Lillyan Pan HW4

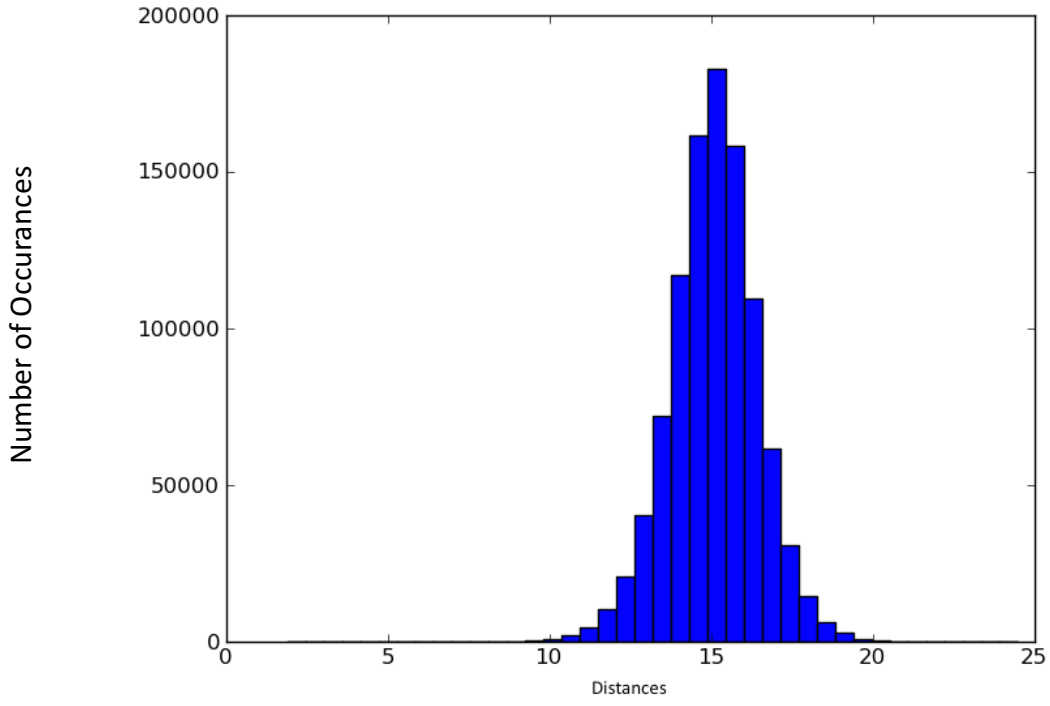
1a. The average s(i) values are as follows

| **k** | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| **s(i)** | -0.000814839 | -0.000676588875109 | -0.000392372690561 | -0.000526042596 09 | -0.000424267403506 | -0.000556933956077 | -0.000481157360302 |

1b. The k value that maximizes the s(i) score provides the best partitioning of the network as the average s(i) over all data of a cluster is a measure of the quality of the clustering. Hence, a maximal s(i) score is one that maximizes inner cluster cohesion and outer cluster separation. Thus, in this average, I found the best k values to be k = 5.

1c. Euclidean distance quickly tends towards a constant on data with lots of noise. (The number of dimensions continue to increase and become useless). In general, cosine similarity has been shows to be less affected from dimensionality. Note, Euclidean distance measures absolute differences between vectors. For some experiments, you may want relative distances instead to better see an overall trend in the data. Also, note clustering does not work very well for complicated data.

2a. **Histogram of all pairwise distances using 40 bins**




2b. Looking at the histogram (and knowing that there are 1000 nodes), I tried to pick a distance that would allow for about 5000 – 10000 edges (as the number of occurrences correlate to the

number of edges in the network.) Thus, I chose a threshold distance of 11.5, which produced the optimal amount of 5-10 edges when analyzing the network.

