Cutoff Exploration

frederic.guintu

November 2024

1 Introduction

To update on my progress for the week, I explored the network of the edges provided by the cutoff chosen last week to see if the small fraction of data would provide valuable information to cluster with. Additionally, I also looked at the sampling of the data by varying the cutoff to find a better threshold.

Figures 1 and 2 show the result of the data above when varying the cutoff. Figure 1 shows the number of pairs from both the real-real/real-shuffled distributions above each cutoff. Figure 2 shows data for the real-real distribution above the varying cutoff, where the blue line shows the percentage of the real-real data out of all the data above the cutoff and the orange line shows the fraction of real-real data above the cutoff over all of the real-real pairs.

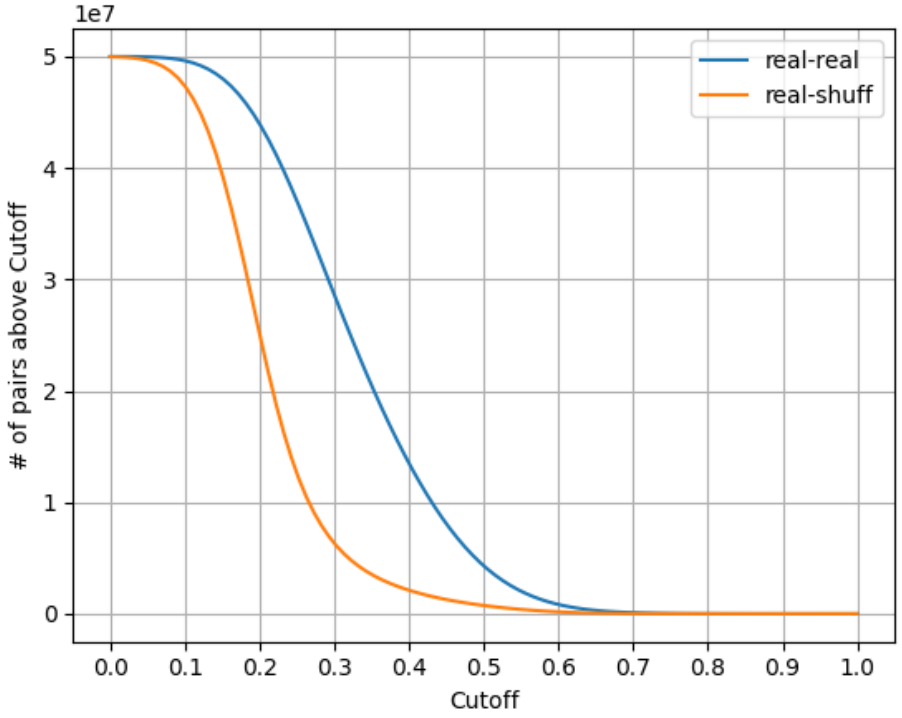

Figure 1: Number of pairs from each distribution above cutoff.

If we decide to switch the cutoff to include more real paired data, I think a value between 0.4 to 0.6 as the cutoff would be good as it maintains a balance between the majority of data above the cutoff being from the real protein pair distribution and on average 10% of the total pairwise data being kept.

