CS896 Introduction to Web Science
Fall 2013
Report for Assignment 6

Corren G. McCoy

October 31, 2013

Contents

List of Figures

List of Tables

1 Question 1

1.1 Problem

We know the result of the Karate Club (Zachary, 1977)[4] 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 document your answer with all supporting equations, code, graphs, arguments, etc.

1.2 Response

We will proceed with a direct proof to determine whether a logical sequences of steps can reasonably generate the desired community structure. We are given the final network structure shown in Figure 1 which identifies membership in the two clubs after the division by the shaded and unshaded nodes. Even though the two groups are very interconnected, we must determine whether the structure itself provides enough information to predict the division between the group members.

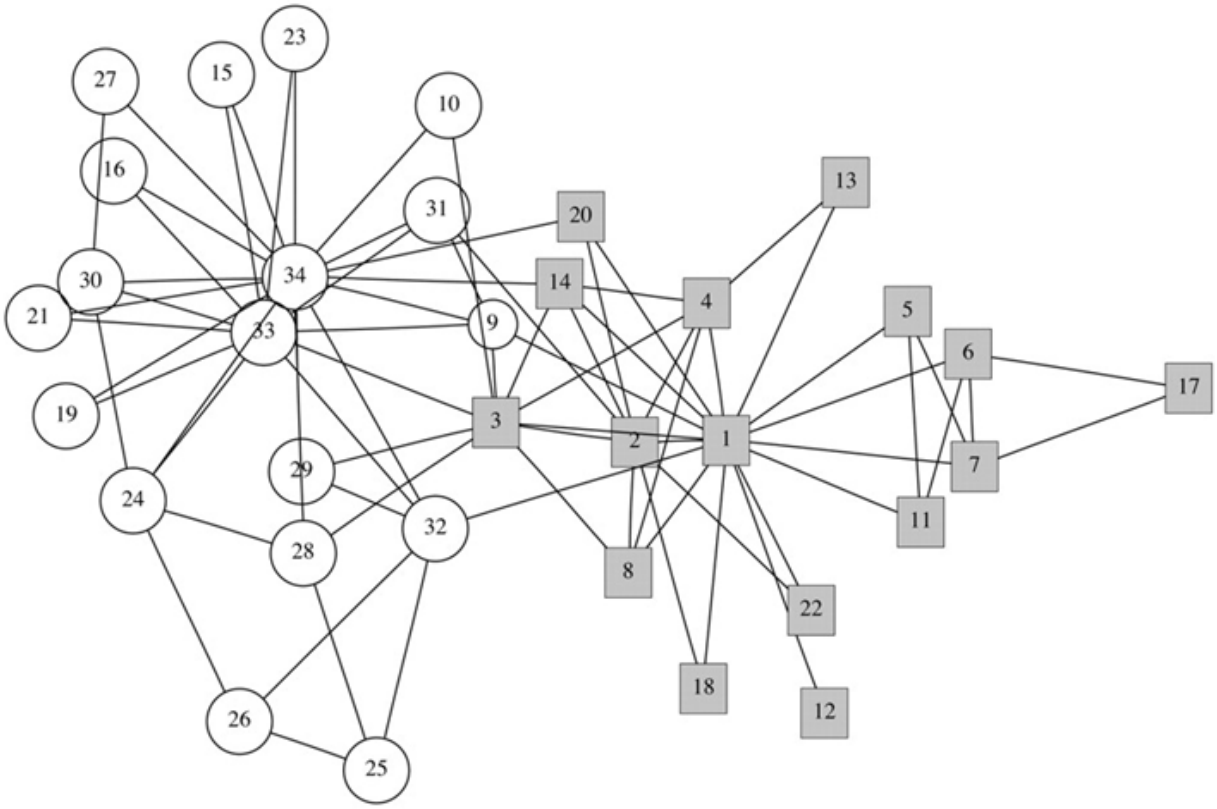

Figure 1: Zachary Karate Club

1.3 Methodology

To properly identify the division in the karate club, we need to carefully examine the edges between the groups. For the analysis of the graph partitions, we will employ a divisive method proposed by Girvan and Newman [2]. This method defines an abstract concept of traffic on the network and looks for the edges that carry the most of this traffic. For each pair of nodes A and B in the graph that

