Brandon Barrett
25 November 2019

1. **For the "small" data set, for each correlation threshold, plot the degree distribution**

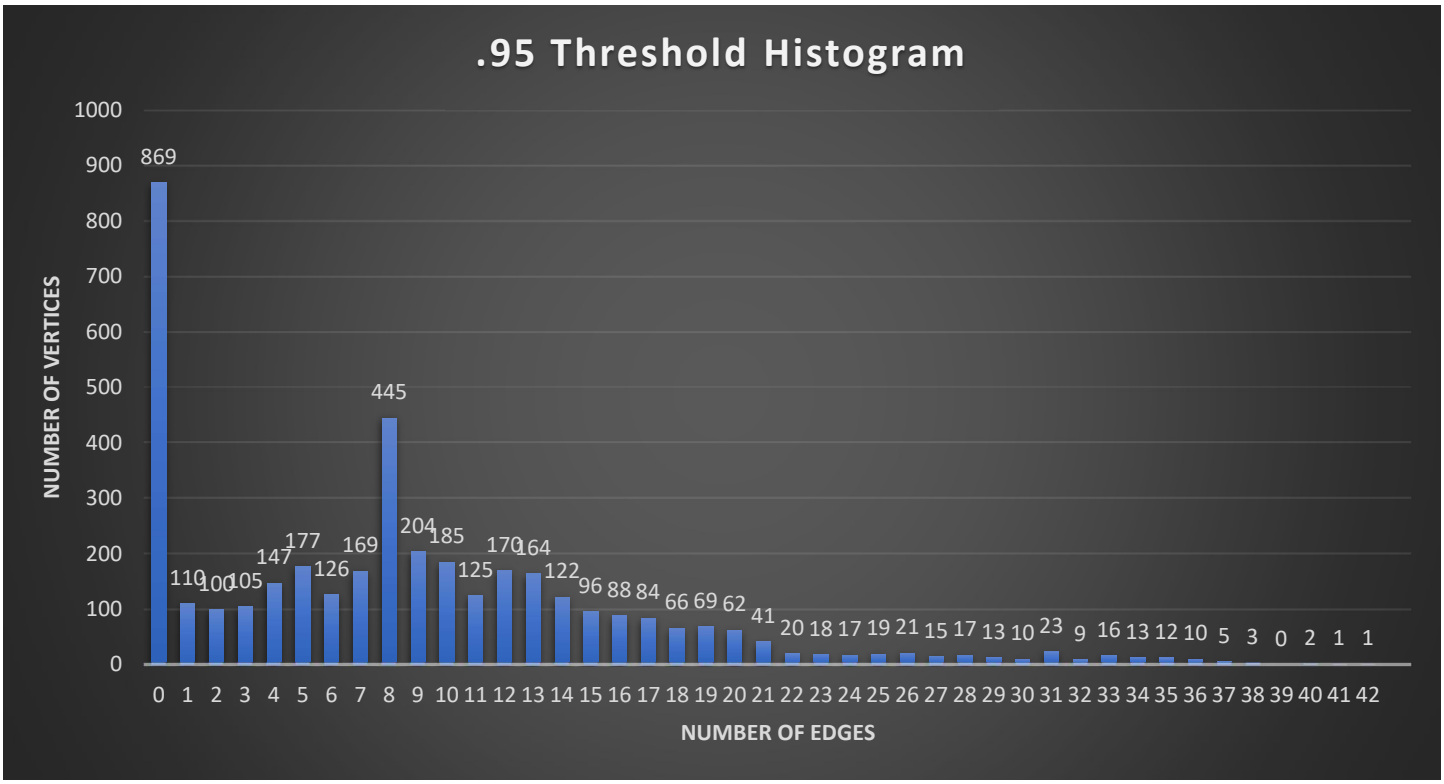

Figure 1: Graph of .95 Threshold with the number of nodes per edge

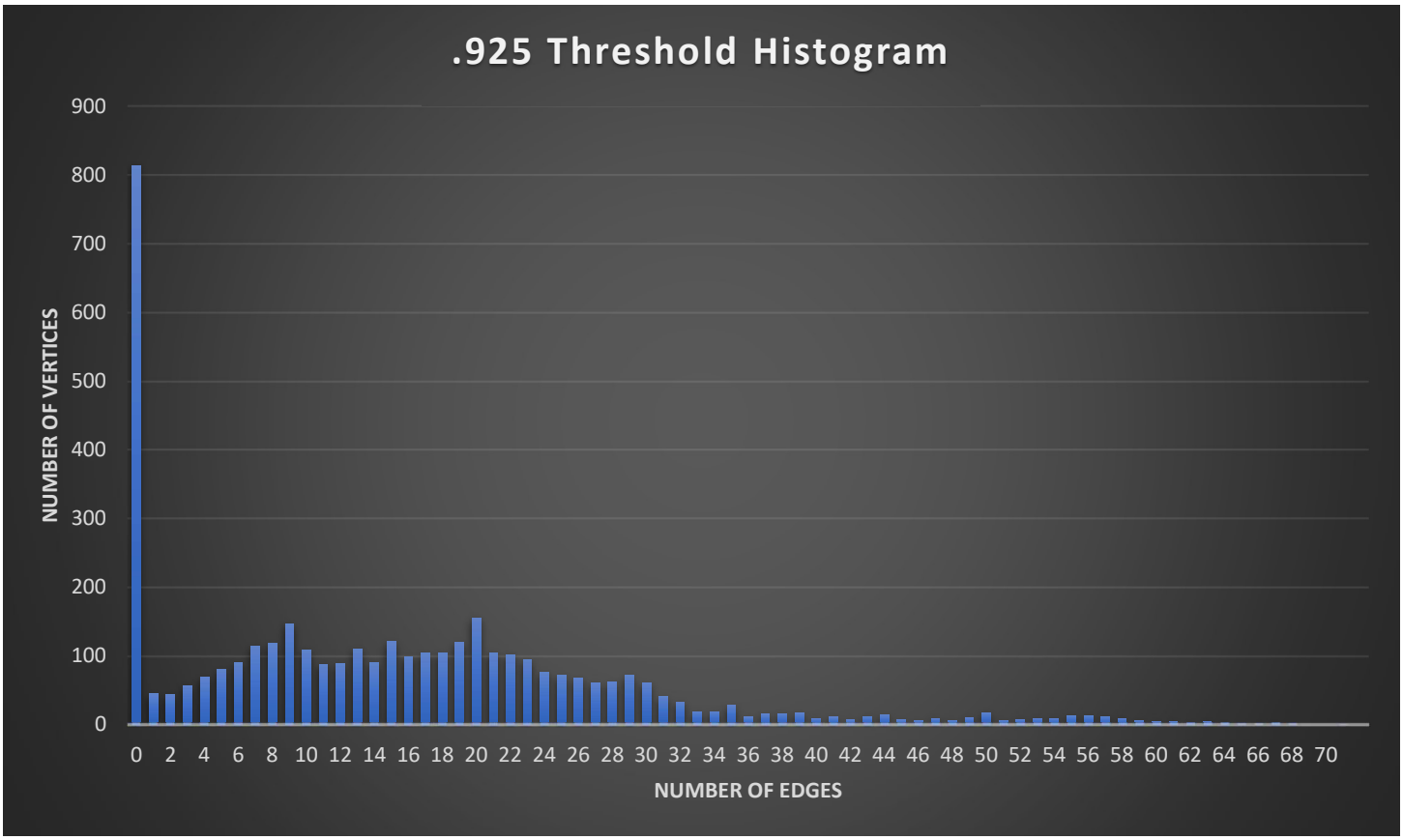

