Word Adjacency Analysis

Michael Polonio

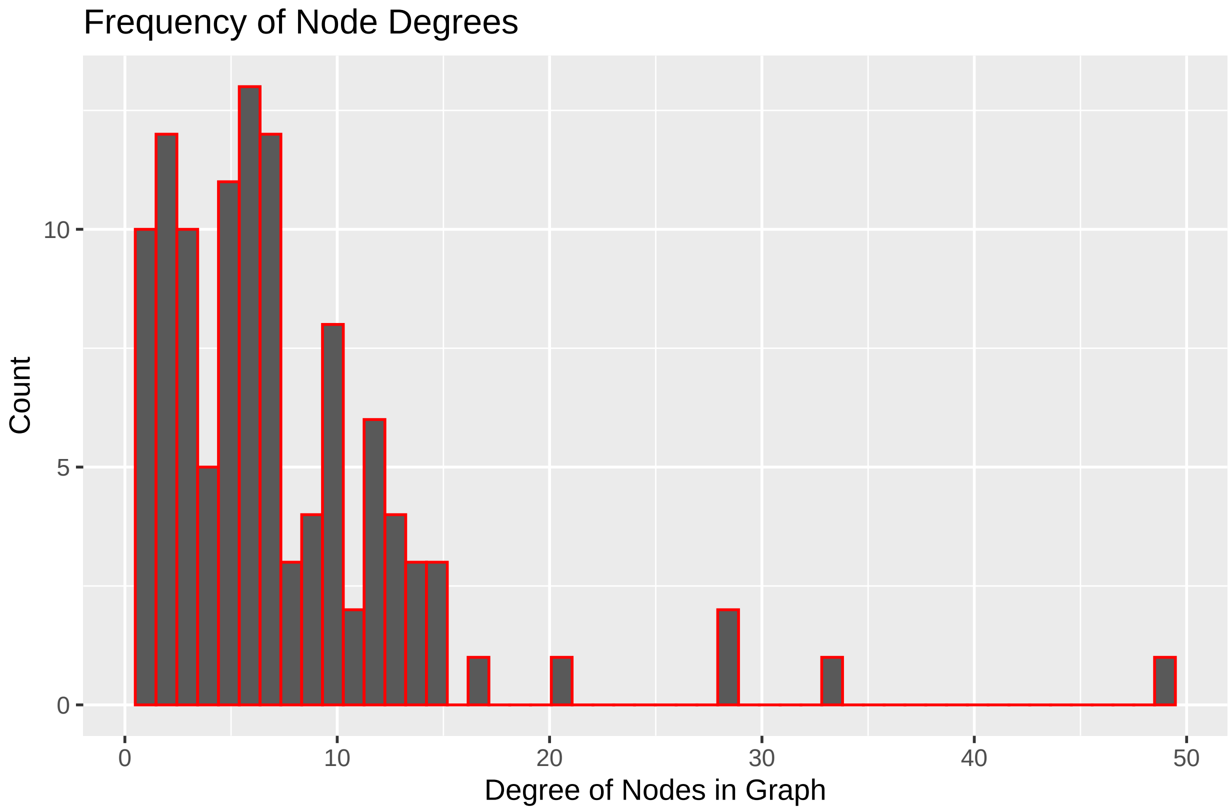
Introduction

Data visualization can provide us with a deeper understanding of the relationships that exist between words. The main objective of this paper is to explore and visualize data about the network of common adjective and noun adjacencies for the novel “David Copperfield” by Charles Dickens, as described by M. Newman. Nodes represent the most commonly occurring adjectives and nouns in the book and the edges between them indicate they appear consecutively in the novel.

Methods

The igraph package is used in this analysis. Node values are 0 for adjectives and 1 for nouns. Edges connect any pair of words that occur in adjacent positions in the text of the book. The dataset contains 112 nodes and each node has three attributes: id, label, and value (indicating if the word is a noun or an adjective). This analysis uses methods such as ggplot to show statistics, splitting the graph into subgraphs based on the type of word (noun/adjective), graphing based on betweenness, and outlining communities within the graph.

