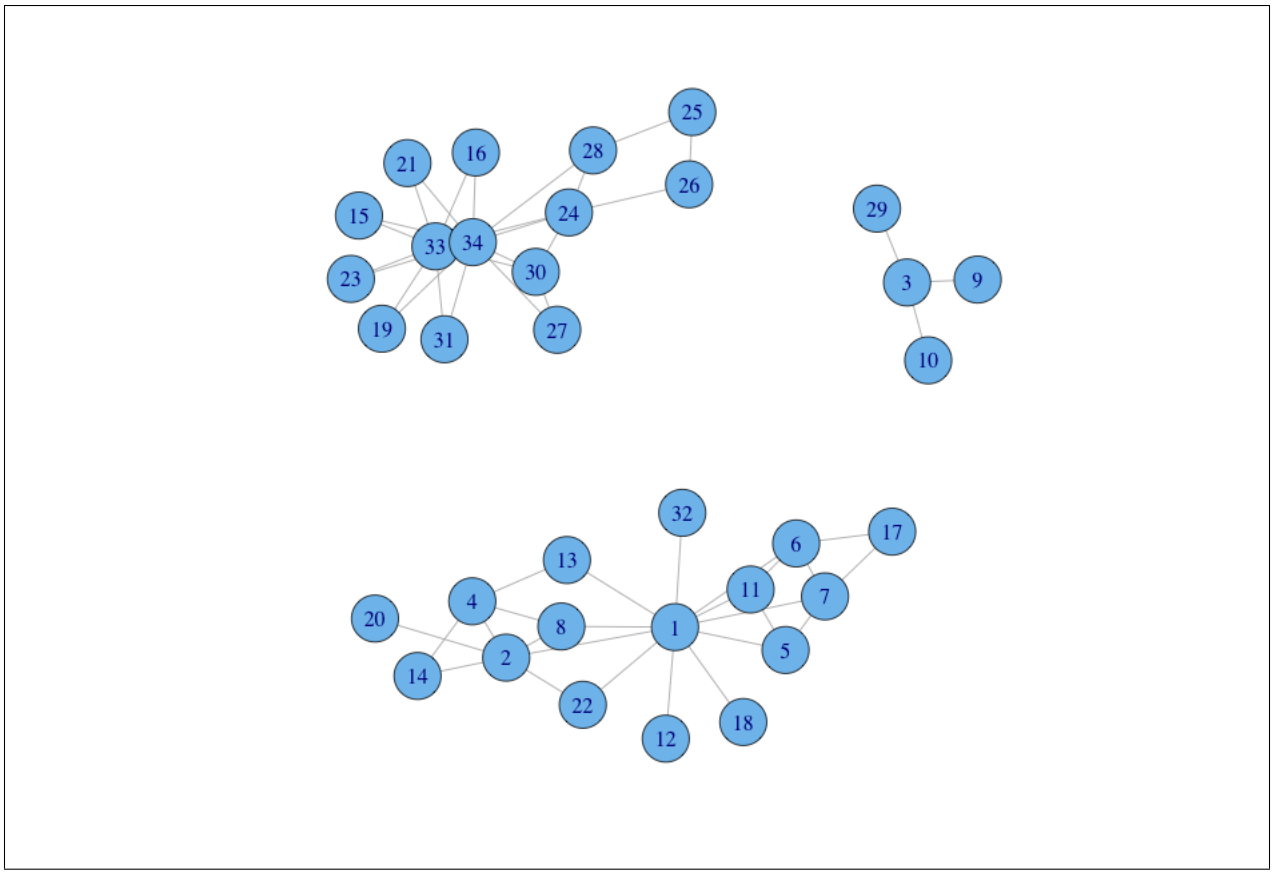

Figure 4: Karate Club with Three Factions

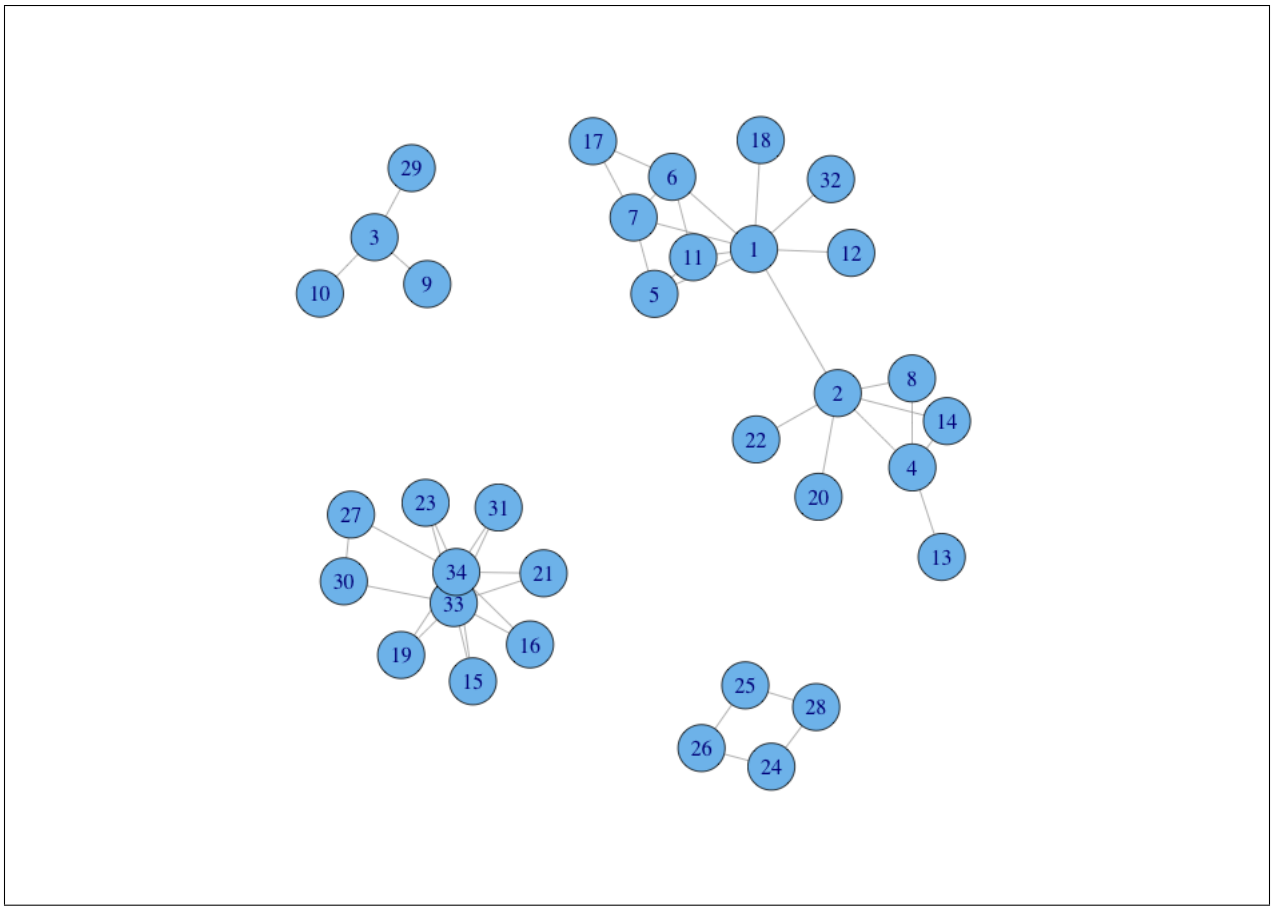

Figure 5: Karate Club with Four Factions

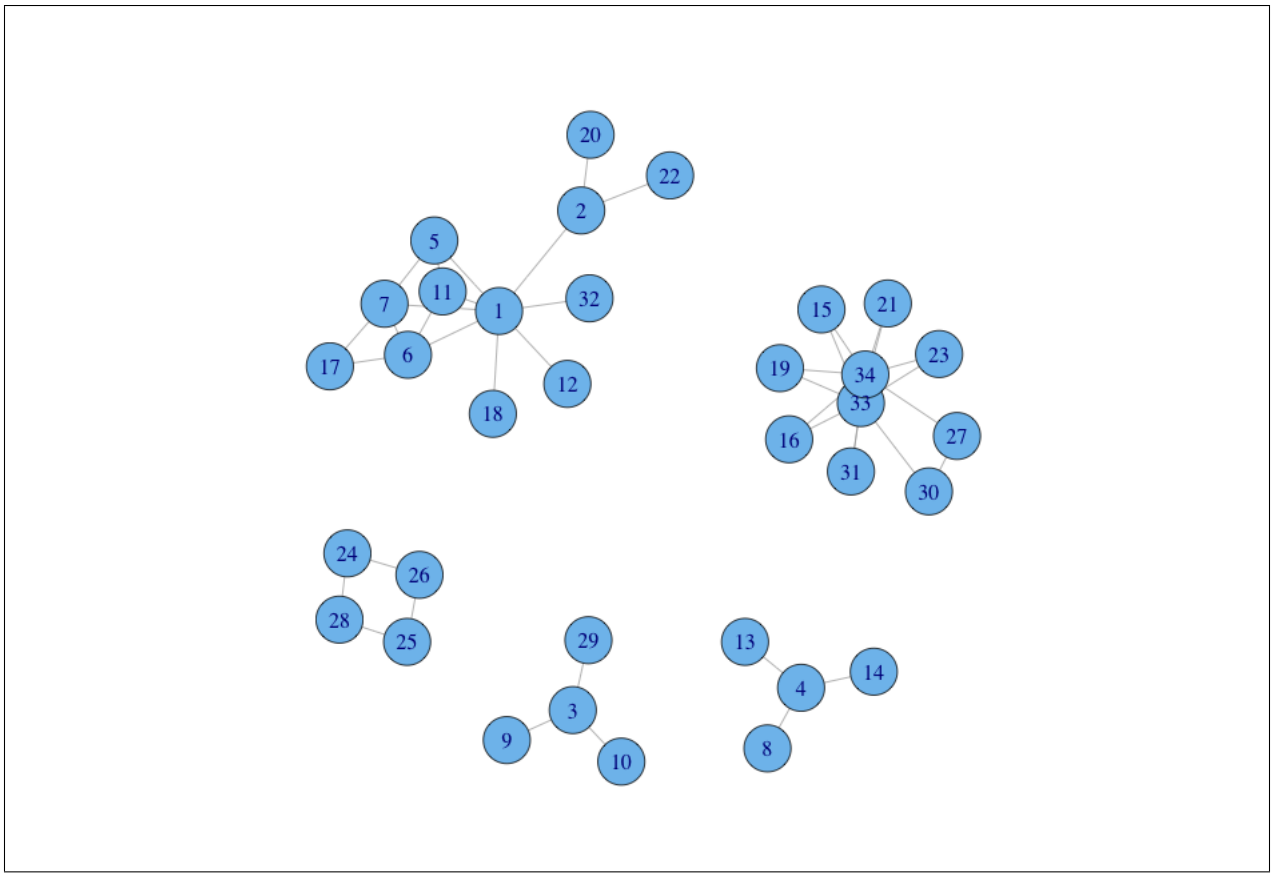

Figure 6: Karate Club with Five Factions

Clusters	Mr. Hi	John	Others
3	16	14	4
4	16	10	4
5	12	10	4

Table 2: Comparison of Three Different Cluster Sizes

Resources

- Csardi, Gabor. Gaining information about graph structure. http://igraph.sourceforge.net/doc/R/structure.info.html
- Csardi, Gabor. Method for structural manipulation of graphs. http://igraph.sourceforge.net/doc/R/graph.structure.html
- Csardi, Gabor. Network Analysis and Visualization. http://igraph.sourceforge.net/doc/R/00Index.html
- Csardi, Gabor. Network Analysis with igraph. http://igraph.sourceforge.net/igraphbook/index.html
- Csardi, Gabor. Vertex and edge sequences and iterators. http://igraph.sourceforge.net/doc/R/iterators.html
- Poulson, Barton. R Statistics Essential Training. http://www.lynda.com/course20/R-tutorials/R-Statistics-Essential-Training/142447-2.html
- Rice, Ken & Lumley Thomas. Writing Loops. http://faculty.washington.edu/kenrice/sisg/SISG-08-05.pdf
- Stack Overflow. Are there implentations of algorithms for community detection in graphs? http://stackoverflow.com/questions/5822265/are-there-implementations-of-algorithms-for-community-detection-in-graphs
- Stack Overflow. How can I remove an element from a list http://stackoverflow.com/questions/9998667/r-syntax-for-selecting-all-but-two-first-rows-in
- Stack Overflow. What are the differences between community detection algorithms in igraph? http://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989
- Zachary, Wayne. An Information Flow Model for Conflict and Fission in Small Groups. http://aris.ss.uci.edu/~lin/76.pdf