Keith Folsom

MSDA Math Workshop Final

July 2015

Data Science in Social Network Analysis

Who is the central person in a network and how do you identify this individual? To answer these types of questions, we will explore a use case employing social network analysis to examine peer influence within a network. This examination will show that social network analysis (SNA) employs a powerful combination of advanced math and visualization techniques. As we proceed through the use case, we will introduce network analysis using matrices, eigenvectors, and visualizations in R using gglot2 and igraph.

The particular use case will focus on the topic of peer influence of doctors in the medical field. In the medical industry, specialists will seek out the assistance of another specialist who is more knowledgeable in specific area within the specialty – in other words, someone who is a known expert and has considerable influence within a particular medical community. Pharmaceutical companies frequently target these experts as means to more quickly disseminate information to members within a particular social network.

Social Networks and Adjacency Matrices

A social network can be represented as a set of vertices or nodes connected by lines or edges. The nodes in this particular case represent physicians. Edges or lines between the nodes signify relationships between physicians in the network. Nodes and edges can be represented in what is referred to as adjacency matrix (figure 1). The data used in this example is sourced from the "network" R package and uses the flo data set to create an adjacency matrix representing a fictitious network of 16 physicians. This particular adjacency matrix is undirected, meaning the resulting matrix is symmetrical and is a mirror image across the diagonal.

Within the adjacency matrix (figure 1), you'll notice that the diagonal is comprised of all 0 values (in red) because an edge is not created for a doctor to him or herself. A value of 1 indicates a connection between two doctors. Other adjacency matrices exist such as directed and weighted matrices. These types of matrices provide direction and weight to the relationship and are asymmetrical. However, for the purposes of this use case exploration, we will be using an undirected adjacency matrix.

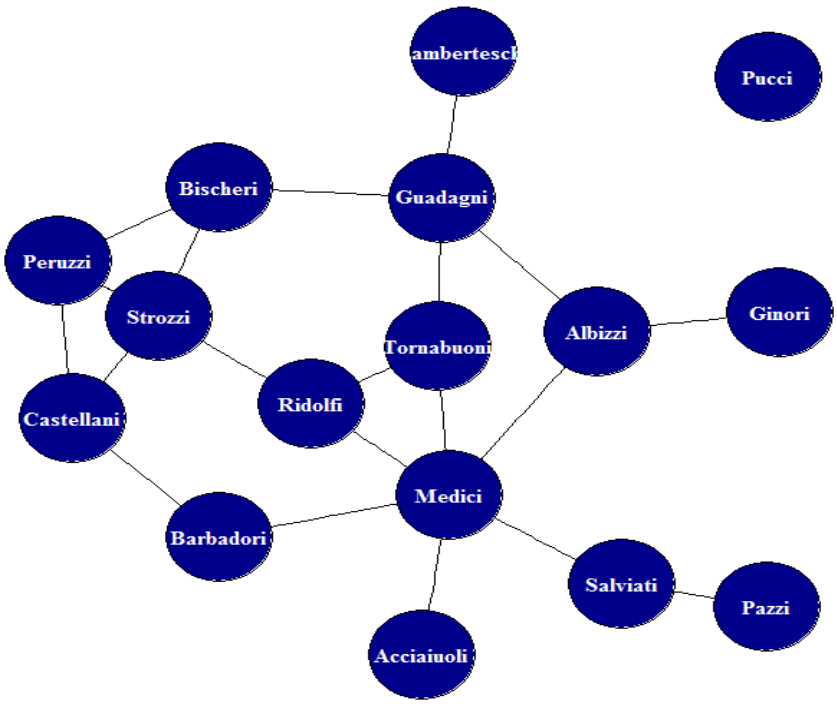
Figure 1 – Adjacency matrix representing physicians in a network

	Acciaiuoli	Albizzi	Barbadori	Bischeri	Castellani	Ginori	Guadagni	Lamberteschi	Medici	Pazzi	Peruzzi	Pucci	Ridolfi	Salviati	Strozzi	Tornabuoni
Acciaiuoli	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Albizzi	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
Barbadori	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Bischeri	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
Castellani	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
Ginori	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Guadagni	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
Lamberteschi	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Medici	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1
Pazzi	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Peruzzi	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
Pucci	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ridolfi	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
Salviati	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Strozzi	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0
Tornabuoni	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0