Figure 2: Graph of .925 Threshold with the number of nodes per edge

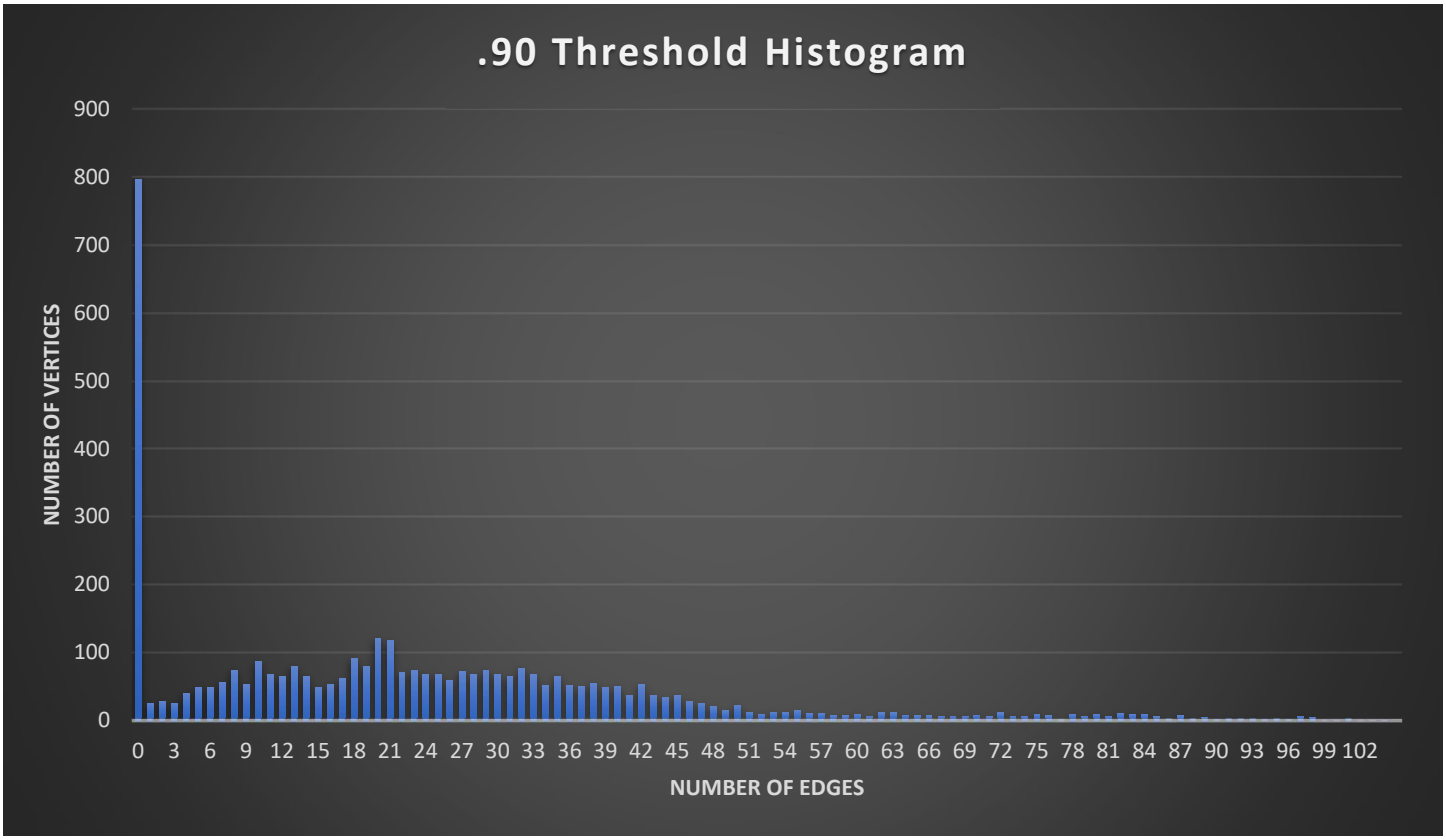

Figure 3: Graph of .90 Threshold with the number of nodes per edge

The results per the histogram seem to support the data's distribution of a small-world graph. There is a clear long-tail distribution present in all three of the histograms (figure 1, 2, 3). In this context, long-tailed meaning that that most of the data is clustered on the left half of the graph and several other results significantly smaller in number trailing off on the right side.