are connected, we calculate the flow along the edges. Girvan-Newman defines the betweenness of an edge to be the total amount of flow it carries when considering all pairs of nodes using this edge. Now, based on the premise that the nodes with high traffic are the most vital edges for connecting different regions of the network, the algorithm tries to remove these nodes first. The basic idea of this method is that two communities should be connected by edges that operate as bridges, through which the traffic should be relatively high. Therefore, by ranking all edges according to their betweenness and then removing the edges with the largest value, we should obtain community partitions. By iteratively recalculating betweenness, we can implement a divisive algorithm for detecting community structure in the undirected graph. This concept is the underlying premise of the Girvan-Newman method which is summarized using the following steps:

1. First, calculate the betweenness of all existing edges in the network. For any given node, compute the total flow along the shortest path from that node to all others.
2. The edge with the highest betweenness is removed. If there is a tie, remove both edges from the graph.
3. The betweenness of all edges affected by the removal is recalculated.
4. Repeat steps 2 and 3 until either no edges remain or the desired number of communities is obtained.

To apply the algorithm, we used the weighted graph representation of the karate club which is available in the igraphdata[1] package in "R". The data in this package is the same as the quantified matrix of social interactions defined in Zachary's[4] Figure 3. For reference, the same figure is included here in Appendix B. To apply the Girvan-Newman method, we implemented a function in "R" which performs the following tasks:

- Accepts a parameter for the desired number of communities. The default is two;
- Loads the karate club data into a graph object;
- Applies the edge.betweenness.community[2] algorithm to the graph object;
- Plots the graph object and highlights the top three candidates for removal;
- Removes the edge with the highest value for betweenness;
- Calculates modularity on the resulting network; and
- Continues processing until either all edges are removed or the desired number of communities is obtained.

The source code for the function is shown in Appendix A. The associated plot for each iteration of the algorithm is shown in Table 1. When viewed from left to right, we see the edge with the highest level of betweenness is highlighted in red. The subsequent plot shows the impact of this edge's removal on the graph and the recalculation of betweenness. The final plot in the series shows a definitive separation of the graph into two distinct communities.

Table 2 shows a comparison of the nodes in each cluster using Girvan-Newman to the known structure in Figure 1. In the graph, node 1 is the instructor, Mr. Hi, while node 34 is the student officer. Three discrepancies in the assignments are noted with members 3, 9 and 14. The graph theory cannot predict subjective behavior or the affect of outside influences that cannot be measured. In this case, we know that the student represented by node 9 joined the instructor's club because he was only