The adjacency matrix can be depicted using a network graph created by the igraph package in R (figure 2).

```
require(network)
require(igraph)

# use the flo data set from the network package
data(flo)

plot.igraph(g,vertex.label=V(g)$name,,vertex.size=30,vertex.label.color="white",
vertex.label.font=2,vertex.color="darkblue",edge.color="black")
```

Figure 2. Network graph of the physician social network



We're now able to visualize this network but how do we determine who the most important physician in this network is? If we wanted to target one physician to have the most influence on the rest of the group, who would it be? To determine this, we need to explore the concept of centrality, which is a measure of the relative importance of a node within a network. The more central a node (or physician in this case) the more influence the node has over the other nodes in the network.

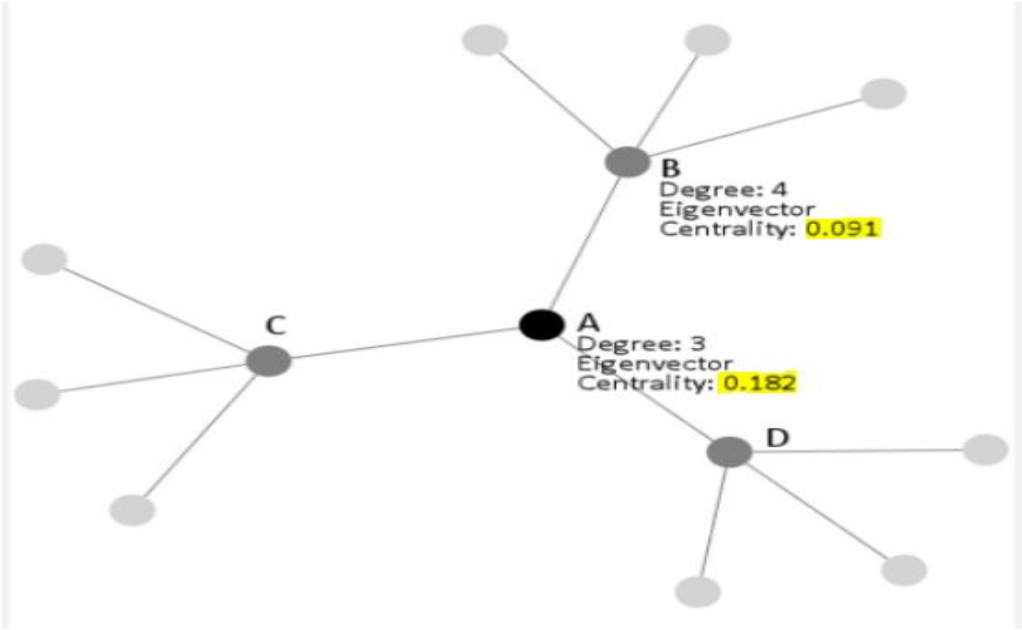
Key Concepts of Centrality in Social Network Analysis

Centrality: The relative importance of a node within a graph. There are various approaches to measure centrality in social network analysis. This example will focus on three different measures of centrality – degree, eigenvector, and betweenness.

Degree centrality: This may be the simplest centrality calculation. Degree centrality is calculated by the number of edges attached to a node. "Though simple, degree is often a highly effective measure of the influence or importance of a node: in many social settings people with more connections tend to have more power. " (Newman)

Eigenvector centrality: A more sophisticated method to calculate centrality incorporating both the number and quality of a node's connections. Eigenvector centrality factors in to the calculation the fact that not all nodes are equal. A node which is connected to nodes who are themselves more influential will have a higher eigenvector centrality than a node which is connected to less influential nodes (Newman). In the network diagram below (figure 3), we see that node B has a degree of 4 and node A has a degree of 3. However, node A has a higher eigenvector centrality value because of its connections to nodes B and C.

Figure 3. Network diagram with Degree and Eigenvector Centrality



Betweenness centrality: The represents the amount of control a node has on the flow information between other nodes as well provides a measure of how quickly a node can transfer information over the network.

