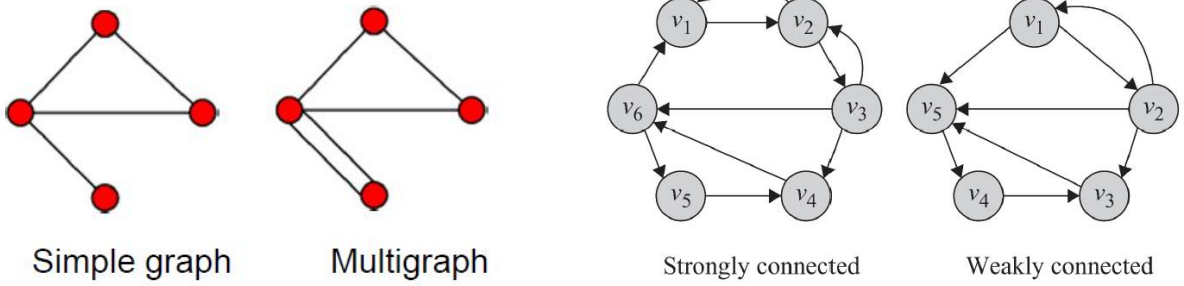
# Maciej Medyk – CAP6776 – Information Retrieval – Social Network Analysis

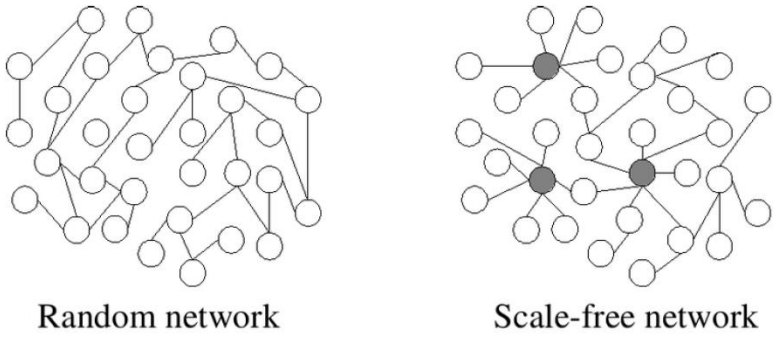
## Report on how to perform analysis of Social Network and analytic tool called Gephi.

---

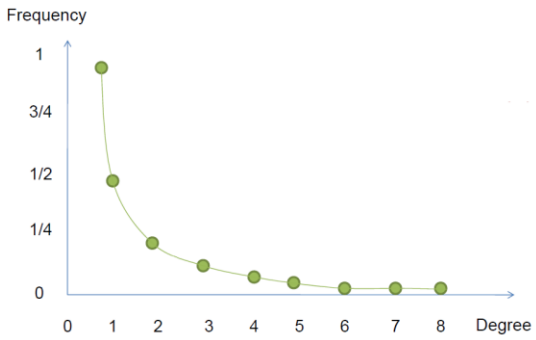
Social Networks became very popular ways for people to interact with each other. It provides ability for people to connect with each other based on friendships, interest, hometown, place of employment, or place of education. As computer scientist, I would like to focus on how social networks are constructed, how they are analyzed, and what tools we can use to analyze a simple social network. Social Networks are build out of nodes which contain all the information about user and are linked to other users through connection links called edges. Social Network can be represented mostly as undirected graph with nodes and edges connecting them. Graphs can be stored in form of adjacency list or adjacency matrix. When evaluating the graph and connections between the nodes we would start with assessing if the graph is a simple or multi graph while evaluating how well connected nodes are with each other (strongly or weakly connected).



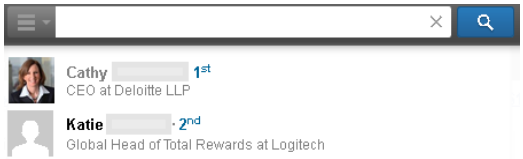

At that moment, we also want to check how the nodes are actually connected and clustered. Within network we can sometimes observe egocentric behavior where nodes consist of focal node that many other nodes are connected to. Assessment of this type would try to evaluate degree of distribution of nodes. Depending on the social network we can either observe a network in which nodes are connected randomly or a network with scale free distribution where some nodes are displaying egocentric behavior.