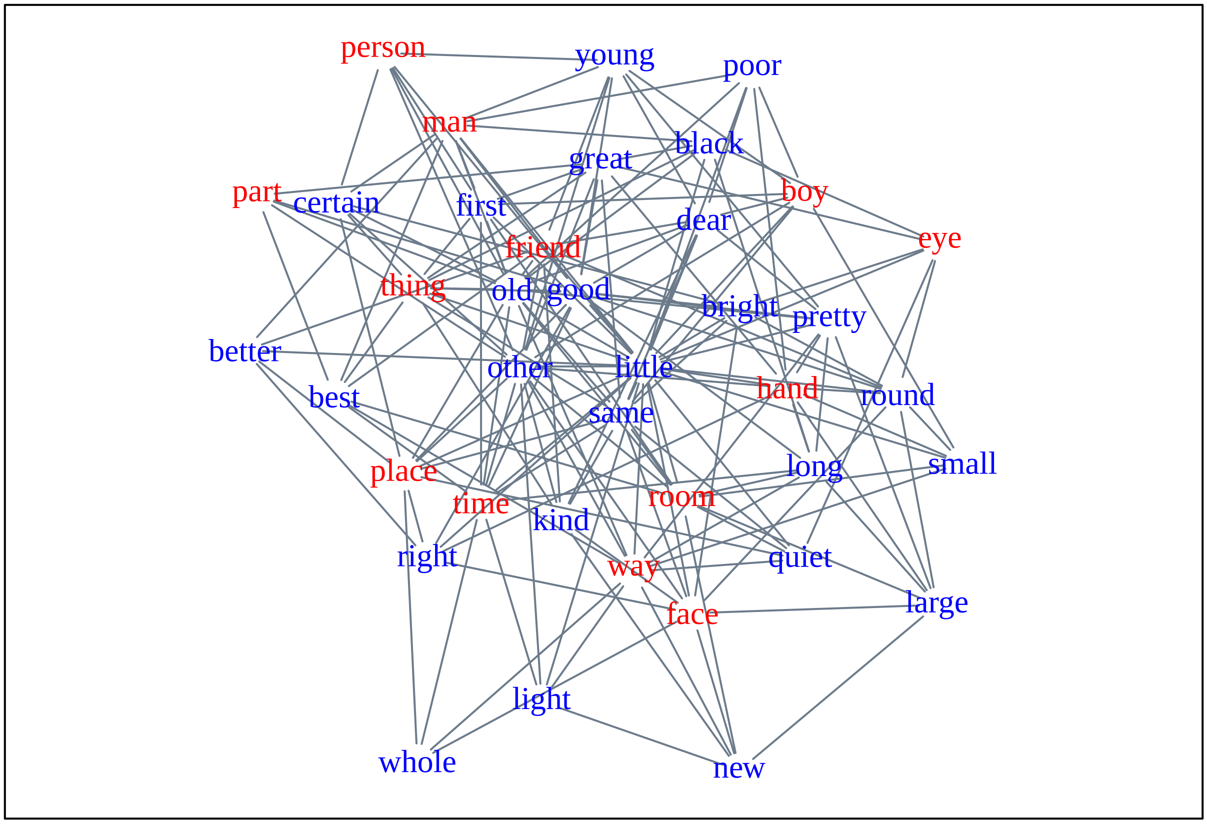
```r
library(igraph);library(dplyr);library(ggplot2);library(expss)

par(mfrow=c(1,1),mar=c(5,5,5,5))
lmdeg <- as.data.frame(degree(lm))

lmdeg %>% ggplot(., aes(x=degree(lm))) + scale_fill_brewer(palette = "Spectral") +
  geom_histogram(bins = 50, colour=’red’) +
  labs(title="Frequency of Node Degrees",
       caption = "Source: http://www-personal.umich.edu/~mejn/netdata",
       x = "Degree of Nodes in Graph",
       y = "Count")
```