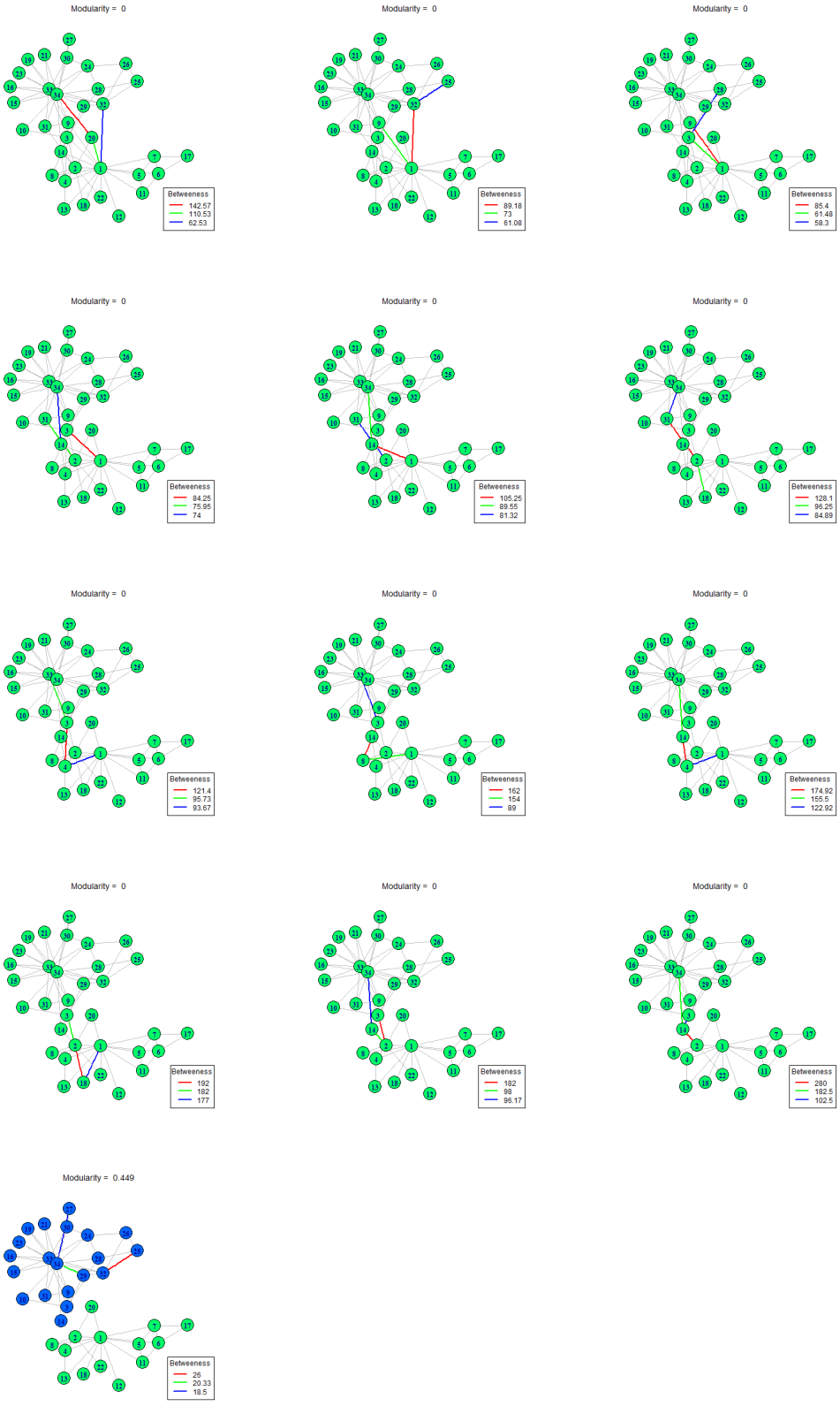

Table 1: Girvan-Newman Community Detection

a few weeks away from obtaining his black belt. To join the officer's club, as suggested by the graph, would have meant starting over; thereby forfeiting three years of training. No exceptional circumstances were noted for members 3 and 14. Despite these small discrepancies, the Girvan-Newman algorithm achieved a 91% hit ratio in assigning members to their respective community.

As a final step, we would like to also apply an objective measure of the quality of the community divisions we obtained. Newman [3] defines modularity (Q) as one "measure of the quality of a particular division of a network." In his analysis of complex networks, Arenas [1] explains further that "the modularity of a given partition is the probability of having edges falling within modules in the network minus the expected probability in an equivalent (random) network with the same number of nodes." Basically, we want to determine if the model performed better than a random assignment. A larger value for modularity indicates a strong community structure. While modularity can have a range of -1 to +1, Newman [3] indicates that Q values "typically fall in the range from about 0.3 to 0.7. Higher values are rare." The modularity function which we invoked in "R" is defined by the following mathematical equation where e is an edge and a is an element of the matrix. Our final plot in Table 1 shows a maximized Q value of 0.449 which when combined with our high hit ratio does indicate that the mathematical model is a reasonable representation of the actual karate club split.

$$Q = \sum (e_{ij} - a_i^2) \tag{1}$$

2 Question 2 Extra Credit

2.1 Problem

We know the group split in 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?

