**Write up for graph statistics of db5**

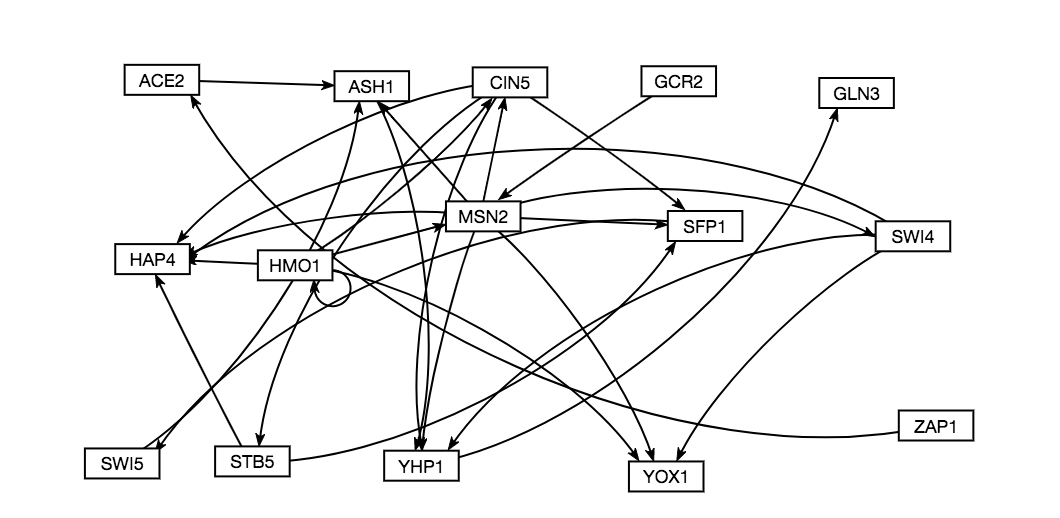


Table 1. Compilation of graph statistics as computed by Gephi for the network db5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Closeness Centrality | Betweenness Centrality | Eigencentrality | Eccentricity |
| ACE2 | 0.5 | 3 | 0.008418 | 3 |
| ASH1 | 0.666667 | 10 | 0.575118 | 2 |
| CIN5 | 0.636364 | 5 | 0.249597 | 3 |
| GCR2 | 0.458333 | 0 | 0 | 3 |
| GLN3 | 0 | 0 | 0.8377 | 0 |
| HAP4 | 0 | 0 | 0.861994 | 0 |
| HMO1 | 0.55 | 0 | 0.11352 | 3 |
| MSN2 | 0.769231 | 14 | 0.121938 | 2 |
| SFP1 | 0.4 | 9 | 0.605438 | 4 |
| STB5 | 0.375 | 0 | 0.248138 | 5 |
| SWI4 | 0.8 | 0 | 0.136077 | 2 |
| SWI5 | 0.5 | 7 | 0.52969 | 3 |
| YHP1 | 1 | 11 | 1 | 1 |
| YOX1 | 0 | 0 | 0.392633 | 0 |
| ZAP1 | 0.4 | 0 | 0 | 4 |

**Closeness Centrality**

Closeness centrality is calculated using the following equation:

Where n = the number of shortest paths going through the node; y = the node in question, x = the node passing through node y.

Closeness centrality calculates the average length of the shortest path between the node and all other nodes in the graph. The more central a node is, the higher the closeness centrality value.

Closeness centrality is an unweighted measure, where the value of ranges between 0-1, where 0 means that no shortest paths pass through the node y, and 1 means the node is fully connected to all other nodes in the graph.

Based on this measure, the following can be determined:

* YHP1 – with a measure of 1, that means every other node has a shortest path going through YHP1. Looking at the graph, this seems like a false measure, as a cursory observation of the path seems like the n-1 might be causing this fraction to be 1.
* SWI4 might be a similar case where the (n-1) value is making the statistic higher as there are not many inputs and outputs going through the node.
* MSN2, CIN5, and ASH1 seem to be most accurate in being central to the graph, based on looking at the number of inputs and outputs for each.

Overall, it seems as though going through and comparing the inputs and outputs for each node might be a good indicator of the closeness centrality value of a given node, since the statistic is unweighted. In looking at closeness centrality in conjunction with other graph statistics, a clearer image of what this statistic might be telling us may become clear.

**Betweenness Centrality**

Betweenness centrality is calculated using the following equation:

Where v is the node of interest, and sigma is the number of paths from nodes st, and sigma(v) is the number of paths from st that pass through node v.

Betweenness centrality calculates how often a node appears on the shortest path between other nodes in the network. The higher the betweeness centrality value, the more it is being used as a stepping-stone from one node to the next.

Betweeness centrality is an unweighted measure, with values being expressed as fractions, or simplified to integers. The higher the betweenness measure, the greater the number of shortest paths that go through the node of interest. One downside to using betweeness centrality as a measure of importance or connectedness of a node is that the measure requires an input and output for each node in order for the measure to be calculated. So a node that might be the start of transcription regulation, and a highly important node in the network would still get a betweenness measure of 0, because there is nothing regulating that node, so it cannot be on the shortest path for any other nodes.

