Enron Network Analysis Tutorial

r date()

Enron Tutorial

We provide this Enron Tutorial as an appendix to the paper in Journal of Statistical Education, *Network Analysis with the Enron Email Corpus*. The paper describes the centrality measures in detail, and we go through the steps in the R analysis here.

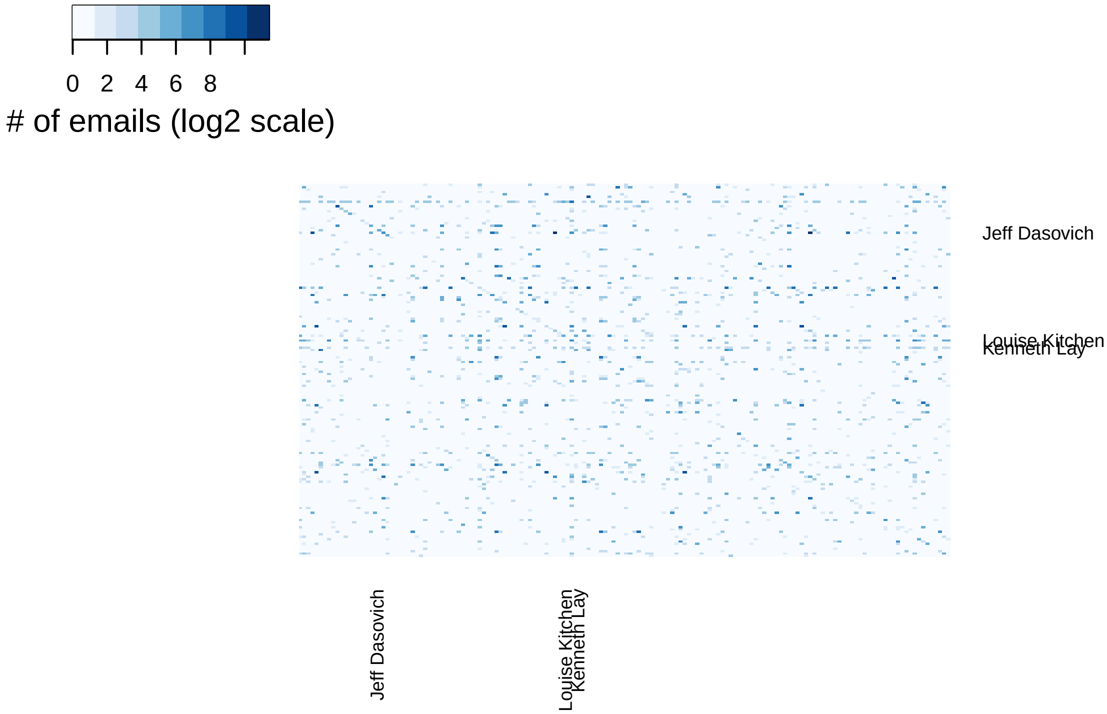
As in the .Rmd file with the R code, be sure to install the pacakges *WGCNA* and *igraph*.

The first step of the analysis is to import the data and create the adjacency matrix of our choice. AM represents emails sent from node i to node j with messages sent via CC weighted as described in the paper. The transpose of the matrix AMt represents emails recieved by node i from node j (again, with CC values weighted differently than messages sent directly to an individual). The sum of the two matrices, AM2, represents the total email correspondence (sent and received) between nodes i and j.

```
setwd("H:/teaching/ResearchCircle/Spring2014 - DataScience/JSE.DSS")
AM = as.matrix(read.csv("Final Adjacency Matrix.csv",
                        sep=",", header=TRUE, row.names=1))  # sent emails
AMlist = read.csv("Final Adjacency Matrix.csv",
                  sep=",", header=TRUE, row.names=1)  # sent emails as list
# employee information might be interesting to analyze for considering relationships
#   within the company
# enronemployees = read.table("Enron Employee Information.csv", sep=",", header=T)
AMt = t(AM)  # received emails
AM2 <- AM + t(AM) - 2*diag(diag(AM))  # sent and received emails
```

We can represent the adjacency matrix graphically using a heatmap.

```
AM.names=c(rep(NA,20), row.names(AM)[21],rep(NA,44), row.names(AM)[66],
           rep(NA,2), row.names(AM)[69], rep(NA,87))
heatmap.2(log2(AM+1), Rowv=FALSE, Colv= FALSE, dendrogram="none",
          col = (brewer.pal(9,"Blues")),scale="none", trace="none",
        labRow=AM.names,labCol=AM.names, colsep=FALSE,
        density="none", key.title="", key.xlab="# of emails (log2 scale)" ,
        mar=c(8,8))
```