Frequency of Node Degrees

Source: http://www−personal.umich.edu/~mejn/netdata

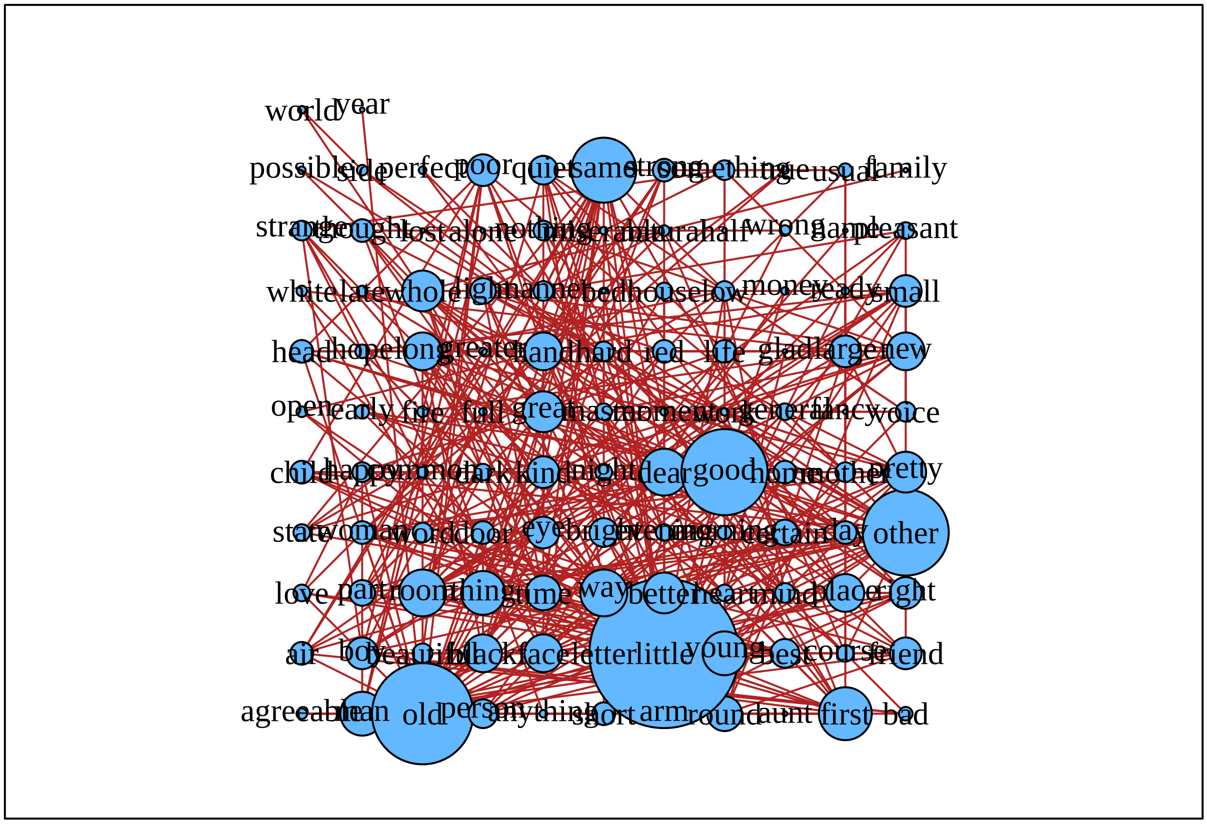
To get more familiar with the data we look at the histogram above to see the frequency of the degrees of the nodes of the graph. The degree of a node tells us how many edges it is connected to. In our case an edge between nodes indicates that the two words are adjacent to each other in the novel “David Copperfield”. We can see one word has a degree of 49, making it the most commonly occurring word in the book.

```r
V(lm)$degree <- degree(lm, mode="all")
cut.off <- mean(V(lm)$degree)
sub <- induced_subgraph(lm, which(V(lm)$degree>cut.off))

par(mfrow=c(1,1),mar=c(.31,.31,1,1))
set.seed(50)
plot(sub, vertex.shape="none", vertex.size=10,
     vertex.color = "blue",
     vertex.label.color=ifelse(V(sub)$value=="0", "blue", "red"),
     edge.color = "slategray4",
     layout=layout_with_fr,
     frame = TRUE)
```



