Let’s list the pros and cons of the two different ways we currently have of returning graphing info from R

# Option 1: Correlation Matrix

* This method corresponds to the canonical way of storing a graph as a matrix.
* Does not allow us to know neighbours and is pretty much useless for creating a custom layout such as clustering.
* Perhaps we could extend this method to return additional information in addition to the correlation matrix. This info could be lists of which nodes are first neighbours and which are second neighbours (based on the genes of interest selected in the main tab).
  + The issue with this approach is that we need to hardcode the reading in of more properties from the R output. Our current method on the server called getWEightsAndDegreesFromROutput already has quite a few layer of reading to it
  + Suppose that the new info we are returning from R are lists of genes indicating which genes are first neighbours and which genes are second neighbours.
  + Furthermore, we would still have to create some sort of O(n^2) method on the server side that iterates through our nodes and checks if they are in one of these lists. If so, a node is positioned accordingly. Mind you that our method that creates all of the edges is already O(n^2) so if we were to stick to this correlation matrix approach, it isn’t to big of a deal to check whether or not our nodes are first or second neighbours

# Option 2: Lists of Weights and Degrees

* This method has some added complexity since generating edges out of it is not entirely trivial.
* Currently, we are only using this method when we want to dissect the neighbours of a selected node. So we pick a node and we get its neighbours, then we can pick one of its neighbours and get that node’s neighbours. When creating the edges, we know which nodes to create the edges from since those are the nodes that the user selected. We also know which nodes to make the edges go to since those are the weights and corresponding names that the R script returns for each level of neighbours.
* This starts to break down after the third or fourth level since one of the nodes that we need to draw an edge from could have been a couple of levels back.

# Option 3: Return Edges and Degrees from R

* An option we have yet to explore is returning edges from R. These edges are basically triples indicating source, target, and strength. On the server side, we could parse these and easily create cytoscape edges from them.
* The degrees we would be returning from R would be the degrees of the selected nodes and their neighbours. The question remains, is degree referring to their degrees in the chosen subgraph or their degree in the overall graph.

For the sake of being consistent and having general methods as opposed to specialized methods for each little thing, we should choose one of these approaches and stick with it. I think that we will stick with the first choice for now. The reason being that it’s probably better to not have to create a list for every single edge in R. What we can do is return a list of lists when it comes to creating the nodes. The first sub-list will be our selected genes, the second sub-list will be the first neighbours, and the third sub-list will be the second neighbours. This will not be difficult to do four our getRelevantSubmatrix script, but it might be a bit harder to do for our findCorrelations script.

It turns out that implementing the findCorrelations script by having it return a correlation matrix gets complicated. The complication arises due to not knowing if we should index into a column or row. We managed to get past this, however, the matrix returned contains some values that should be zeroed out. Everything that is not a column or row that was specified should be zeroed out.

For example, we specified c(“TBP-E”, “UBE2C-S”, “STAT3-E”) and got the following matrix:

UBE2C-S TBP-S EBNA1BP2-S CDC45-S TFDP1-S

TBP-E 0.4943829 0.6379173 0.5289383 0.4917726 0.0000000

STAT3-E -0.4870864 0.0000000 0.0000000 0.0000000 -0.4615142

VPS72-E 0.4930360 0.0000000 0.0000000 0.0000000 0.0000000

UBE2C-E 0.7949256 0.0000000 0.0000000 0.4967731 0.0000000

EBNA1BP2-E 0.5264976 0.0000000 0.7561553 0.0000000 0.0000000

CDC45-E 0.5020195 0.0000000 0.0000000 0.8033812 0.0000000

However, what we want is the same matrix but with only the first two rows and first column allowed to have non-zero entries. This does seem like a somewhat flimsy solution though. Let’s see if it solves the problem that occurs when doing past the second level. It does seem to solve the issue since for selected genes of: selectedGenes = c('VPS72-E', 'UBE2C-S', 'TBP-E', 'CDC45-S')

We get the following back

UBE2C-S CDC45-S VPS72-S TBP-S EBNA1BP2-S

VPS72-E 0.4930360 0.0000000 0.6086204 0.0000000 0.0000000

TBP-E 0.4943829 0.4917726 0.0000000 0.6379173 0.5289383

UBE2C-E 0.7949256 0.4967731 0.0000000 0.0000000 0.0000000

EBNA1BP2-E 0.5264976 0.0000000 0.0000000 0.0000000 0.7561553

CDC45-E 0.5020195 0.8033812 0.0000000 0.0000000 0.0000000

STAT3-E -0.4870864 0.0000000 0.0000000 0.0000000 0.0000000

SUZ12-E 0.0000000 0.4968252 0.0000000 0.0000000 0.0000000

For our new method of getting neighbours that involves returning edges as opposed to retuning a correlation matrix, let’s specify the algo and make sure that it doesn’t suffer from the same issues as the previous way.

The point is that we need to ensure that the list of degrees we are returning is unqiue, i.e doesn’t have duplicate values.

What we actually need to do is ensure that for our function that generates edges, don’t take any exclusions into account.

Looks like the only issue we’re having right now with the method that returns edges from R is that we have duplicate edges for the selected genes at each level. This makes sense since we aren’t usingand exclusions in our edges creating function in R. Let’s try using exclusions to see if that works.

Looks like the following path fails:

Clear

VPS72-E

UBE2C-S

TBP-E

CDC45-S

SUZ12-E

PPIC-S

EGFR-E

At step EGFR-E due to the getDegreesFromNeighbourNames not working properly. Probably has to due with the fact that the gene has only degree 1.

Fixed this issue. The new method seems to be working very well for the neighbor exploration tool. Now we need to stardardize the getRelevantSubmatrix method. We need to add exclusions to this tool somehow. Say we pick a single gene. WE get its first neighbours, then we get the neighbours of those first neighbours. We need to exclude that selected gene when getting the second neighbours since we already have edges coming from the selected gene to the first neighbours, we don’t want to potentially repeat any of those edges when getting the second neighbours.

We also need exclusions for the nodes. Say we pick gene A and gene B as genes of interest. WE need to ensure that we are not making duplicate nodes for their neighbours. I.e. if gene A has neighbours C1, C2, C3 as the top three interactions, and gene B has neighbours C1, D1, D4 as the top three interactions, then we don’t want to have two instances of C1 in the graph. So as we’re creating the neighbor names and the degrees of those neighbours, exclusions should be any previous neighbours. So when getting the neighbor nodes for gene B, we should be exluding all of the previous nodes we saw for gene A: C1, C2, C3.