Eigenvector Centrality

The first measure of centrality that we use is eigenvector centrality; the *evcent* function is available in the *igraph* package. Degree, Betweenness, and Closeness centrality measures are also given in the *igraph* package.

```
# eigenvalue centrality (on both directed graphs),
# degree, betweenness, and closeness
eng <- graph.adjacency(do.call(rbind,AMlist))    # creates a network graph using the adjacency matrix
engt <- graph.adjacency(do.call(cbind,AMlist))    # creates a network graph using the transpose of the a

eigcent <- igraph::evcent(eng, directed=TRUE)     # eigenvalue centrality
eigcentt <- igraph::evcent(engt, directed=TRUE)  # eigenvalue centrality on transpose of graph
dcent <- igraph::degree(eng)                    # degree centrality
bmeas <- igraph::betweenness(eng)               # betweenness
cmeas <- igraph::closeness(eng)                  # closeness

# TOM
AM2 <- AM2 / max(AM2)                            # set values between 0 and 1
TOM <- TOMsimilarity(AM2)                        # create TOM
```

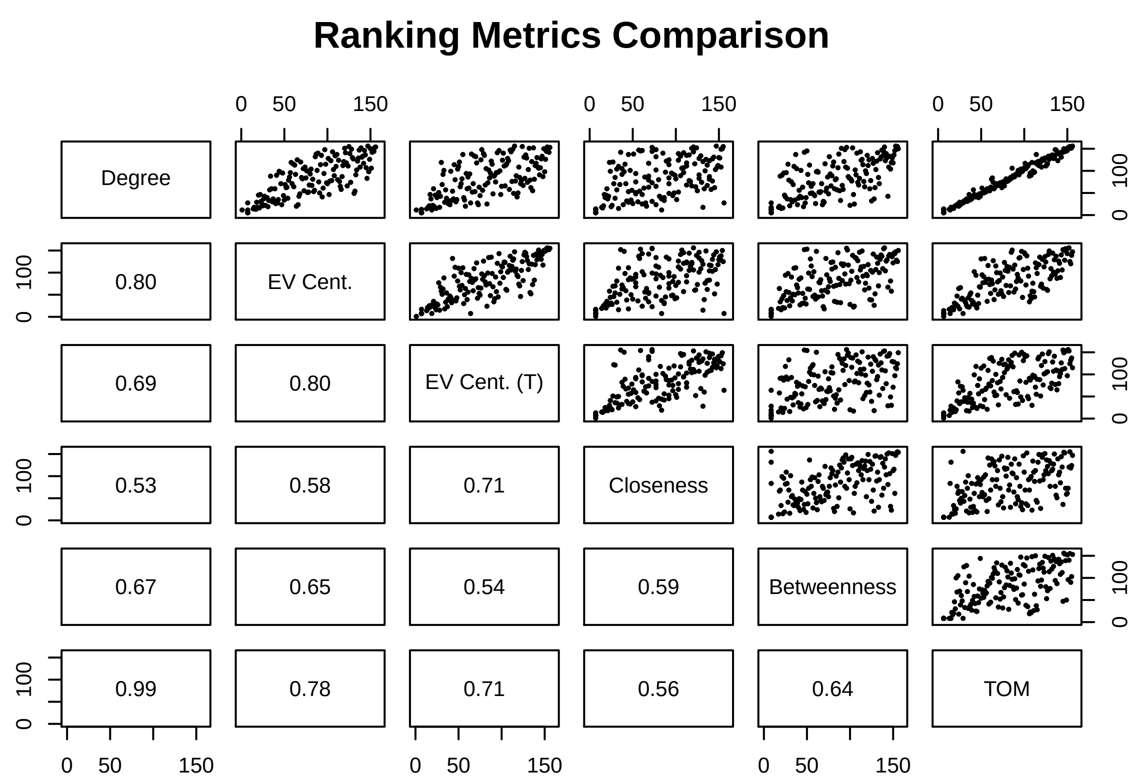
```
## ..connectivity..
## ..matrix multiplication..
## ..normalization..
## ..done.
```

```
TOMrank <- as.matrix(apply(TOM,1,sum))          # grab its row-sums
rownames(TOMrank) <- rownames(AM)
colnames(TOMrank) <- "value"
```