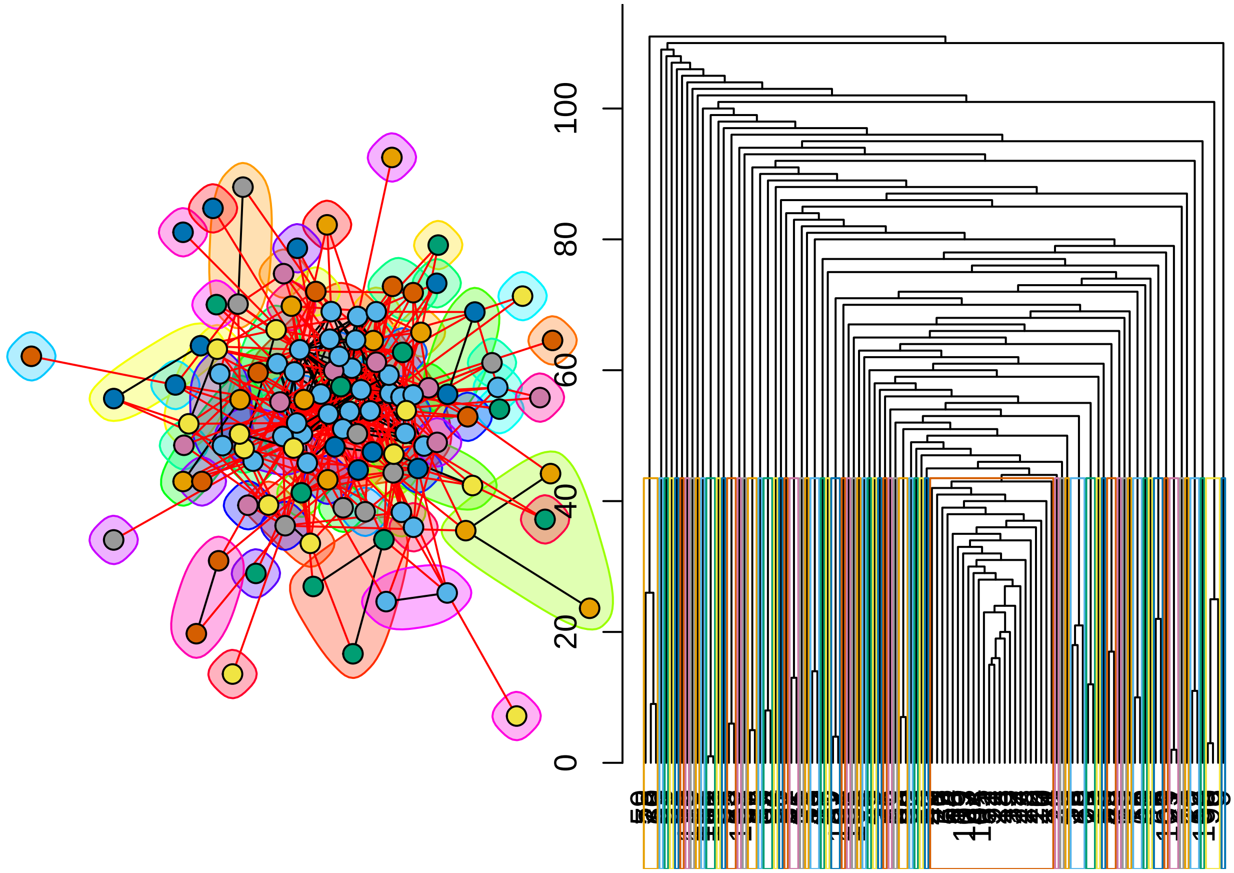
Above we first split the dataset, only looking at nodes with a degree greater than the mean of the degree of all nodes. These would be the most popular words in our dataset. The words are colored based on being a noun or adjective.

```
par(mfrow=c(1,1),mar=c(.31,.31,1,1))

plot(lm, vertex.size=degree(lm)  + .51,
     layout=layout_on_grid(lm),
     vertex.color="steelblue1",
     edge.color = "firebrick",
     vertex.label.color="black",
     frame = TRUE)
```



In the graph above we have the nodes organized in a grid patter with the size of the vertices being proportional to the degree of the vertex. We can see the most used words in decreasing order are “little”, “old”, “other”, “good”, “same”, and so on.

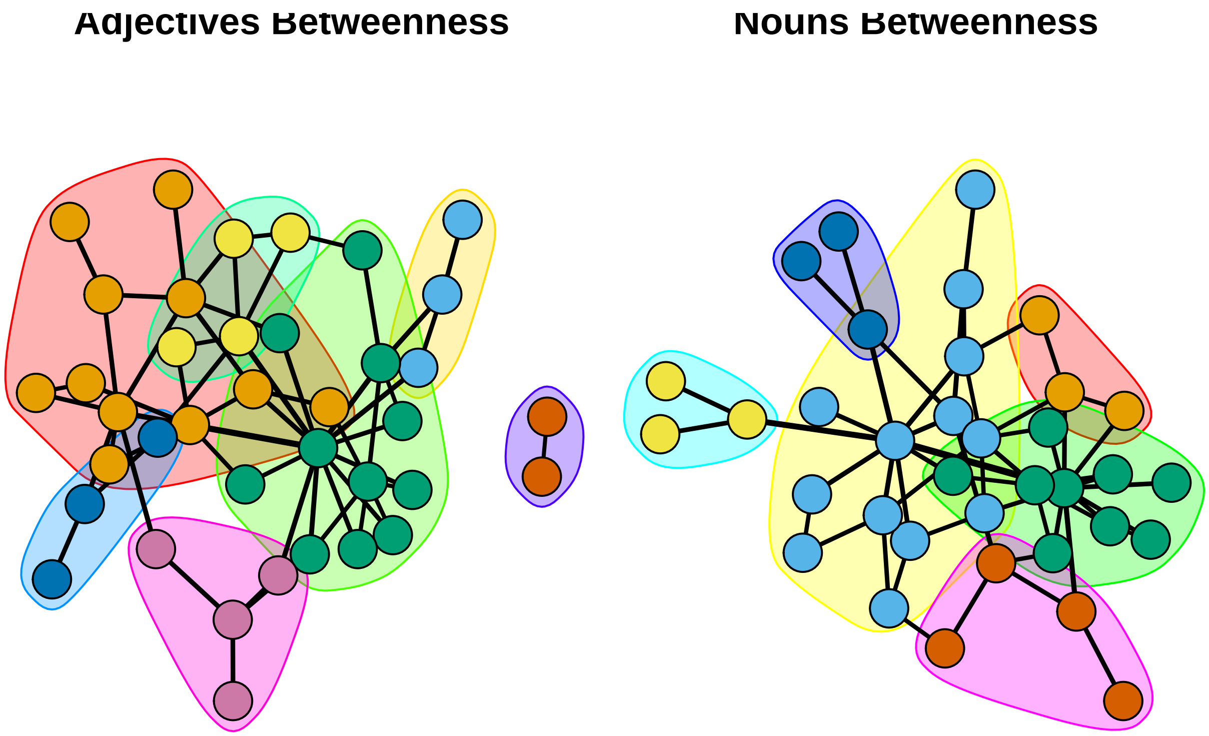
```
par(mfrow=c(1,2),mar=c(0,0,0,0))

g<-cluster_edge_betweenness(lm)


set.seed(70)
plot(g,lm,
     layout =  layout.fruchterman.reingold,
     vertex.label=NA,#V(lm)$label,
     vertex.size = 7)
dendPlot(g)
```



