Exercise 2 – Filtering and selecting nodes/edges

Data

For this exercise we are going to use a network consisting of a subset of human protein-protein interactions published by Stelzl, Ulrich, et al. "A human protein-protein interaction network: a resource for annotating the proteome." Cell 122.6 (2005): 957-968.
It is a small subset of a larger human interaction dataset consisting of protein-protein interactions, some of them supported with topological and GO criteria, and some of them verified biochemically.
Nodes in the network are labelled by numeric Entrez IDs, which are the IDs employed by NCBI (www.ncbi.nlm.nih.gov).
Interactions in the network have evidence from varying sources: high-confidence yeast two-hybrid interactions (lacz4) and low-confidence yeast two-hybrid interactions (sd4), coAP and GST pull-down interactions, and supporting evidence of interaction reported in literature.

The data is provided in tree files:

- **STELZL.sif** ➔ contains interactions
- **STELZL.txt** ➔ contains the HUGO gene symbol associated to the proteins

Load the network and attributes

Load the network and attributes into Cytoscape.

Hint: Menu bar -> File -> Import -> Network > File...
Menu bar -> File -> Import -> Table -> File...

Remember when importing the attributes that the "shared name" column is always string. When you try to import the attributes, Cytoscape will consider the Gene ID a number and you will get an error ("Types of keys selected for tables are not matching"). To fix it you need to force the Gene ID field to be considered as a string.

Hint: Right click on "Gene ID" and select "String"

Filtering

Inferred gene networks tend to be "hairballs" and it is necessary to clean them in order to extract the relevant information.

In this part of the exercise we will clean the STELZL dataset removing the low confidence edges (SD4), the self-loops and the duplicated edges.

Select the high confidence interactions.

Create a new filter named "high-confidence" and configure it to select every edge that is not of SD4 type.

Hint: Left panel -> Select -> Create new filter
Left panel ->+ -> Column Filter

Next, select also the nodes linked by those interactions.

Hint: Menu bar -> File -> Select → Nodes → Nodes connected by selected edges

Finally, create a new network containing the selected nodes and edges.

Hint: Menu bar -> File → New → Network -> From selected nodes, selected edges

At this point you should have a new network called STELZL.sif(1) with 1139 nodes and 2038 edges.

Remove self-loops

Have you notice the nodes with self-loops? Filter those out and make sure that you select the STELZL.sif(1) network so other networks in Cytoscape are not affected.

Hint: Menu bar -> File -> Edit -> Remove self-loops

49 edges should be removed.

Remove duplicated edges

Have you notice the duplicated edges? Sometimes they might be important but some times they can be removed. Filter those out.
Once again, make sure that you select the STELZL.sif(1) network so other networks in Cytoscape are not affected.
The network is undirected so don't forget to tick the "Ignore edge direction" box.

Hint: Menu bar -> File -> Edit -> Remove duplicated edges

143 edges should be removed

Compared to the network you started with, the network you have now has fewer edges, but all the edges are determined through by higher-confidence experimentation or by literature-based methods. For some types of analysis, this is a more appropriate set of edges.

Choosing nodes based on adjacent nodes

It is often important to select a node or an edge based on properties of the nodes it connects to, not based on the properties of the node or edge itself. For the sake of illustration, the STELZL data has a column tagging a node as "Suspicious", as judged from some process outside of Cytoscape.

In this part of the exercise we will identify "non-suspicious" genes connected to at least three "suspicious" nodes in three steps:

Select all suspicious nodes and any nodes they connect to

Similarly to what we did above, create a new filter called "Suspicious" that selects suspicious nodes (nodes where the suspicious attribute is "true").

Hint: Left panel -> Select -> Create new filter
Left panel ->+ -> Column Filter

Use select menu to select the adjacent edges and neighbors nodes.

Hint: Menu bar -> Select -> Edges -> Select adjacent edges
Menu bar -> Select -> Nodes -> Nodes connected by selected edges

Create a new network from the selected nodes and edges.

Hint: Menu bar -> File -> New -> Network -> From selected nodes, all edges

Remove non-suspicious nodes connected to less than three suspicious nodes

Create a new filter called "Relevant-non-suspicious" that will contain:

- A degree filter that selects node with a degree between 0 and 2
- A column filter that selects nodes that are non-suspicious

Remove the selected nodes.

Hint: Menu bar -> Edit -> Delete Selected Nodes and Edges

Remove suspicious nodes connected only to other suspicious nodes

Similarly to what was done above, create a filter called "Non-suspicious" that select all the non-suspicious nodes

Select the adjacent edges and the nodes connected by those edges.

Hint: Menu bar -> Select -> Edges -> Select adjacent edges
Menu bar -> Select -> Nodes -> Nodes connected by selected edges

Create a new network from the selection.

Hint: Menu bar -> File -> New -> Network -> From selected nodes, all edges

The final network has only non-suspicious nodes that connect to three or more suspicious nodes, and the suspicious nodes they relate to.