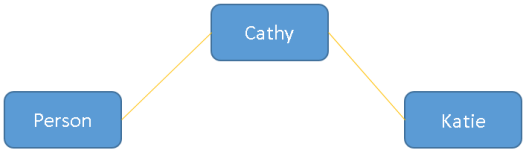



However, certain behavior can often be observed in many networks in which there are very small number of nodes with extremely large amount of connection and very large number of nodes with extemly low level of connections.

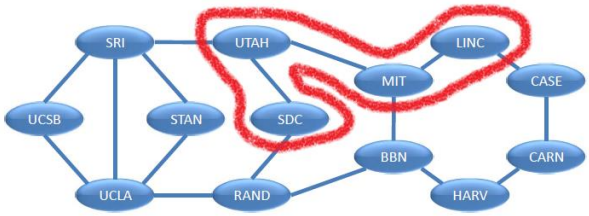


To further evaluate the connectivity between nodes we need to investigate the lengths of the paths between nodes to find shortest paths between nodes and to find network diameter. The shortest path between nodes is simply the smallest amount of edges it takes to travel from one node to the next. This is a statistic often used in link prediction and we can see this behavior in LinkedIn where we see the degree of connections when searching or being adviced on new potential connections. In example below, we can see that person's immediate connections are connections of 1st degree while the friends of person's immidiete connections are connections of 2nd degree.

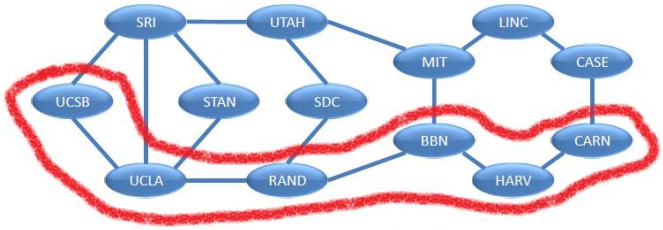



Diameter of the network on the other hand is the maximum shortest path found in entire network. It defines the shortest path between the least connected two nodes. This measurement defines how closely connected are all the nodes within the network.

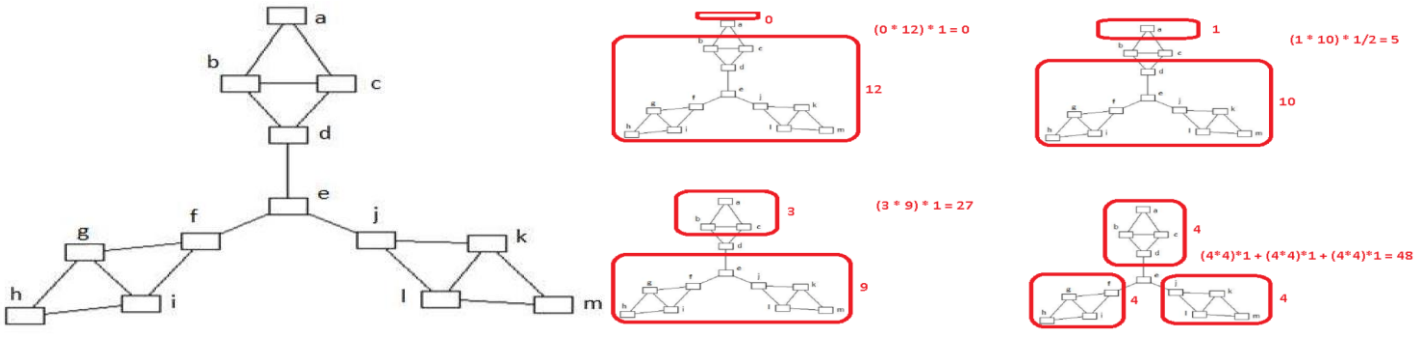



Distance between LINC and SDC is 3.




Diameter is 5.