Network Examination using Eigenvector Centrality

Let's examine the physician network using eigenvector centrality. The R code below (Bogard, 2012) calculates the eigenvector centrality values for each of the 16 physicians in the fictitious network for our use case.

```
require(network)
require(igraph)
require(dplyr)

data(flo)

# load the flo data set as a matrix
# this represents an undirected/symmetrical adjacency matrix
m <- flo

# compute eigenvalues and eigenvectors of the adjacency matrix m
EV <- eigen(m)

physician.names <- colnames(m)

# get the eigenvector associated with the largest eigenvalue
centrality <- cbind(physician.names, data.frame(EV$vectors[,1]))

names(centrality) <- c("Physician", "EV_Centrality")

# sort the phsycians using the absolute value of the eigenvector centrality
# the physician at the top of the list will have the highest eigenvector centrality value
arrange(centrality, desc(abs(EV_Centrality)))
```

```
      Physician EV_Centrality
1        Medici   -0.43030809
2       Strozzi   -0.35598045
3       Ridolfi   -0.34155264
4    Tornabuoni   -0.32584230
5      Guadagni   -0.28911560
6      Bischeri   -0.28280009
7       Peruzzi   -0.27573037
8    Castellani   -0.25902617
9       Albizzi   -0.24395611
10    Barbadori   -0.21170525
11     Salviati   -0.14591720
12   Acciaiuoli   -0.13215429
13 Lamberteschi   -0.08879189
14       Ginori   -0.07492271
15        Pazzi   -0.04481344
16        Pucci    0.00000000
```

Physician	EV_Centrality
Medici	0.430308094
Strozzi	0.355980448
Ridolfi	0.341552644
Tornabuoni	0.325842301
Guadagni	0.289115599
Bischeri	0.282800087
Peruzzi	0.275730374
Castellani	0.259026167
Albizzi	0.243956109
Barbadori	0.211705251
Salviati	0.145917196
Acciaiuoli	0.132154295
Lamberteschi	0.088791888
Ginori	0.074922708
Pazzi	0.044813436
Pucci	0

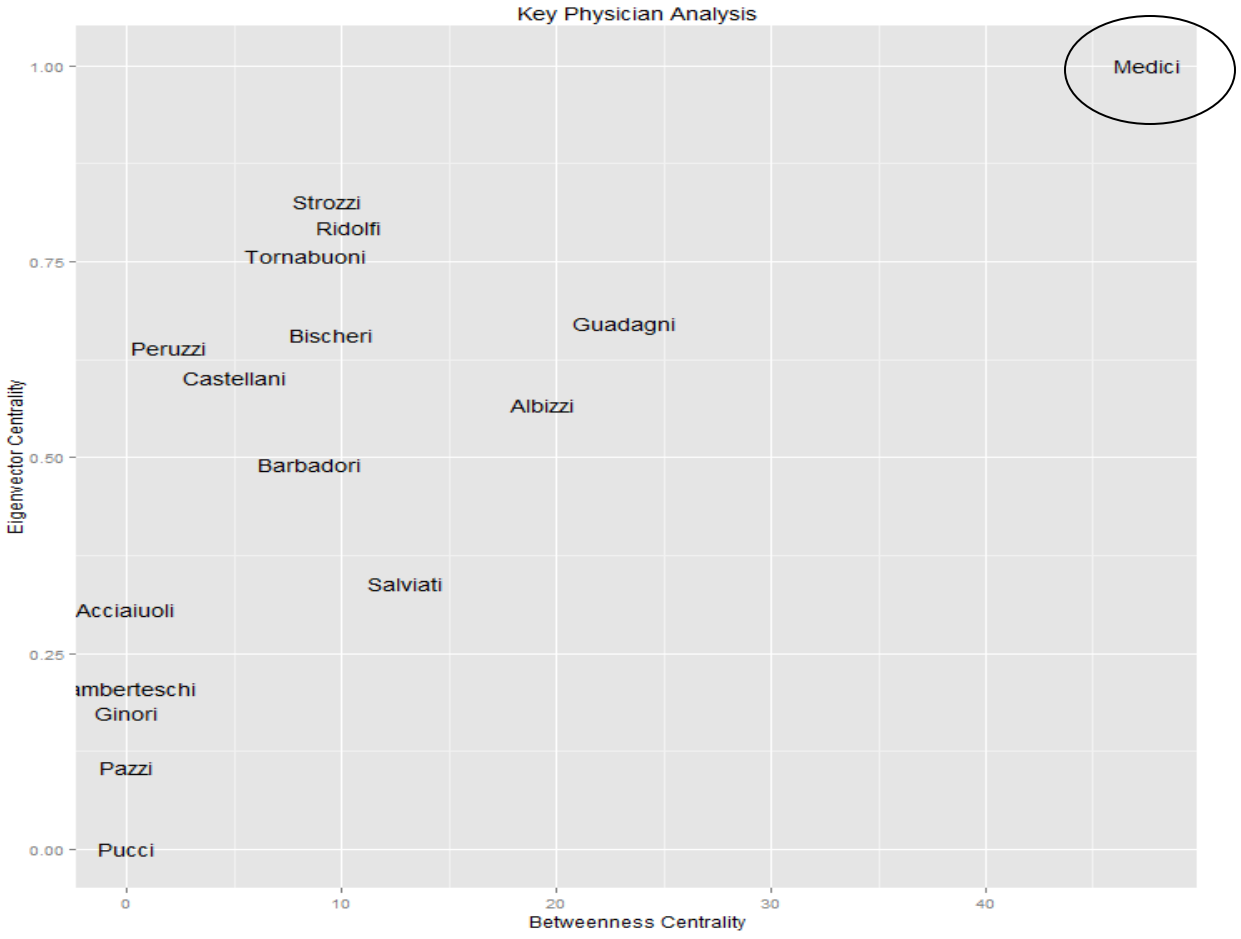
Sorting the list of physicians by the absolute value of the eigenvector centrality calculation, we see that Dr. Medici has the highest EV value, indicating that he is very influential within this network. In contrast, Dr. Pucci has a 0 eigenvector centrality value which is expected since he was not connected to any other physicians in the network.