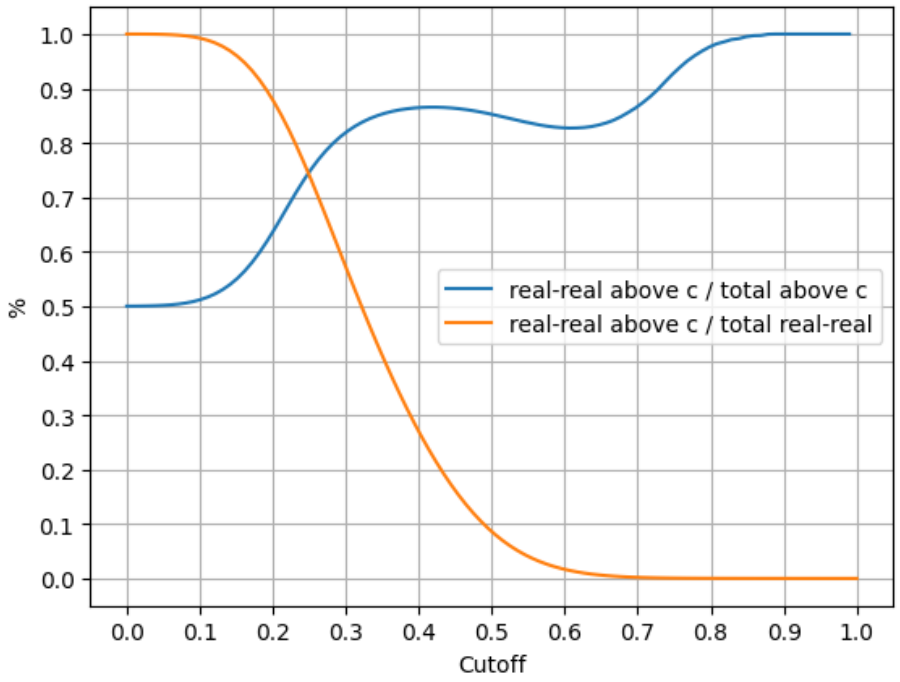

Figure 2: Percentages of real-real data above the cutoffs.

The next four figures show the network analysis of the real protein pairs above the cutoff of 0.77. To reiterate, this cutoff produces the following:

```
Cutoff = 0.77
% of real-real above c out of total above c = 95.16%
% of real-shuf above c out of total above c = 4.83%
% of real-real above c out of total real pair = 0.0411%
```