At this moment. we may want to get deeper in understanding the individual nodes within the network and see the importance of each nodes. Here we can use two measures: betweeness centrality and closeness centrality. Betweeness centrality defines how important the node is in connecting other nodes and the larger the number the more disconnecting effect the node would do. If we look at the graph below, if we would take out node A that would not affect shortest path between any other nodes; however, if we take out node D or E we would decoupe the graph either to two or three separate components. We can easily see that the most crucial node in the graph is node E as it connects the biggest number of nodes and if it would be eliminated then one graph would become three separate graphs. When performing betweeness centrality calculations we can easily see that node E receives the highest score of 48 or 0.727 when normalized. Normalization factor is calculated by 1 / ((# of nodes -2) * ((# of nodes – 1) / 2)) which in this case is 1 / (11 * 6) = 0.0152. Second measure is called closeness centrality which looks at how easy it is to reach other nodes from the node being evaluated. In this case, we also see that node E being central to the graph has the shortest paths to all other nodes on average. The longest shortest path on node E is 3 while for example the longest shortest path for node A is 6. To assess the closeness centrality of each node we have to calculate all shortest paths between all nodes and then sum them up. Node E has 24 total when node A has 47 total. When calculating the score, we would use formula sum / (# of nodes – 1); therefore, score for node E would be 24 / 12 = 2.00 while score for node A would be 47 / 12 = 3.917



Betweenness centrality

| NODE | SCORE | NORMAL |
|---|---|---|
| A | 0 | 0.000 |
| B | 5 | 0.076 |
| C | 5 | 0.076 |
| D | 27 | 0.409 |
| E | 48 | 0.727 |
| F | 27 | 0.409 |
| G | 5 | 0.076 |
| H | 0 | 0.000 |
| I | 5 | 0.076 |
| J | 27 | 0.409 |
| K | 5 | 0.076 |
| L | 5 | 0.076 |
| M | 0 | 0.000 |

Closeness centrality

| | A | B | C | D | E | F | G | H | I | J | K | L | M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 5 | 4 | 5 | 5 | 6 | |
| B | 1 | 0 | 1 | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 4 | 4 | 5 | |
| C | 1 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 4 | 4 | 5 | |
| D | 2 | 1 | 1 | 0 | 1 | 2 | 3 | 4 | 3 | 2 | 3 | 3 | 4 | |
| E | 3 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 3 | |
| F | 4 | 3 | 3 | 2 | 1 | 0 | 1 | 2 | 1 | 2 | 3 | 3 | 4 | |
| G | 5 | 4 | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 3 | 4 | 4 | 5 | |
| H | 6 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 4 | 5 | 5 | 6 | |
| I | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 1 | 0 | 3 | 4 | 4 | 5 | |
| J | 4 | 3 | 3 | 2 | 1 | 2 | 3 | 4 | 3 | 0 | 1 | 1 | 2 | |
| K | 5 | 4 | 4 | 3 | 2 | 3 | 4 | 5 | 4 | 1 | 0 | 1 | 1 | |
| L | 5 | 4 | 4 | 3 | 2 | 3 | 4 | 5 | 4 | 1 | 1 | 0 | 1 | |
| M | 6 | 5 | 5 | 4 | 3 | 4 | 5 | 6 | 5 | 2 | 1 | 1 | 0 | |
| SUM | 47 | 37 | 37 | 29 | 24 | 29 | 37 | 47 | 37 | 29 | 37 | 37 | 47 | SUM(A:M) |
| SCORE | 3.917 | 3.083 | 3.083 | 2.417 | 2.000 | 2.417 | 3.083 | 3.917 | 3.083 | 2.417 | 3.083 | 3.083 | 3.917 | SUM(A:M) / N-1 |
| NORMAL | 0.255 | 0.324 | 0.324 | 0.414 | 0.500 | 0.414 | 0.324 | 0.255 | 0.324 | 0.414 | 0.324 | 0.324 | 0.255 | N-1 / SUM(A:M) |