Network Examination using Eigenvector Centrality and Betweenness

Let's examine the same physician network using eigenvector centrality and betweenness. Recall that betweenness is a measure of a node's ability to transfer information across the network based on the number of shortest paths to other nodes. A node with high betweenness centrality has the ability to quickly pass information through the network.

The R code below (Bogard, 2012) calculates the eigenvector centrality and betweenness values for each of the 16 physicians in the fictitious network. The eigenvector centrality and betweenness are plotted on the x and y axes respectively for a visualization which quickly allows the reader to determine who the most influential doctor is within the network.

```
require(ggplot2)
require(igraph)

G=graph.adjacency(m, mode="undirected",weighted=NULL,diag=FALSE)

cent<-data.frame(bet=betweenness(G),eig=evcent(G)$vector)

res<-as.vector(lm(eig~bet,data=cent)$residuals)

cent<-transform(cent,res=res)

p <- ggplot(cent,aes(x=bet,y=eig,label=rownames(cent))) +xlab("Betweenness Centrality") + ylab("Eigenvector Centrality")

p + geom_text() + labs(title="Key Physician Analysis")
```



Conclusion

Based on this analysis, we see that Dr. Medici is the most influential doctor within this network due to having both a high eigenvector centrality value and a high betweenness value. In the use case of a pharmaceutical company looking to target a single physician for highest impact, Dr. Medici would be the obvious choice as he is the most central person in this network. We also see that Dr. Guadagni would be a good choice as a second option after Dr. Medici.

References

Matt Bogard. "An Introduction to Social Network Analysis with R and NetDraw". April 2012.

Available at: http://econometricsense.blogspot.com/2012/04/introduction-to-social-network-analysis.html

M. E. J. Newman. "The mathematics of networks"

Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1040

Available at: http://www-personal.umich.edu/~mejn/papers/palgrave.pdf

Dr. Cecilia Mascolo. "Social and Technological Network Analysis".

Available at: https://www.cl.cam.ac.uk/teaching/1314/L109/stna-lecture3.pdf

Leo Spizzirri. "Justification and Application of Eigenvector Centrality". March 2011.

Available at: https://www.math.washington.edu/~morrow/336_11/papers/leo.pdf