Of important note is that the vast majority of nodes with 0 edges at any of the three tested thresholds are the 783 land-based nodes. Most every other node has at least 1 edge besides those results. Because these results are not obviously visible on the graphs compassing such a large range of edges, I have included the raw data in the project document to verify the results.

On the opposite side of the spectrum to these completely disconnected land nodes are the super nodes which have more edges than any other node. For the .95 and .925 thresholds these were the same nodes located just off the shore in the North central section of the 63x63 grid. This was an area located in the middle of the sea away from most of the landmasses present. It's likely this produced more consistent sea ice results across all the years along with surrounding cells.

For the .90 threshold these super nodes were found in roughly the same area, but slightly more North and East relative to the rest of the map. Again, this was another area in the middle of the sea without any nearby landmasses and this likely helped produce consistent results across the years.

The consistency of these super nodes can be validated by the correlation coefficient formula used to calculate them which attempts to connect nodes with similar outcomes over time.

2. **For the "small" data set, for each correlation threshold, compute the number of connected components in G_r and their size.**

For threshold .95 I would expect to see many forests present in the small-world graph. The vast majority of these would likely be very small, maybe 0, 1 or 2 connected vertices per forest. With a few larger forests encompassing hundreds of vertices that all have very similar correlation coefficients. The .95 threshold is a very hard one to break into meaning that only those vertices that are nearly perfectly similar will be connected leading to more disjoint graphs.

For the .925 and .90 thresholds I would expect to see a lot of 0 connected vertices from the landmasses and then 1 or 2 super large trees with thousands of vertices each. The Although the difference in threshold from .95 to .925 isn't a steep one it is very significant in a square area only 315km x 315km which is already predisposed to be highly similar.