Another method used to calculate how the degree of connectivity of nodes in social networks is called Graph Edge Density. This method measures the portion of potential connection in the network that are actual connections. Graph Edge Density measures the percentage of completeness of the graph. In complete graph, all nodes are connected to all other nodes. Graph edge density uses formula that calculates (# of edges) / (# of nodes * ((# of nodes – 1) / 2)). In our network, we can see the completeness of the graph being around 23.1%. The more connections we will establish the higher Graph Edge Density score will be.

| ENTIRE NETWORK | |
|---|---|
| EDGE DENSITY | 18 / (13 * (( 13 - 1 ) / 2 )) = 18 / (13 * ( 12 / 2 ) = 18 / (13 * 6 ) = 18 / 78 = 0.230769 |

When evaluating social networks we also want to evaluate completeness of connections between your neighbours. This can be done by evaluating clustering coefficient. Cluster can be defined as triangle and to simply explain if a person has two friends and those two friends are also friends with each other then that would be considered cluster. In evaluation of Clustering Coefficient, we try to understand how many nodes have complete relationships with neighbouring nodes. In our example, social network, we can see that node A has two friends which are B anc C and both of them are also friends. Therefore; A has a 1 cluster and its clustering coefficient is 100%. When evaluating node B, we can see it has three friends, but only two clusters. The cluster friends are A B C and B C D. Potential for existance of another cluster is A B D; however, since out of three potential clusters only two exist the clustering coefficient for node B would be 66.7%. Lastly, if we look at node E which has 3 friends we can see that none of them are friends of each other and clustering coefficient of node E is 0%. Clustering coefficient is calculated by (# of clusters) / ((#number of edges) * ((# of edges - 1) / 2)). To calculate clustering coefficient of the network we average clustering coefficients of all nodes.

| CLUSTERING COEFFICIENT OF NODES | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K | L | M |
| DISTRIBUTION | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 2 |
| CLUSTERS | 1 | 2 | 2 | 1 | 0 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 |
| (DISTRIBUTION - 1) / 2 | 0.5 | 1 | 1 | 1 | 1 | 1 | 1 | 0.5 | 1 | 1 | 1 | 1 | 0.5 |
| COEFFICIENT | 1 | 0.6667 | 0.6667 | 0.3333 | 0 | 0.3333 | 0.6667 | 1 | 0.6667 | 0.3333 | 0.6667 | 0.6667 | 1 |

| CLUSTERING COEFFICIENT OF NODES | |
|---|---|
| COEFFICIENT | ( 1 + 0.6667 + 0.6667 + 0.3333 + 0.3333 + 0.6667 + 1 + 0.6667 + 0.3333 + 0.6667 + 0.6667 + 1 ) / 13 = 0.615385 |

Last method to evaluate social network is calculation of degree distribution of the network. This method describes what is an average expected number of connection per random nodes based on all degrees of all nodes within the network. To calculate this statistic, we have to assess how many edges each node has and sum them up and later divide the sum by number of all nodes. In our example, we can see that we have 13 nodes in which we can see that all nodes have either 2 or 3 edges. After calculation, the degree distribution of the network is 2.77 connections per node.

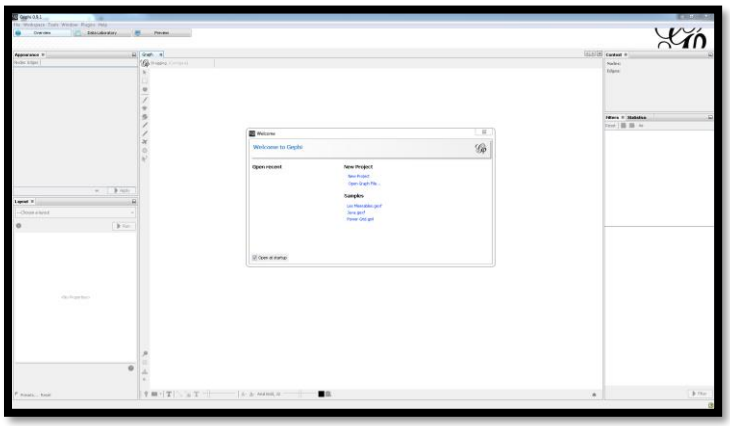
| INDIVIDUAL NODES | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K | L | M |
| DISTRIBUTION | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 2 |