Initially, we plot the ranks of the individuals based on the different measures of centrality. The ranks are clearly correlated, but we can also see that they seem to be measuring different qualities of the email correspondence matrix.

```
comptable <- matrix(ncol=6, nrow=dim(AM)[1])
comptable[,1] <- rank(dcent)
comptable[,2] <- rank(eigcent$vector)
comptable[,3] <- rank(eigcentt$vector)
comptable[,4] <- rank(cmeas)
comptable[,5] <- rank(bmeas)
comptable[,6] <- rank(TOMrank)

pairs(comptable[,1:6],pch=20,main="Ranking Metrics Comparison",
      labels=c("Degree","EV Cent.", "EV Cent. (T)","Closeness","Betweenness", "TOM"),
      cex=.5,xlim=c(0,160),ylim=c(0,160),lower.panel=panel.cor)
```



Next, we lists the top 10 most central individuals for each metric. Note that we use the negative of the centrality measure so that the order function produces the first individual as the most central.

```
rankedEnron <- data.frame(Degree = rownames(AM)[order(-dcent)],
                          EVcent = rownames(AM)[order(-eigcent$vector)],
```

```r
                              EVcentT = rownames(AM)[order(-eigcentt$vector)],
                              Close = rownames(AM)[order(-cmeas)],
                              Between = rownames(AM)[order(-bmeas)],
                              TOM = rownames(AM)[order(-TOMrank)])
rankedEnron[1:10,]
```

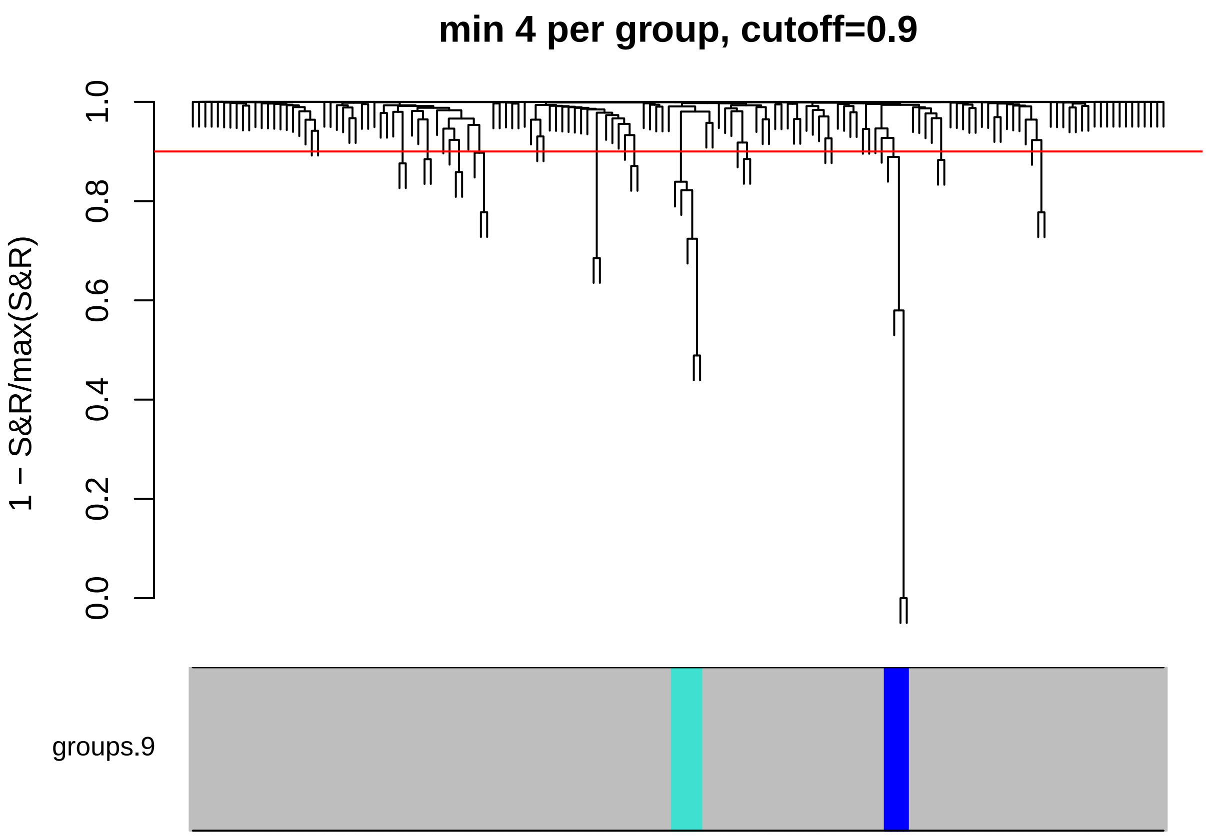
```
##             Degree          EVcent          EVcentT            Close
## 1    Jeff Dasovich      Tana Jones  Sara Shackleton    Robert Benson
## 2     Mike Grigsby Sara Shackleton     Susan Bailey     Mike Grigsby
## 3       Tana Jones Stephanie Panus      Marie Heard   Louise Kitchen
## 4  Sara Shackleton     Marie Heard       Tana Jones  Kevin M. Presto
## 5  Richard Shapiro    Susan Bailey  Stephanie Panus      Susan Scott
## 6   Steven J. Kean        Kay Mann  Elizabeth Sager       Scott Neal
## 7   Louise Kitchen  Louise Kitchen   Jason Williams   Barry Tycholiz
## 8      Susan Scott Elizabeth Sager   Louise Kitchen    Greg Whalley
## 9   Michelle Lokay  Jason Williams Jeffrey T. Hodge Phillip K. Allen
## 10   Chris Germany   Jeff Dasovich     Gerald Nemec    Jeff Dasovich
##            Between             TOM
## 1  Louise Kitchen   Jeff Dasovich
## 2    Mike Grigsby Richard Shapiro
## 3     Susan Scott  Steven J. Kean
## 4   Jeff Dasovich    Mike Grigsby
## 5       Mary Hain      Tana Jones
## 6      Sally Beck Sara Shackleton
## 7     Kenneth Lay       Mary Hain
## 8      Scott Neal     Marie Heard
## 9      Kate Symes Stephanie Panus
## 10 Cara Semperger     Susan Scott
```