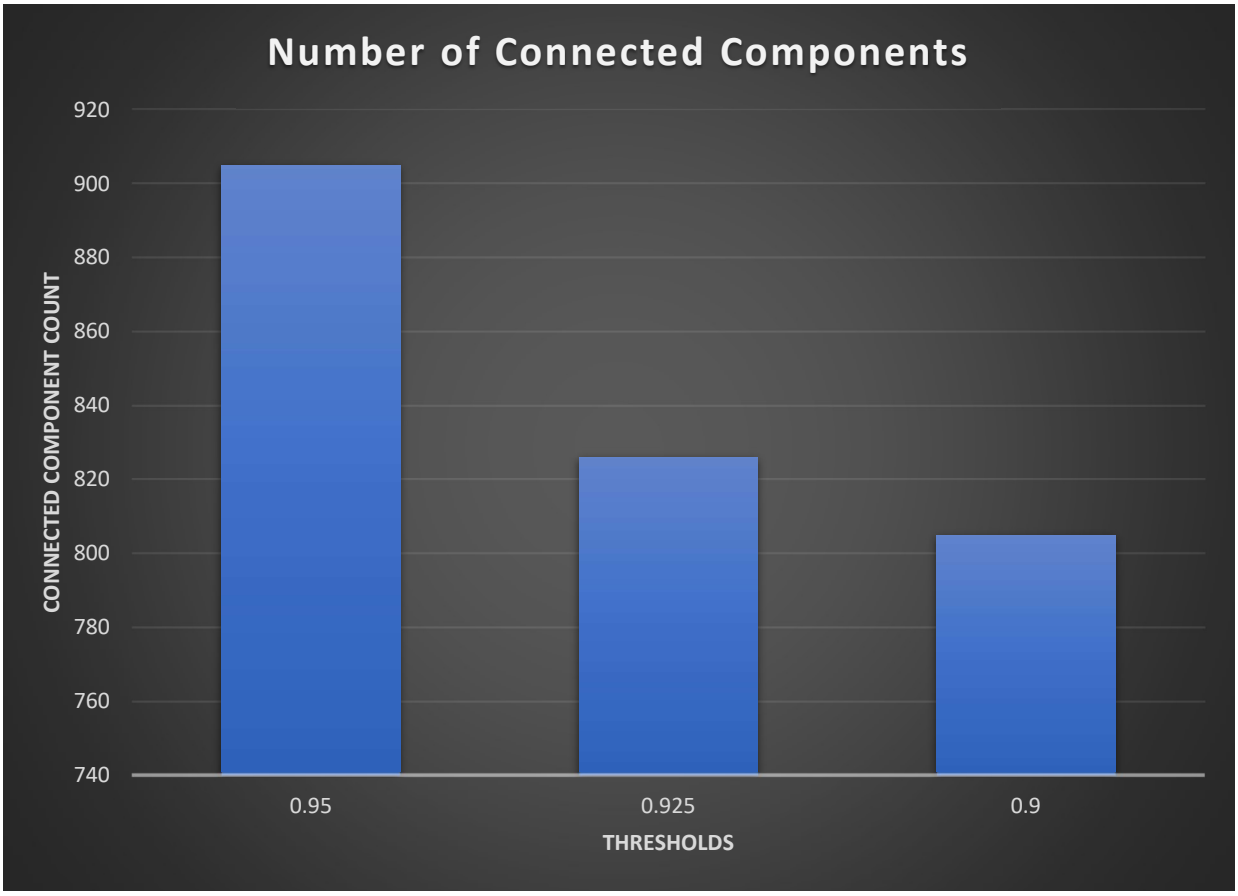

Figure 4: Number of Connected Components by Threshold

It appears that my hypothesis was mostly supported by the raw data provided. When the clusters were graphed the result is what appears in figure 4. There are significantly more connected components in the .95 threshold and then .925 and .90 show very little overall change indicating that the graph is relatively well connected at that point.

If the 783 land vertices are removed from the graph as well it becomes even more clear just how well connected the system becomes (figure 5). The number of small separate clusters diminishes greatly, and we're left primarily with a couple very large trees and some relatively small trees as outlying forests.