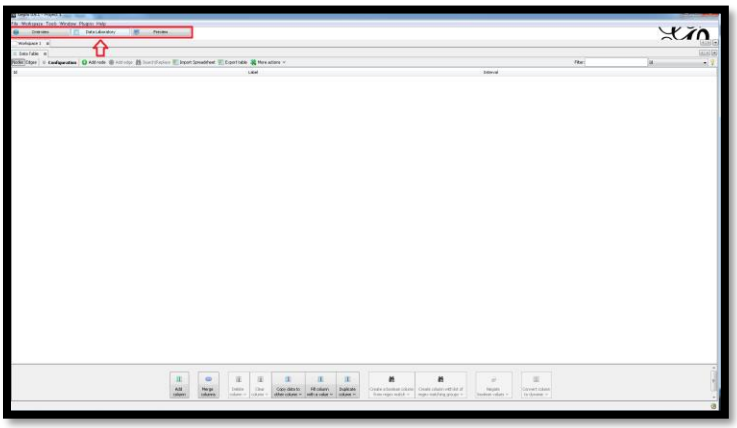
| ENTIRE NETWORK | |
|---|---|
| DISTRIBUTION | (2+3+3+3+3+3+3+2+3+3+3+3+2)/13=2.769231 |

There are many tools that allow for easy evaluation of the social networks and when small networks of 13 nodes can be easily measured by calculations by hand; on the other hand, large networks having thousands or millions of nodes require a program that can do those calculations much quicker and much more accurately. The program that is very easy to use is called Gephi and its obtainable at website link https://gephi.org/ where you can download a copy of the program from the main page. Gephi is available for many operating systems including Windows, Mac OS X, and Linux

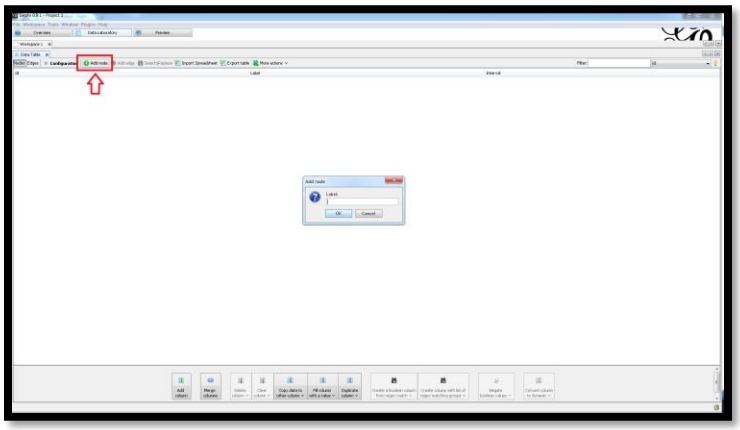


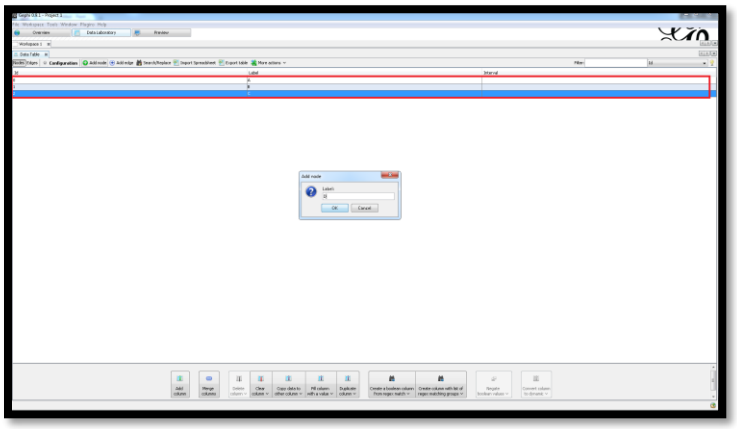
Once you download and finish installation of the program you will be able to start new project or open an existing one. In our tutorial we will assume you will start a new project and add nodes and edges. Once in the program you want to access data laboratory tab on far top left of the screen where you can add nodes and edges.