Hierarchical Clustering

Below, we create the heirarchical cluster with both the symmetric (sent and received) adjacency email matrix as well as the TOM adjacency build from the symmetric measures. After building the dendrogram, we find groups of employees who are strongly linked and report the names of the individuals.

```r
# dissimilarity is 1 - number of sent and received / max of sent and recieved over all individuals
dissAM2=1-AM2
# Create the heirarchical clustering
hierAM2=hclust(as.dist(dissAM2), method="average")
groups.9=as.character(cutreeStaticColor(hierAM2, cutHeight=.9, minSize=4))

# Plot results of all module detection methods together:
plotDendroAndColors(dendro = hierAM2,colors=data.frame(groups.9),
                    dendroLabels = FALSE, abHeight=.9,
                    marAll =c(0.2, 5, 2.7, 0.2), hang=.05,
                    main ="min 4 per group, cutoff=0.9",ylab="1 - S&R/max(S&R)")
```




```r
table(groups.9)
```

```
## groups.9
##      blue      grey turquoise
##         4       147         5
```

```r
row.names(AM)[groups.9=="turquoise"]
```

```
## [1] "Susan Bailey"     "Marie Heard"      "Tana Jones"       "Stephanie Panus"
## [5] "Sara Shackleton"
```

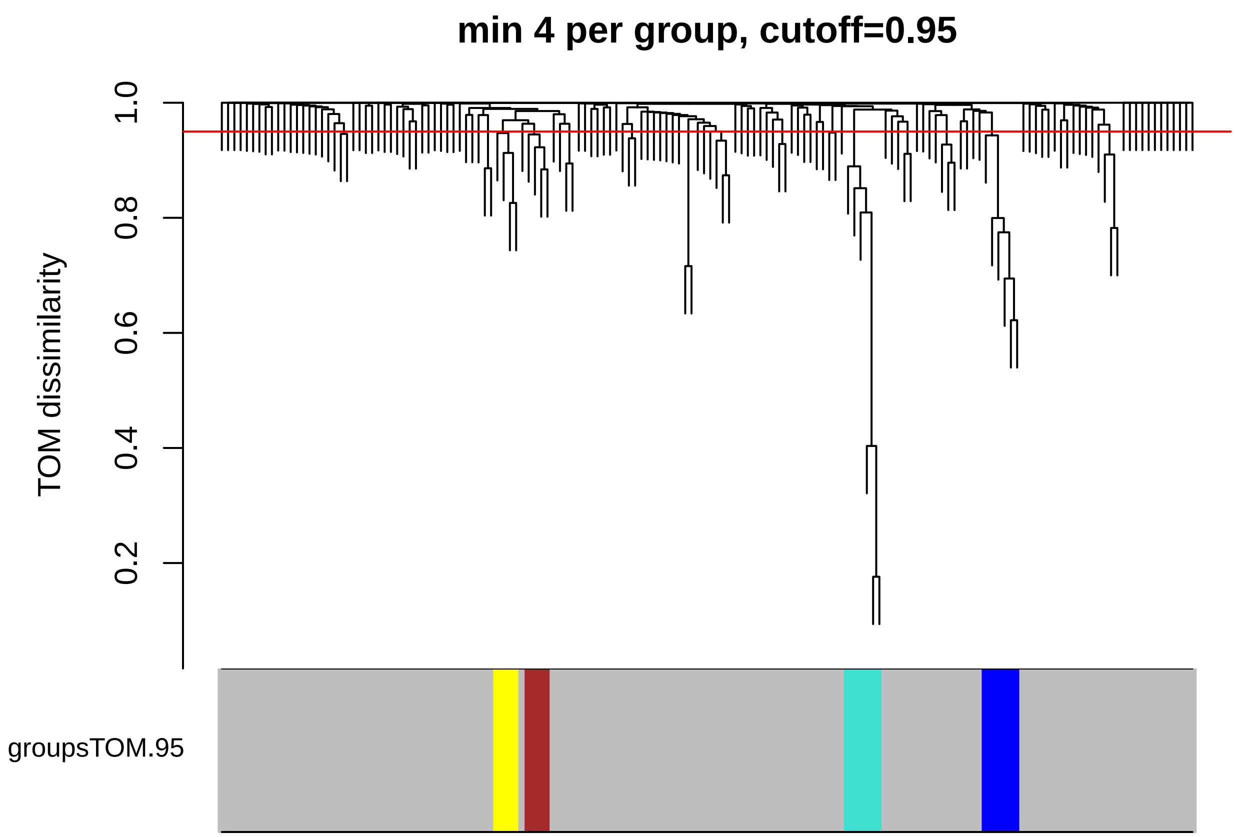
```r
row.names(AM)[groups.9=="blue"]
```

```
## [1] "Jeff Dasovich"    "Mary Hain"        "Steven J. Kean"   "Richard Shapiro"
```

```r
###  Now cluster with TOM
# for the next plot, dissimilarity uses TOM metric to encorporate neighbors
dissTOM=TOMdist(AM2)