2.2 Response

For this question, we used the same methodology and "R" function as stated in question 1. We passed a parameter to the "R" function to indicate the desired number of communities. The results are shown in Figures 2, 3 and 4. A strong community structure is evidenced by the high modularity values of 0.567, 0.597 and 0.658 respectively,

[1] `http://cran.r-project.org/web/packages/igraphdata/igraphdata.pdf`
[2] `http://www.inside-r.org/packages/cran/igraph/docs/edge.betweenness.community`

Member #	Zachary's After Split	Our Model	Hit / Miss
1	Mr. Hi	Mr. Hi	Hit
2	Mr. Hi	Mr. Hi	Hit
3	Mr. Hi	Officer	Miss*
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	Officer	Miss*
10	Officer	Officer	Hit
11	Mr. Hi	Mr. Hi	Hit
12	Mr. Hi	Mr. Hi	Hit
13	Mr. Hi	Mr. Hi	Hit
14	Mr. Hi	Officer	Miss*
15	Officer	Officer	Hit
16	Officer	Officer	Hit
17	Mr. Hi	Mr. Hi	Hit
18	Mr. Hi	Mr. Hi	Hit
19	Officer	Officer	Hit
20	Mr. Hi	Mr. Hi	Hit
21	Officer	Officer	Hit
22	Mr. Hi	Mr. Hi	Hit
23	Officer	Officer	Hit
24	Officer	Officer	Hit
25	Officer	Officer	Hit
26	Officer	Officer	Hit
27	Officer	Officer	Hit
28	Officer	Officer	Hit
29	Officer	Officer	Hit
30	Officer	Officer	Hit
31	Officer	Officer	Hit
32	Officer	Officer	Hit
33	Officer	Officer	Hit
34	Officer	Officer	Hit
Total (Actual)			31 Hits 3 Misses
Total (Percentage)			91.2% Hits 8.8% Misses