Figure 3 is a graph visualization of of the nodes and edges above the cutoff. It features many single pair connections with fewer smaller clusters. Further analysis in figures 4-6 of the degrees of the vertices show that the lower half of the data has degree less than three with the other half being above 3, with outliers of highly connected vertices.

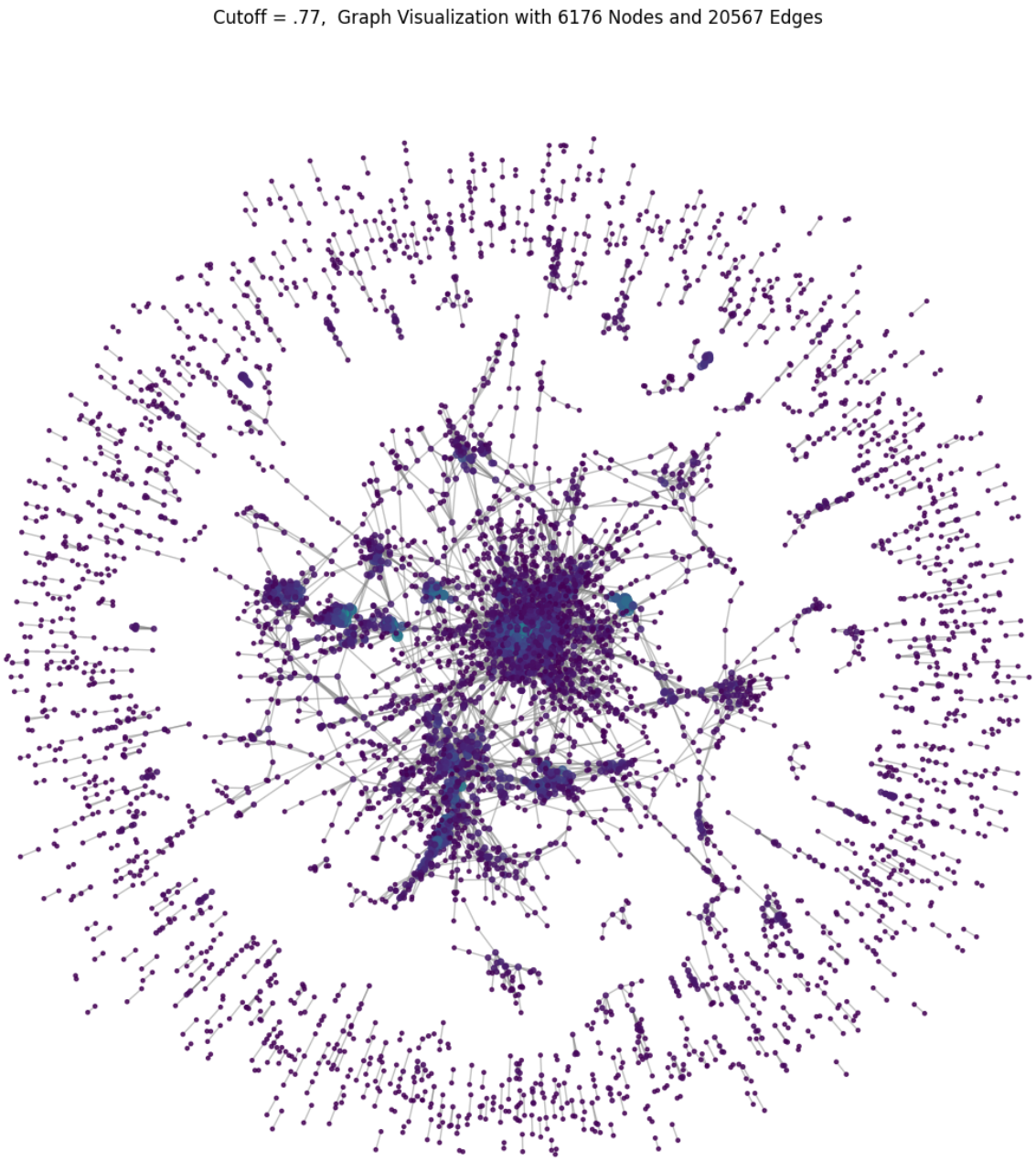
Cutoff = .77, Graph Visualization with 6176 Nodes and 20567 Edges

