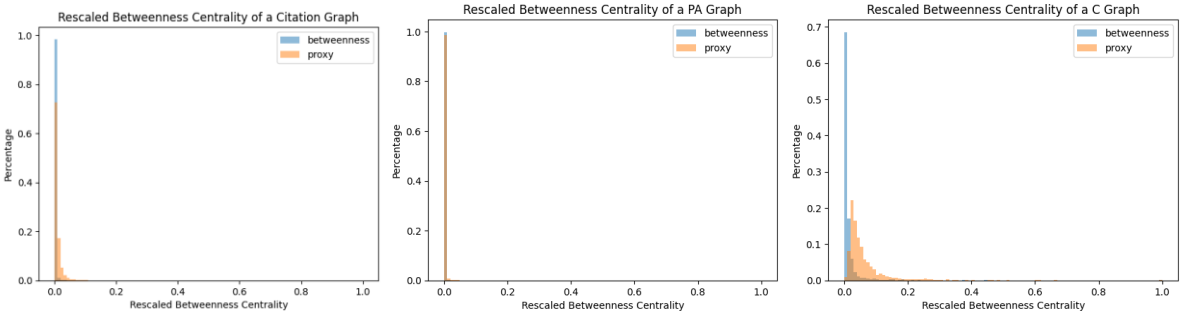
Question 2a

Betweenness centrality is a metric between 0 and 1. On the other hand, our proxy betweenness centrality is generally much bigger than 1 (some percent of the given node's degree, determined by its clustering coefficient). As such, we normalise our proxy betweenness centrality to between 0 and 1 to ensure that the distributions are comparable. This is a linear rescaling that will not affect the relative differences between values, so our analysis is still valid.

Betweenness centrality is a metric to highlight the most influential vertices in a graph based on the number of shortest paths between nodes that pass through it. Intuitively, our proxy is a good alternative to betweenness centrality since it still highlights the most influential vertices. An influential node is likely to have a high degree. An influential node is also likely to have a low clustering coefficient as this would require more shortest paths between nodes to pass through it (less alternatives). As such, multiplying the degree by 1 minus the clustering coefficient is likely to be a reasonable proxy for centrality.

Graph	RMSE	SI	Kendall Tau	P Value
Citation Graph	0.008	5.823	0.001	0.811
PA Graph	0.002	3.427	0.206	0.000
C Graph	0.047	3.360	0.398	0.000



As we can see from the graphs above, our proxy metric closely follows the true betweenness centrality distribution. However, our proxy metric may not accurately calculate values for specific nodes. If we look at the root mean squared error (RMSE) comparison of our proxy metrics, it obtains very low scores which would suggest a high accuracy. However, we observe the values that we obtain are always strongly weighted towards 0. Therefore, a small RMSE could actually mean quite a large error with respect to our values. As such, we can normalise the RMSE with respect to the mean to get the scatter index (SI) which is given as a percentage of error.

If we instead consider our metrics as ranking the vertices based on their importance, we can compare the rankings using Kendall rank correlation coefficient – where -1 represents strong negative correlation, 0 represents no correlation, and 1 represents strong positive correlation. As such, we see that our proxy doesn't accurately rank nodes for a citation graph. On the other hand, we see that our proxy more similarly ranks vertices based on their importance for the PA and C graphs. The only consideration is that the p value for the Kendall Tau value for the citation graph is very high so the confidence in the value obtained is very low.

Question 2b

Our proxy betweenness centrality may be preferred to the standard betweenness centrality since it is much quicker to compute. In fact, the standard metric is so slow to compute that for our analysis above we use a subset of 100 nodes. Specifically, when using the Brandes algorithm for betweenness centrality, the computation has a high complexity of O(v*e + v^2 * log(v)).