Table 2: Evaluation

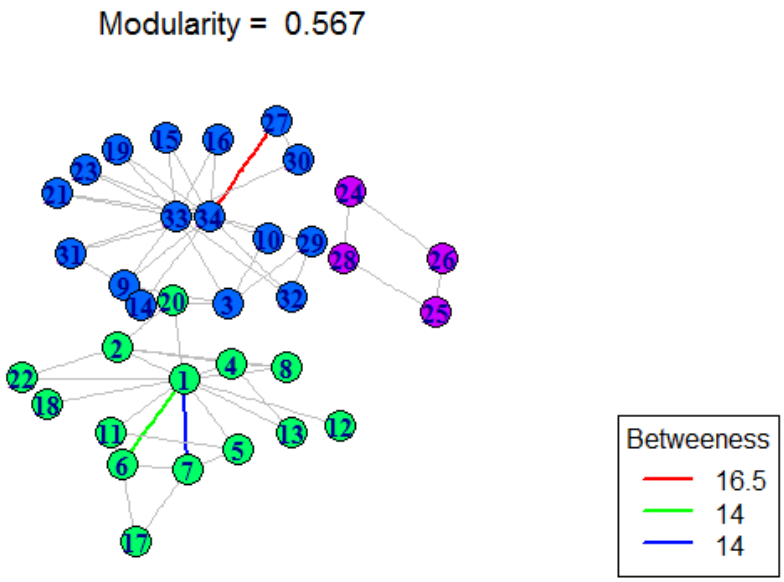

Figure 2: Karate Club - Three Groups

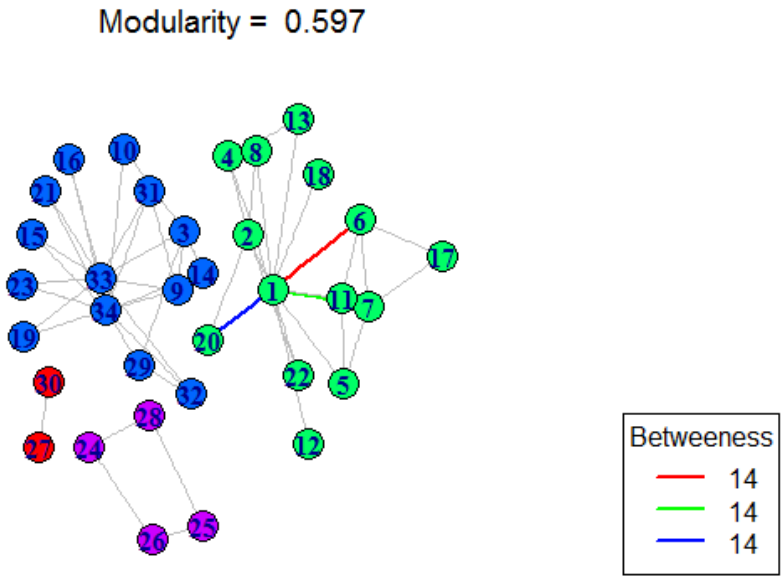

Figure 3: Karate Club - Four Groups

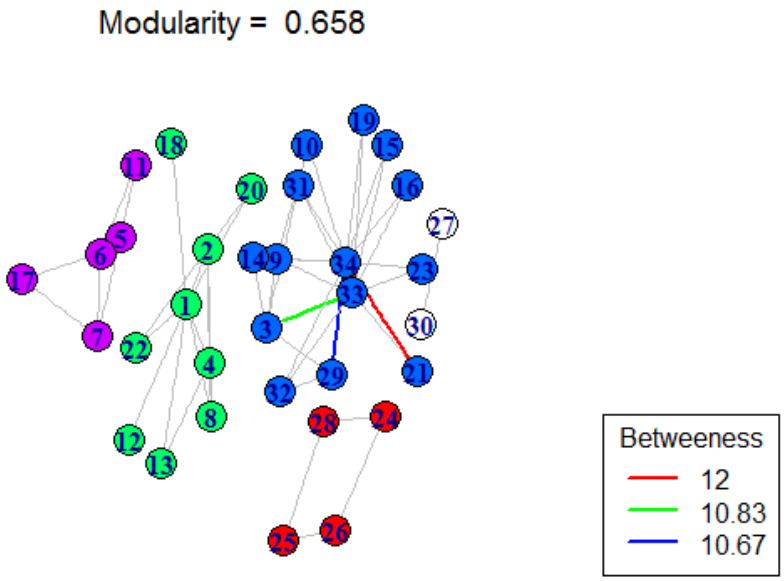

Figure 4: Karate Club - Five Groups

Bibliography

[1] A. Arenas, A. Fernandez, and S. Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, 2008.

[2] M. Girvan and M. E. Newman. Community structure in social and biological networks. *arXiv preprint cond-mat/0112110*, 2001.

[3] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[4] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