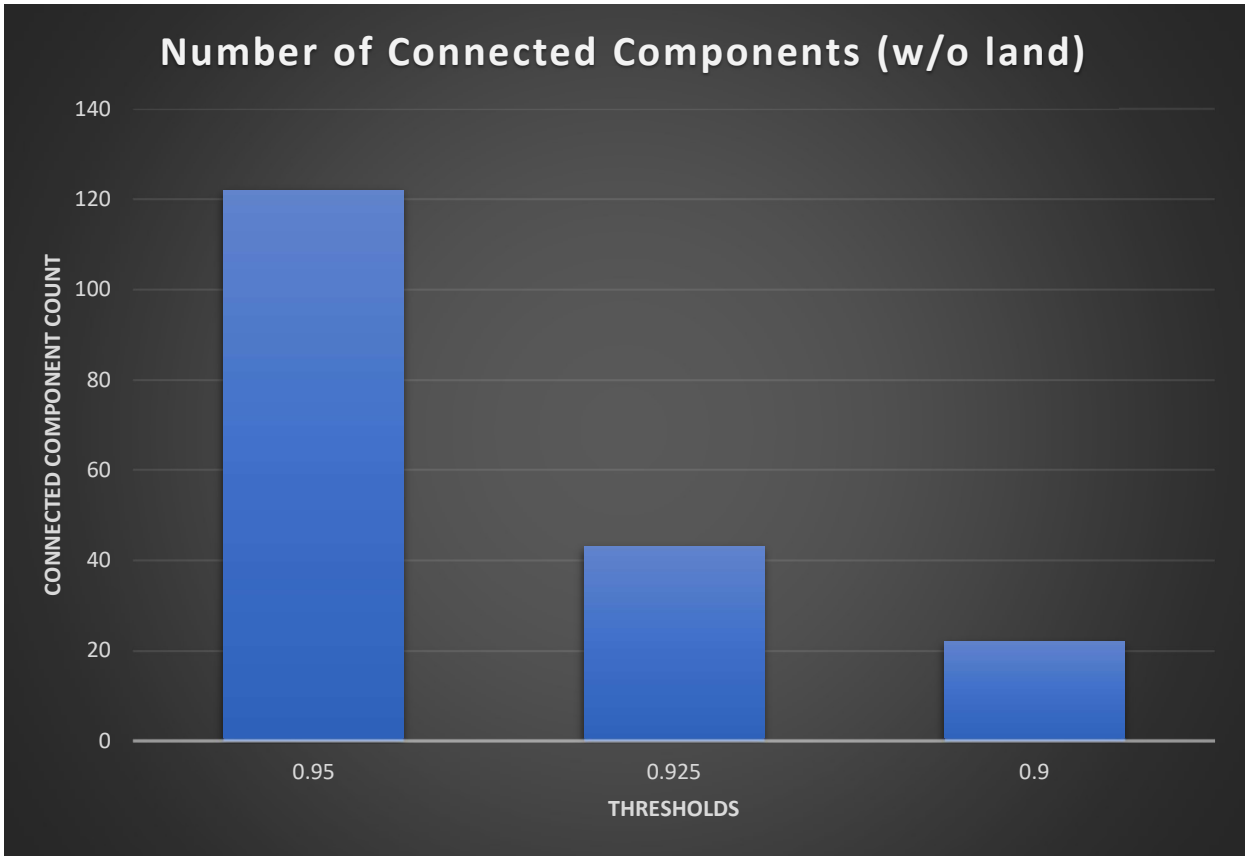

Figure 5: Number of connected components by threshold, excluding land vertices

Judging the individual sizes of the components this trend of increasing connectivity becomes clearer (figure 6). At threshold .95 there are a lot of single vertices making up their own graphs with a few larger graphs and one very large graph. At thresholds .925 and .90 there is very little change in the connectivity between the two, at this point the graphs are very well connected with a single large graph dominating the components.

Thresholds:	**0.95**	**0.925**	**0.9**
Size of Component	**Number of Components**		
1	869	814	797
2	18	4	3
3	8	4	3
4	4	3	0
5	1	0	0
6	1	0	1
7	1	0	0
30	1	0	0
88	1	0	0
2888	1	0	0
3123	0	1	0
3151	0	0	1

Figure 6: Table of the component sizes by threshold

3. **Compute the clustering coefficient and compare to the random graphs of equivalent thresholds.**

