Notes(4): Network Vis, Gephi, R, Text processing

Sagar Parekh 1213332118

I. Gephi:

A) Tutorial from website
1) Import File:
    a) Open a Graph file: File -> Open
    b) Import Report: Click "OK"
    c) You should now see a graph
2) Vizualization: You can
    a) Modify edge thickness
    b) Zoom and pan out the vizualization
    c) Reset the position of graph
3)Layout:
    a)Layout algorithms sets the graph shape, it is the most essential action.
    b)There are a number of different layouts to choose from and also you can control the layout by adjusting the properties of the layout. The purpose of Layout Properties is to let you control the algorithm in order to make a aesthetically pleasing representation.
4)Ranking by color:
    a)Ranking module lets you configure node’s color and size.
    b)You can rank by color. You can also configure the colors.
    c)You can also rank by size.
5)Metric:
    a)We can calculate the average path length for the network. It computes the path length for all possibles pairs of nodes and give information about how nodes are close from each other.
    b)We can find the metrics at the statistics module on the right panel.
    c)Select the metric to calculate and run it. Click OK for general configuration and you have the metric result.
6)Ranking by size:
    a)Metrics generates general reports but also results for each node. We can rank the nodes by size by selecting different rank parameters. We can also set the minimum and maximum node size.
7)Show Labels:
    a)You can display the node labels and set label size.
8)Community Detection:
    a)The ability to detect and study communities is central in network analysis. We would like to colorize clusters in our example. We can find different communities by running the modularity option under statistics.
9)Partition:
    a)Once the communities are created the partition module can be used to colorise different communities. You can select different partitioning parameters.
10)Filters:
    a)You can add different filters to hide nodes and edges.
11)Preview:
    a) Before exporting your graph as a SVG or PDF file, go to the Preview and you can see exactly how the graph will look like and put in end touches.

II. Network Vizualization:

-In Social network vizualization the designer hopes to find out undefined connections. DV delivers somewhat unexpected. Connect memories and bring in all the readers.

-Remember the basic rule: Make the complex things simple, divide big data in small groups, handle them and find out relations.

-The non-linear processing chain of data: Try to interpret data variable and try to filter them and then parse. Present it and if not satisfied then go over the previous stages again. Refine it and try to make it better. All these stages need different set of skills.

-Orderness in data: How to appreach it more systematically. Different levels: macro and micro which can be seen as local and global. Now lets group it by dimensions so as to sort the messy data. You can also sort the data by timestamp.
- First level is overview. Zoom in and out for orderness. In stats we look at certain data called the qualitative data: social, semantic , geographic. In quantitative data you can find groups and connections within them. IF there are connections then you can represent by linked nodes to describe the patterns.

-What is a network?
-A network is any collection of objects in which some pairs of these objects are connected by links. One more layers of entities which can be any objects.

- Network encoding: how we transform representations into code. Vertices are the nodes, edges will be the connections between a pair.
- Types of network encoding: textual, tabular, graphical, XML. These help us to organise the data in a systematic way.

- Math representations: Tabular form. (Matrix)
- The thickness of the connections depend on the weights.
- Decompose the matrix to understand further. Math helps to viz the patterns.
- Easy to convert table - matrix - network.
- The goal is to understand the undefined relations.

-Basic guidelines on what ot focus depends on number of nodes:
1-100: discuss semnatics, discuss meanings of the relations.
100-1000: focus on particular attributes, explain their structure,
1000-50000: how the structure explains attributes.
>50000: more computational power to zoom in, algo to filter the nodes.

-Different types of Network:
- Undirected: Facebook friendships.
- Directed: Twitter following.
- Multimode: Amazon user products.
- Weighted: Facebook likes.

- Measures:
- Node level:

-Centrality: In graph analytics, Centrality is a very important concept in identifying important nodes in a graph. It is used to measure the importance (or “centrality” as in how “central” a node is in the graph) of various nodes in a graph. Now, each node could be important from an angle depending on how “importance” is defined.

-Clustering coefficient:

-(In/Out) Degree: How many people can this person reach directly?

-Betweenness Centrality: Betweenness centrality measures the number of times a node lies on the shortest path between other nodes. This measure shows which nodes are ‘bridges’ between nodes in a network. It does this by identifying all the shortest paths and then counting how many times each node falls on one. Betweenness is useful for analyzing communication dynamics, but should be used with care. A high betweenness count could indicate someone holds authority over disparate clusters in a network, or just that they are on the periphery of both clusters.

-Closeness Centrality: Closeness centrality scores each node based on their ‘closeness’ to all other nodes in the network. This measure calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths. For finding the individuals who are best placed to influence the entire network most quickly. Closeness centrality can help find good ‘broadcasters’, but in a highly-connected network, you will often find all nodes have a similar score. What may be more useful is using Closeness to find influencers in a single cluster.

-Eigenvector Centrality: Like degree centrality, EigenCentrality measures a node’s influence based on the number of links it has to other nodes in the network. EigenCentrality then goes a step further by also taking into account how well connected a node is, and how many links their connections have, and so on through the network. By calculating the extended connections of a node, EigenCentrality can identify nodes with influence over the whole network, not just those directly connected to it. EigenCentrality is a good ‘all-round’ SNA score, handy for understanding human social networks, but also for understanding networks like malware propagation.

-PageRank: PageRank is a variant of EigenCentrality, also assigning nodes a score based on their connections, and their connections’ connections. The difference is that PageRank also takes link direction and weight into account – so links can only pass influence in one direction, and pass different amounts of influence. This measure uncovers nodes whose influence extends beyond their direct connections into the wider network. Because it takes into account direction and connection weight, PageRank can be helpful for understanding citations and authority.

- Graph Level: Diameter (longest path, average path length), Size..