Appendix A

"R" Source Code

```
karateClub <- function(communities=2) {

  # Pass a parameter to identify the desired number of communities.
  # Default to 2 which is the initial split between club members.
  setwd("C:/Users/Corren/Documents/My R Scripts/Assignment 06")

  library(igraph)
  library(igraphdata)

  # The igraphdata package contains the karate club data set along with
  # the edge weights which identify the common activities between
  # club members.

  # Load the dataset into the "R" workspace.
  data("karate",package="igraphdata")
  # Save the package data as an igraph object.
  graphk <- karate

  # Alternative dataset from UCI repository
  #graphk<- read.graph("karate.paj", format="pajek")

  # How many vertices are in the graph? Use this number to
  # control the iterations of the algorithm.
  vertices <- length(V(karate))

  # Using a force-based layout with a large number of iterations (recommended)
  l <- layout.kamada.kawai(graphk, niter=1000)
  #l <- layout.fruchterman.reingold(graphk, niter=1000)
  # Examine the shortest path between all edges
  ebc <- edge.betweenness.community(graphk, weights=E(graphk)$weight)

  # Set up the color palette for the plot (vertices)
  colbar  <- rainbow(6, start=.3, end=.7)
  # Set up the color palette for the calculations (EBC).
  # Display the first ones in color, remaining are black.
  colbar2 <- c(rainbow(3, start = 0.0, end = 0.66), rep("black", 31))
```

```
# Trim off excess margin space (bottom, left, top, right)
#par(mar=c(0.5, 0.5, 0.5, 0.5), bg="white")

# Either iterate over all the vertices or
# stop when we have the desired number of communities.
numClusters <- no.clusters(graphk)
i <- 1
while (i <= vertices && numClusters != communities) {
  # Girvan-Newman: iteratively delete edges
  g2 <- delete.edges(graphk, ebc$removed.edges[seq(length=i-1)])
  # Recalculate the flow between edges
  eb <- edge.betweenness(g2)

  # Node membership in the clusters
  cl <- clusters(g2)$membership
  # Number of Clusters
  numClusters<- no.clusters(g2)
  # Debug:
  print (numClusters)

  q <- modularity(g2, cl)

  E(g2)$color <- "grey"
  # Sort the edges by betweenness. Highlight the top 3
  E(g2)[order(eb, decreasing=TRUE)[1:3] ]$color <- colbar2[1:3]

  # default edge width
  E(g2)$width <- 1
  # Emphasize the candidate edges we might delete
  E(g2)[ color != "grey" ]$width <- 2

  # Save each iteration to a file for our report
  png(sprintf("karateClub-community-%04d.png", i))

  plot(g2, layout=l, vertex.size=15,  vertex.label.font=2,
    edge.label.color="red", vertex.color=colbar[cl+2],
    edge.label.font=2)

  # Add a heading
  title(main=paste("Modularity = ", round(q,3)), font.main=1)

  # Vector for the Y coordinate
  yCoords <- seq(1,by=-strheight("1")*1.5, length=3)
  # Display the top EBC calculation. Color of text matches the
  # corresponding edge in the graph

  #text(-1.5, yCoords, adj=c(0,0.5), round(sort(eb, dec=TRUE)[1:3],2), col=colbar2,
  #  font=2)
```

```
    # Legend for the edge weights
    legend("bottomright",
           legend=c(round(sort(eb, dec=TRUE)[1:3],2)),
           col=c(colbar2), lwd=2, lty=c(1,1,1),
           title="Betweeness",
           horiz=FALSE
           )

    dev.off()
# Pause so we can review each plot as it changes
    if (interactive()) Sys.sleep(4)
    # increment the loop counter
    i<- i + 1
  } # end of loop
} # end function
```

Appendix B

Karate Club Quantified Matrix

FIGURE 3

QUANTIFIED MATRIX OF RELATIVE STRENGTHS OF THE RELATIONSHIPS IN THE KARATE CLUB: THE MATRIX C

Individual Number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	4	5	3	3	3	3	2	2	0	2	3	2	3	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	2	0	0	0	0
2	4	0	6	3	0	0	0	4	0	0	0	0	0	5	0	0	0	1	0	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0
3	5	6	0	3	0	0	0	4	5	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	3	0
4	3	3	3	0	0	0	0	3	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	3	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	5	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	3	0	0	0	2	5	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	2	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	3
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
11	2	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	3	5	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4
17	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
20	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
22	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
28	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
29	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2
31	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
32	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
33	0	0	2	0	0	0	0	0	3	0	0	0	0	0	3	3	0	0	1	0	3	0	2	5	0	0	0	0	0	4	3	4	0	5
34	0	0	0	0	0	0	0	0	4	2	0	0	0	3	2	4	0	0	2	1	1	0	3	4	0	0	2	4	2	2	3	4	5	0