Understanding a network of doctors

M3A50 – Project 2

Hitesh Kumar
Department of Mathematics, Imperial College London

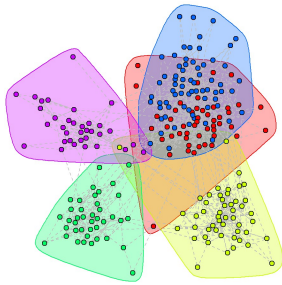
Imperial College London

PURPOSE OF THIS INVESTIGATION

In this investigation we use a data set with information about various "attributes" of doctors from four cities, and focus on what makes information spread faster or better around the network. We will begin by understanding our network and seeing what makes it up (and how to take it apart) and then we will draw conclusions based on simulations we run on it

COMMUNITY DETECTION

- First, instead of relying on cities, we split the doctors into communities based on their connections with other doctors. Using the "Leading eigenvector" method[1]:



- Here we see that our network splits into **5 communities** (recall we started with 4 cities).

- The overlap between red and blue shows how a city can actually be made up of two distinct communities

- If we assume that cities are the "correct" way to split the doctors into communities, then we can find the **accuracy** of our model : 70.2%!

DISRUPTING THE NETWORK

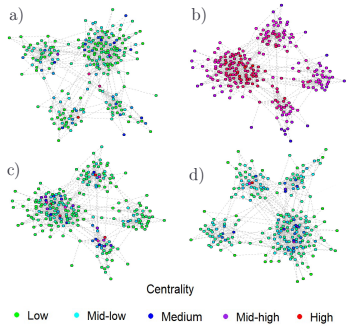
- Its useful to know which nodes are the most important in keeping the network connected, and we can measure this with centrality.

- The task here is to use these measures to identify which nodes, when removed would best "disrupt" the network.

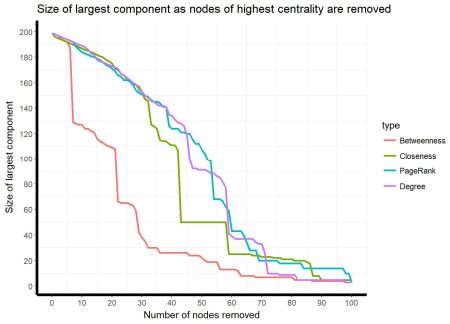
- Because our network is made up of one giant component and multiple small components, we can measure disruption by seeing how the size of that component changes

- We calculate each type of centrality of each node, record the size of the largest component and repeat this process, removing the node of highest centrality and repeating 100 times.

- Then we can plot the data to see which centrality-based removal is quickest at decreasing the size, and the nodes ordered by that measure will be the most disruptive/important.



Network colored by: a) betweenness, b) closeness, c) PageRank and d) degree centralities.



- We see the important nodes are those with the highest bet. centrality, as removing them is the best way to disrupt the network

- In context, this could mean that the most important doctors are those who act as connections between large, separate groups.

SIMULATING AN INFORMATION EPIDEMIC

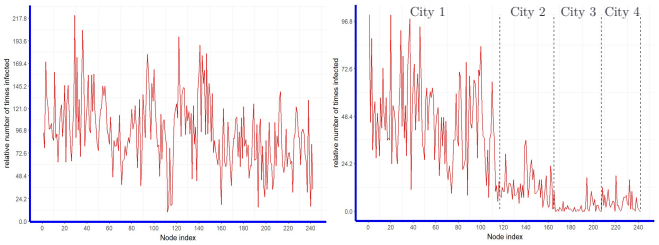
- Now that we better understand the network, we will see how information spreads across it by modelling it as an incurable disease. We use given function "simEpi()" to do this.

- We start by initially infection random nodes and seeing which other nodes usually get infected quickly, by repeating the simulation for a small number of time-steps and counting how often each node is infected.

- Clearly, the nodes that are more prone to infection will much more often be infected.

- We then do the same, but now initially infecting nodes with the highest betweenness centrality. As we see just above, this may be the best way to spread information through the network fast.

- To compare results, we plot the "proneness to infection" against the node index to see how the information has spread



- Its clear there are nodes that receive information quicker than others, and in the plot with betweenness-based initial infection, city 1 finds out about it much faster than other cities do.

- We also do the same by simulating the spread with doctors that read more journals being initially "infected" with information. Observing the attributes of those that got the information quickest, we see that they are:

- Not usually from city 4
- Much more likely to have started between 1920-1940 than earlier or later
- Usually internal medicine specialists

WHY DID SOME ADOPT IT LATER?

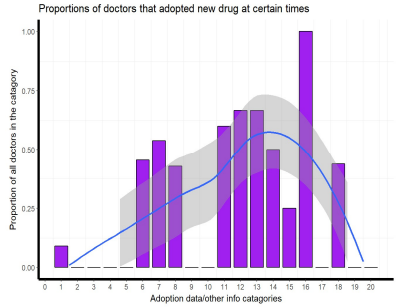
- To answer such a question, we start with the assumption that those who adopted the drug earlier also learnt about it earlier.

- We can run a simulation with these nodes initially infected, and see if the ones who are very unlikely to be quickly infected are the same ones who adopted it later

- We see now that this might actually be the case!

- The plot strongly suggests that those who are slow to be infected are also the ones who prescribed the drug much later.

- And we already know what decides how quickly/slowly doctors get information.



Certain attributes => late exposure => late adoption

And so

Certain attributes => late adoption

REFERENCES

[1] – Gabor Csardi, *leading.eigenvector.community()*, V 0.4.1, [R function], in "R documentation", available from: https://www.rdocumentation.org [Accessed 03/01/18]