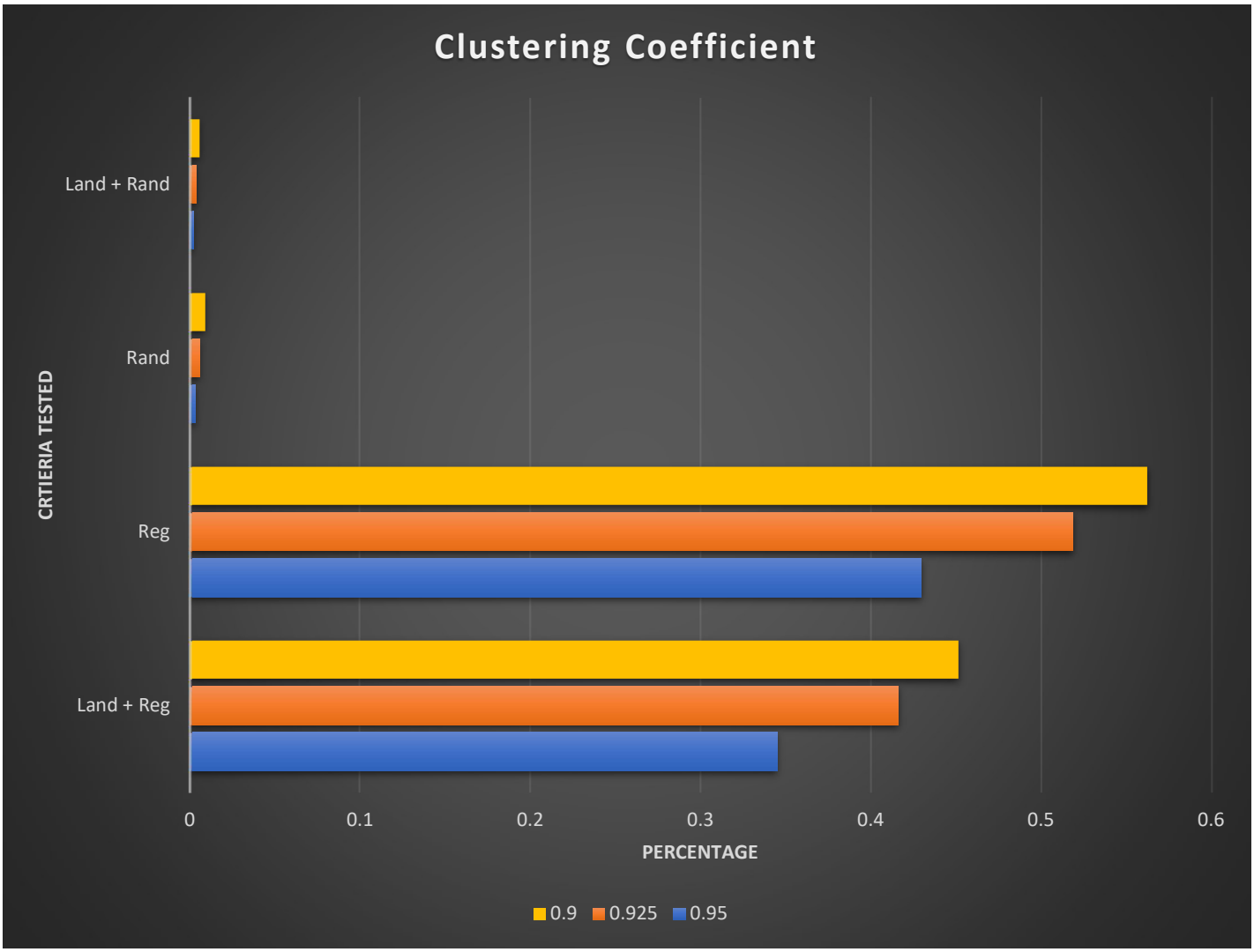

Figure 7: Chart of clustering coefficients by their criteria

From the clustering coefficient it's very plain to see that the sea ice data used in the project is very very well connected (figure 7). At even a .95 threshold there is a clustering coefficient hovering around .35 to .43 depending on if land is included as part of the evaluation. What this indicates is that the nodes are very well connected to each other and can often be found in loose groups even at this relatively high bar for a related system.

This trend carries over to the thresholds of .925 and .90 which become increasingly more connected to each other in tighter groups. This indicates that the sea ice of the Beaumont Sea is consistent from cell to cell and that local clusters are very similar across the broad map of the region.

This is plotted against the random graphs generated from the same number of vertices. These graphs on average are very likely not well connected to each other as per their very low clustering coefficient hovering below .01 even for the loosest connection. This comparison between the random graphs and the sea ice graphs serves to support the idea that the Beaumont Sea cells are very well connected since they are orders of magnitude more clustered together. It also shows that the random graphs are likely generating very disconnected systems, so perhaps an algorithm with a more connected system outcome should be considered for a more accurate comparison.

4. **Compute the characteristic path length and compare to the random graphs of equivalent thresholds.**