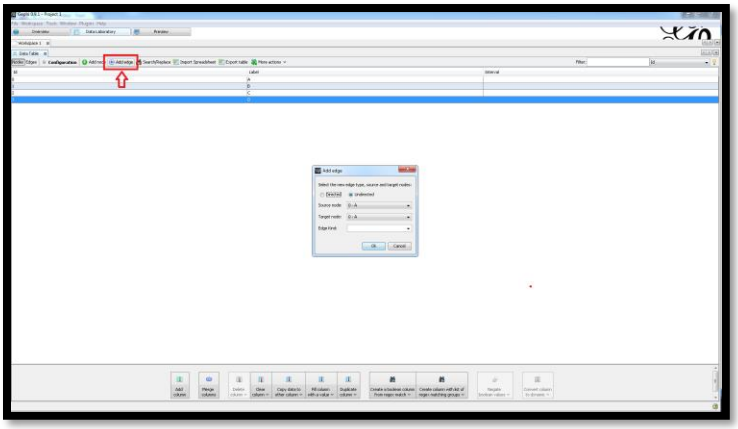


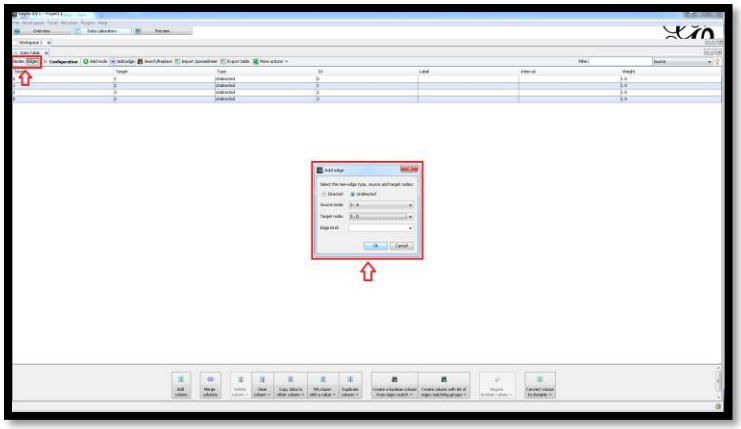
At this moment, you will be able to add nodes first. Once you start adding nodes they will be displayed in the window in the middle. For the purposes of this demo we will add 4 nodes.