This graph shows the betweenness in the dataset, as well as a dendrogram. The edge betweenness score of an edge measures the number of shortest paths through it. The leafs of the dendrogram are the individual vertices and the root of the tree represents the whole graph. (R Documentation)

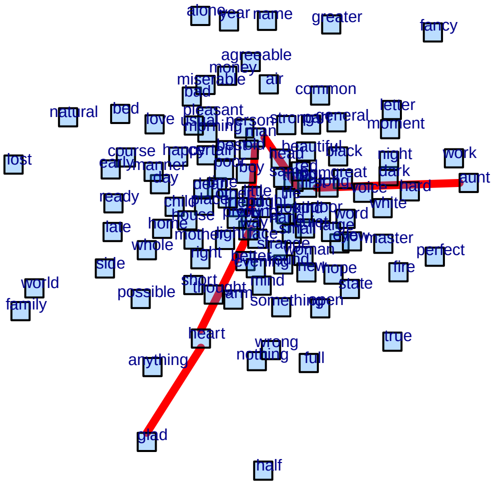
```
par(mfrow=c(1,2),mar=c(5,5,5,5))


adj <- subgraph.edges(lm,
                      eids=which(V(lm)$value==0))

nouns <- subgraph.edges(lm,
                        eids=which(V(lm)$value==1))

g<-cluster_edge_betweenness(adj)
eba<-2+edge.betweenness(adj,directed=F)/100
ebn<-2+edge.betweenness(nouns,directed=F)/100
par(mfrow=c(1,2), mar=c(.4,.4,.4,.4))


set.seed(444)
plot(g, adj,vertex.label=NA,
     edge.width=eba,
     edge.color="black",
     vertex.color=adjustcolor("navy", alpha.f = .5) ,
     main="Adjectives Betweenness")
```

```
g<-cluster_edge_betweenness(nouns)
set.seed(444)
plot(g, nouns,vertex.label=NA,
     edge.width=ebn,
     edge.color="black",
     vertex.color=adjustcolor("firebrick", alpha.f = .5) ,
     main="Nouns Betweenness")
```



In the graph above the dataset is split into two subgraphs, one for nouns and one for adjectives. The graphs are plotted and their betweenness is highlighted.

```
par(mfrow=c(1,1))

gd<-get_diameter(lm)
V(lm)$color <-ifelse(V(lm)$label %in% gd ,"black","gold")

E(lm, path = gd)$color <- "red"
E(lm, path = gd)$width <- 4

set.seed(3)
plot(lm,
     layout=layout_with_fr,
     vertex.label=V(lm)$name2,
     vertex.shape = "csquare",
     vertex.label.family="Helvetica",
     vertex.label.cex=.5,
     vertex.label.dist=.5,
     vertex.size = 8,
     vertex.color=adjustcolor("dodgerblue1", alpha.f = .3))
```



Here we see the longest path along the network. The words along the path are "glad"-"heart"-"way"-"young"-"man"-"first"-"aunt". These two nodes are also the furthest from each other in the whole graph.

Citations

M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Preprint physics/0605087 (2006)

R Documentation