Based on the issue mentioned above, any conclusions or observations based on the betweenness centrality statistic should be viewed through a lens of skepticism, however, from Table 1, the following observations can be made:

* MSN2, YHP1 and ASHI have the highest betweenness centrality measures, which would indicate that they are the most-used hubs in the network.
* With only 7 nodes having a betweeness centrality higher than 0 (MSN2, YHP1, ASH1. SFP1, SWI5, CIN5, and ACE2), this graph has many nodes that are either at the start of regulatory pathways or at the end. With many nodes not being used as a stepping stone between nodes, this also means there are many direct connections between nodes (ie AC, doesn’t need A B C)

Based on these measures, it might make the most sense to only use betweeness centrality to determine which nodes are being used the most as hubs (highest centrality measures) and to determine the structure of the graph, based on comparing in and out degrees for each node to betweenness centrality. Those nodes with no in degree and a betweeness centrality at 0 being at the start of the network, no out degree and betweeness centrality 0 being at the end, and nodes with an in degree and out degree but a betweeness centrality of 0 are not on any shortest paths between nodes. If a network has a high number of nodes with a betweeness centrality of 0 that also have in and out degrees, then that means most edges in the network are direct connections between nodes.

**Eccentricity**

Eccentricity is calculated using an algorithm identifying the ***max{dist(i,j)}*** where *i*  is the node of interest, and *j* is any other node in the network. This algorithm used by Gephi is detailed in a paper by Ulrik Brandes, *A Faster Algorithm for Betweenness Centrality* in the Journal of Mathematical Sociology.

Eccentricity shows how accessible a node is from other nodes, or the distance from the starting node to the farthest node from it in a network. To have a high eccentricity measure means that the node is indirectly connected to other nodes in the network. Nodes with higher eccentricity have a higher impact/influence on other nodes in a network than nodes with a low eccentricity.

Eccentricity is an unweighted, directional statistic which only takes into account a node’s out degree. Eccentricity is expressed as a positive integer, with an eccentricity of 0 indicating that a node has no out degrees. A high integer means the node is highly connected, or has a far reach across the network.

Based on this measure, the following can be determined:

* YOX1, HAP4, and GLN3 are at the end of regulatory pathways, as they have an eccentricity measure of 0, and no out degree.
* STB5 has the highest eccentricity, which means it has the farthest reach, or influence across the network. Based on the network, this does seem to be a fair conclusion to be drawn.
* Most nodes in the network seem to have an eccentricity of 3, which would indicate the reach of nodes is about the same across the network (regardless of what the eccentricity value is, this conclusion could be drawn)

It seems as though the best utilization of eccentricity would be to compound it with an analysis of a weighted network, to determine what type of regulatory relationships are at play in reaching the furthest across the network. If the node in question has a strong negative influence, then there is a large negative reach across the graph. If the node has a strong positive influence, then many nodes would be impacted by the activation by that node.

**Eigenvector centrality**

Eigenvector centrality is calculated by using the adjacency matrix, where the relative centrality of a vertex *v* is defined as:

Where = 1 if vertex *v* is linked to vertex *t* and = 0 otherwise. *M(v)* is a set of neighbors of vertex *v* and is a constant.

Eigenvector centrality measures the influence of a node in a network. The measure assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question thank equal connections to low-scoring nodes. Gephi offers the option to run any number of iterations on Gephi, with the number of iterations automatically set to 100.

Eigenvector centrality is an unweighted statistic, with values between 0 and 1. Based on playing with the iteration count, it appears as though this centrality measure acts similarly to a limit, with the higher number of iterations causing the statistic to reach higher values. A high eigenvector centrality would indicate that the node in question has a high level of influence over the graph. This statistic is a node based measure rather than one that is indicative of edge importance, so it might prove interesting to look at this node in relation to

Based on the above, the following conclusions can be drawn:

* Nodes with a low eigenvector centrality seem to be those that are deemed by other statistics to be the “most central to the graph,” or the starting of regulatory pathways. This might indicate that the level of “node importance” may be partially based on a relationship between in degrees and out degrees shown in the network.
* Nodes with a high eigenvector centrality seem to be those that have fewer out degrees, and thus are more important because more nodes are regulating them.

The eigen vector centrality measure does not seem to be particularly informative to investigating our networks. What is interesting from the graph, is that it might indicate where regulatory pathways in networks end, as the highest measures are those with no out degrees.

What these statistics show overall as well is that there seems to be something odd going on with the node YHP1. There is a statistic of 1 seen for almost all measures, which would indicate that there might be an error generated for most of the centrality measures based on the calculations being done for each statistic. This relationship will be investigated further, and is something to keep in mind before judging the importance of YHP1 in the network.