Figure 3:

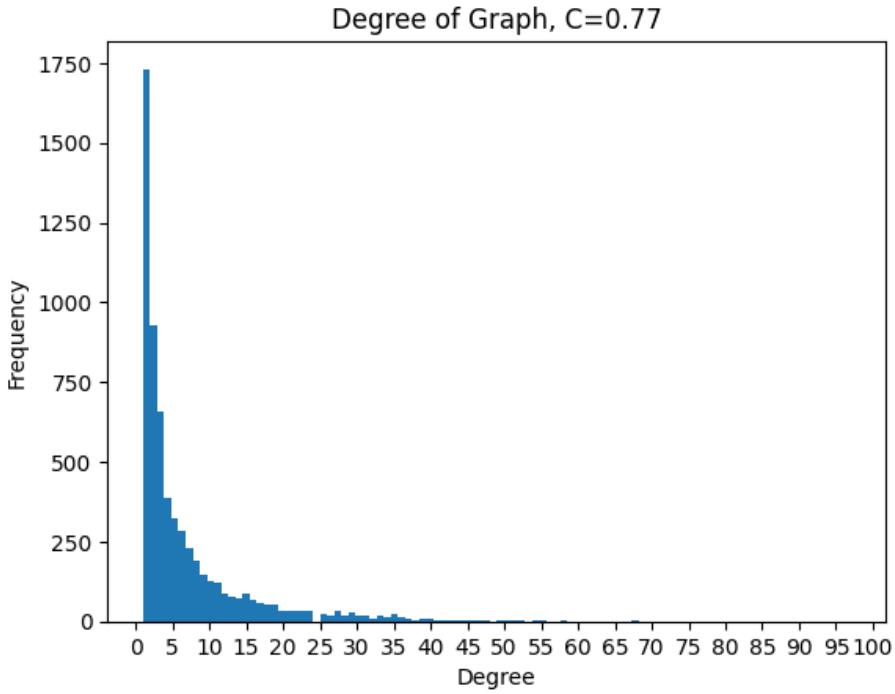

Figure 4: Histogram of Degrees

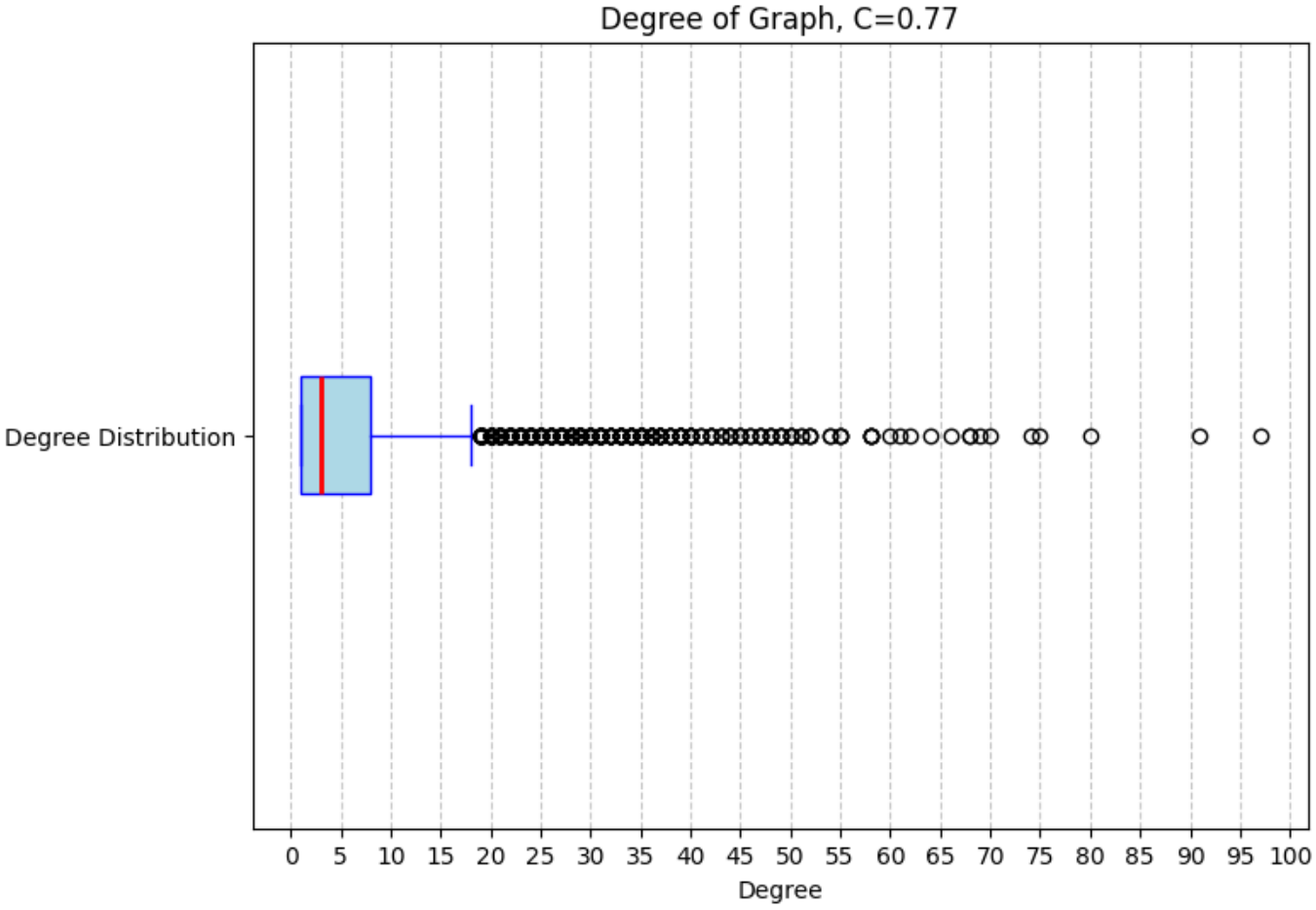