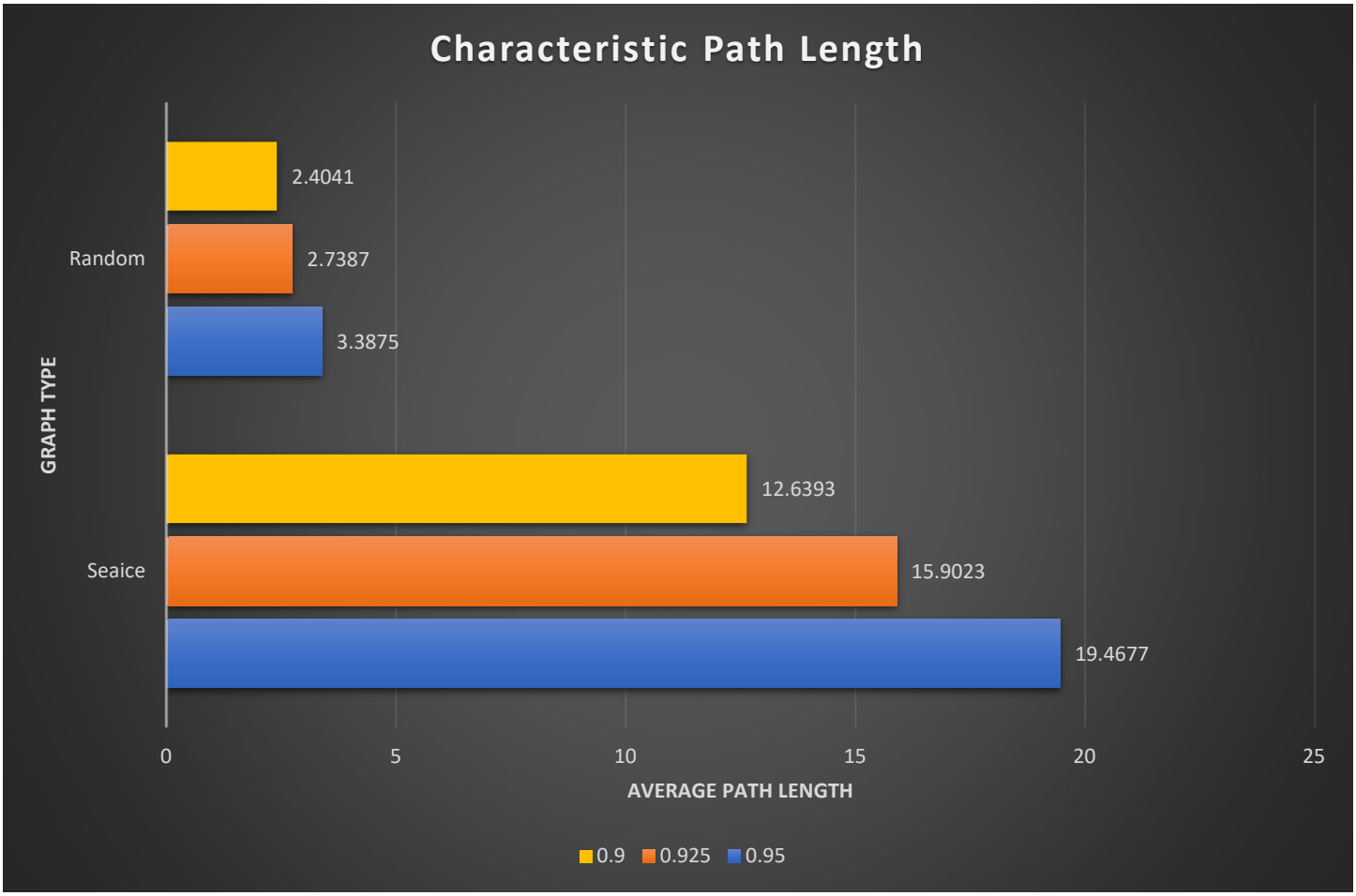

Figure 8: Graph of the characteristic path lengths of the sea ice graphs and random graphs

Like the previous section the characteristic path length also indicates how well connected the sea ice graphs are (figure 8). The characteristic path length for the sea ice graphs is on average about 5 times longer than the random graphs in every single test case. This indicates that the sea ice graphs tend to be connected along most of the system; that they aren't divided into many little forests which would create the shorter characteristic paths seen in the random graphs.

This result can be verified by looking at the path lengths for each node as well. Even as the random graphs have the same number of edges of the same weight as the sea ice graphs, because they are not well connected many of these edges are isolated into small forests rather than extending across a giant forest as seen in the sea ice graphs.

However, this metric is not completely fool proof. If we were to continue lowering the threshold the sea ice graphs would eventually have a near identical characteristic path length to the random graphs at a value of around 1, the weight of each edge. The only reason this discrepancy exists is because high thresholds are being used and that we're solely attempting to determine how well connected the sea ice graphs are via these thresholds.