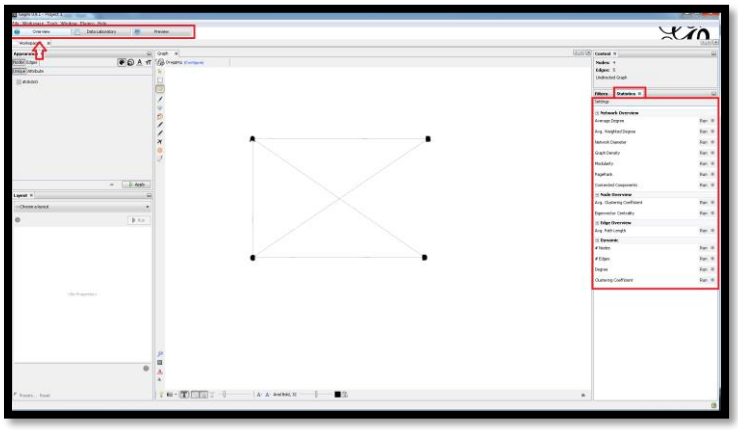


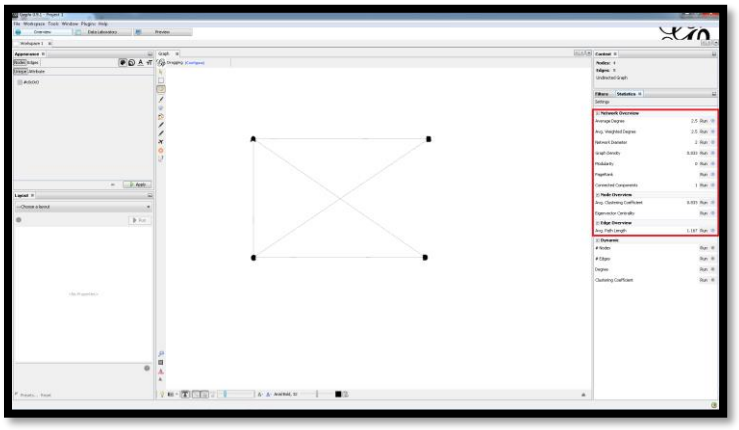
At this moment, we are ready to add edges. Once you click to add edges button you will see a popup box that will ask you if edges are directed or undirected and you need to specify source and destination node. You are also able to switch between views when clicking on either Nodes or Edges tab on far left on the toolbar menu.



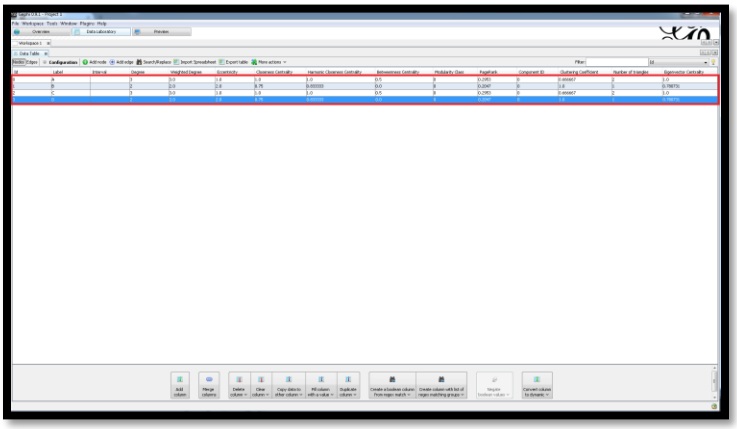


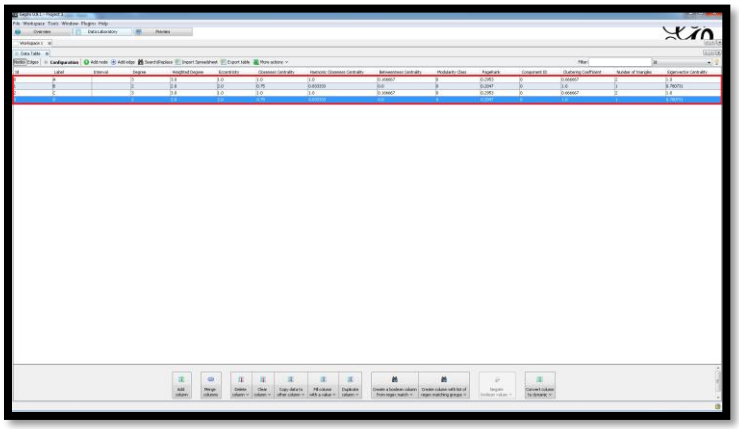
Once you have all the nodes and edges, you can click on Overview to see you graph and do calculations in Statistics panel on far right. Here you will be able to calculate average degree of distribution for the network, netword diameter, graph edge density, clustering coefficient for the network, and both centrality. Betweenness and closeness centrality is accessible on another tab.





To be able to see betweeness centrality or closeness centrality you need to go back to Data Laboratory tab where you will see either normalized or not normalized scores depending on which option you choose when running Network Diameter calculation. Data Laboratory tab will provide you with degree per each node, closeness centrality, harmonic closeness centrality, betweenness centrality, pagerank, clustering coefficient, and eigenvector centrality per each node. In first illustration the scores are not normalized when in the second one I chose option of normalization.





This would be a very basic tutorial on how to get started with Gephi and how to use it in order to perform analysis of larger social networks. Below is the much larger social network with 1589 nodes and 2742 edges. From the analysis we can see that average degree of distribution is 3.451, network diameter is 17, graph edge density is 0.2%, clustering coefficient is 0.878, and we can see that this is not fully connected network but there are 396 separate disconnected graphs here. On second image we can see all the betweenness and closeness centrailities per node along with other attributes.