Figure 5: Boxplot of Degrees

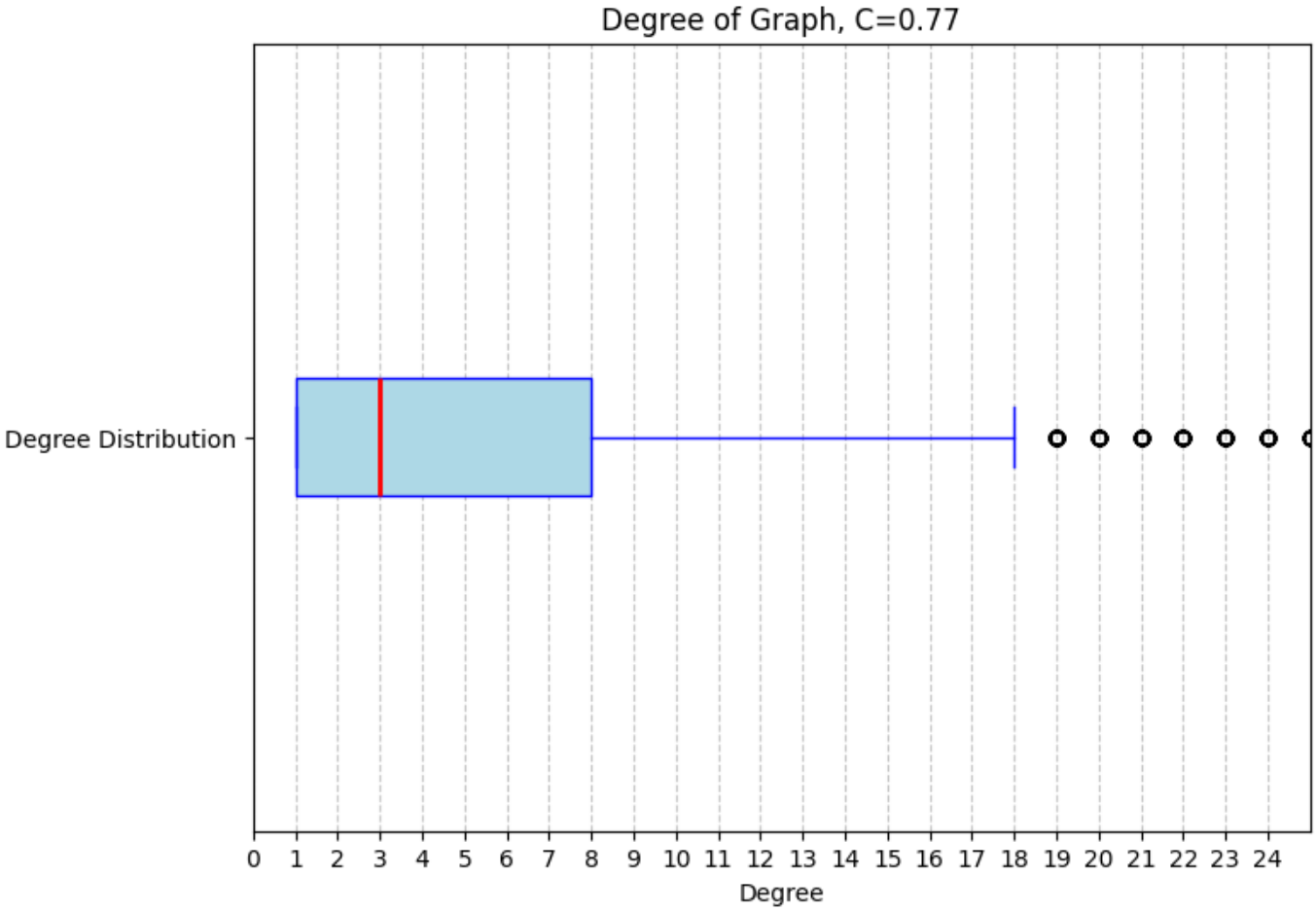

Figure 6: Boxplot of Degrees from range 0 - 25