```

```
## ..connectivity..
## ..matrix multiplication..
## ..normalization..
## ..done.
```

```r
rownames(dissTOM) <- rownames(AM)
colnames(dissTOM) <- rownames(AM)
# Create the heirarchical clustering
hierTOM=hclust(as.dist(dissTOM), method="average")
groupsTOM.95=as.character(cutreeStaticColor(hierTOM, cutHeight=.95, minSize=4))
# Plot results of all module detection methods together:
plotDendroAndColors(dendro = hierTOM,colors=data.frame(groupsTOM.95), abHeight=.95,
                    dendroLabels = FALSE, marAll =c(0.2, 5, 2.7, 0.2),
                    main ="min 4 per group, cutoff=0.95", ylab="TOM dissimilarity")
```




```r
table(groupsTOM.95)
```

```
## groupsTOM.95
##      blue     brown      grey turquoise    yellow
##         6         4       136         6         4
```

```r
row.names(AM)[groupsTOM.95=="turquoise"]
```

```
## [1] "Robert Badeer"    "Jeff Dasovich"    "Mary Hain"
## [4] "Steven J. Kean"   "Richard Shapiro"  "James D. Steffes"
```

```r
row.names(AM)[groupsTOM.95=="blue"]
```

```
## [1] "Susan Bailey"    "Marie Heard"     "Tana Jones"      "Stephanie Panus"
## [5] "Elizabeth Sager" "Sara Shackleton"
```

```r
row.names(AM)[groupsTOM.95=="brown"]
```

```
## [1] "Lindy Donoho"    "Michelle Lokay"  "Mark McConnell"  "Kimberly Watson"
```

```r
row.names(AM)[groupsTOM.95=="yellow"]
```

```
## [1] "Drew Fossum"   "Steven Harris" "Kevin Hyatt"   "Susan Scott"
```