2c. Number of nodes: 756
Number of edges: 4079

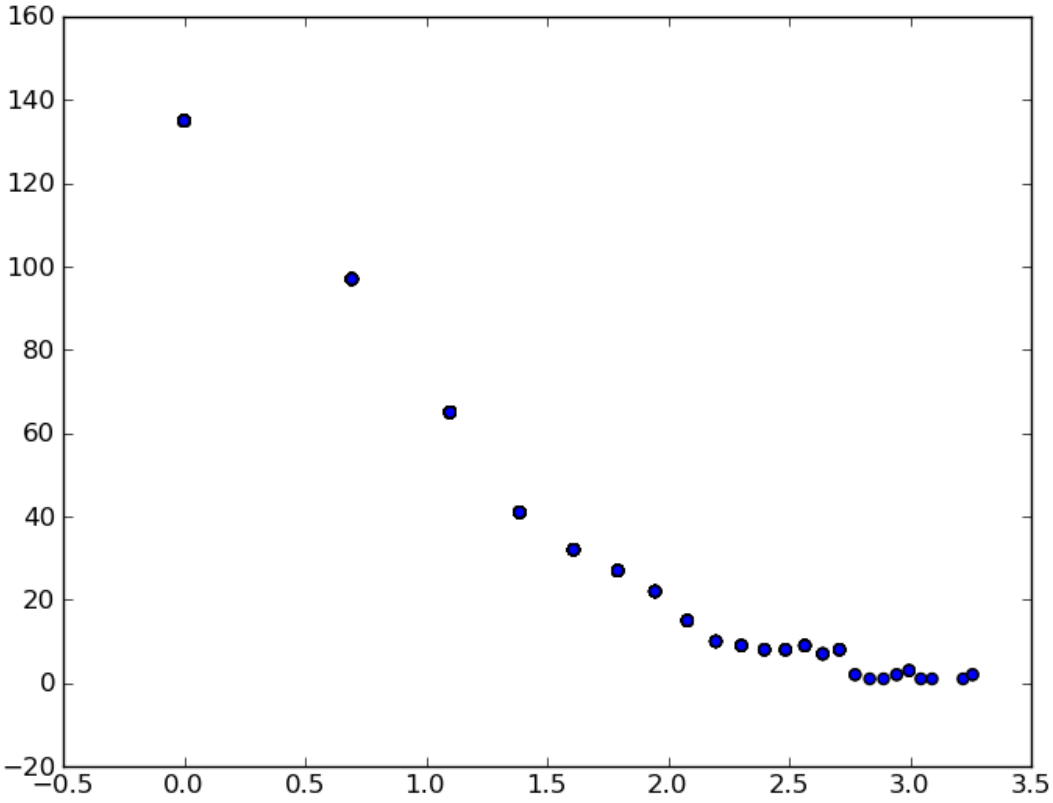
2d. Number of nodes: 443
Number of edges: 1089

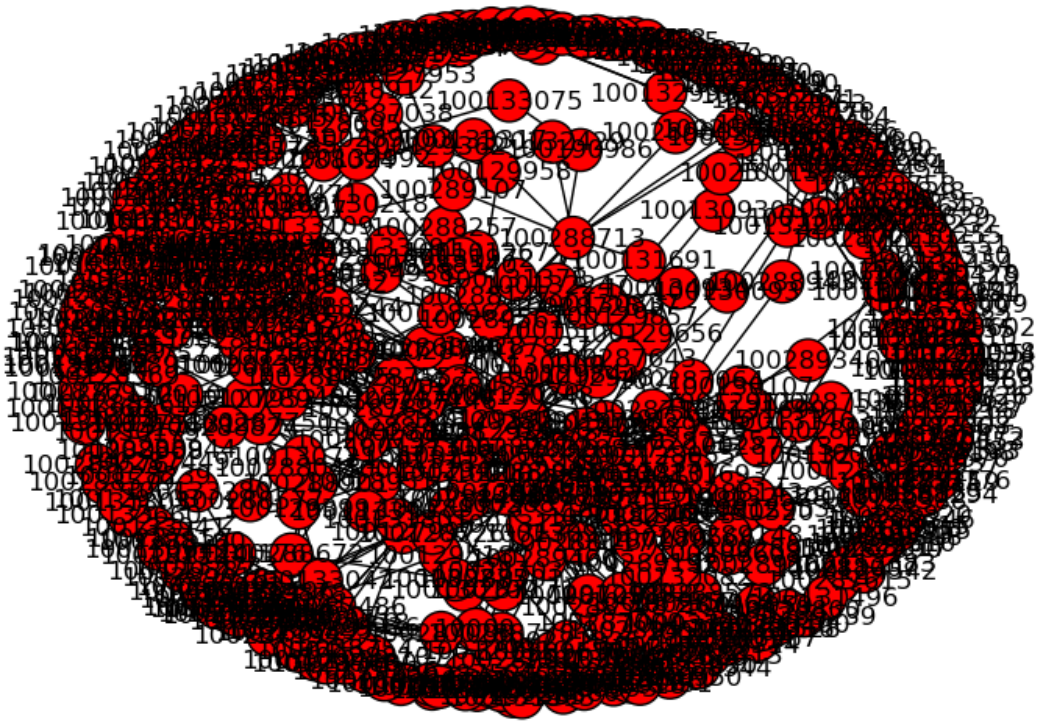
2e. Average clustering coefficient: 0.252422999567

3a. We see there are only a few nodes that have a high density of edges

Frequency

Degree


160
140
120
100
80
60
40
20
0
−20
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5



3b.The clustering was not too accurate. With complicated data, clustering is not the best statistical framework to use to analyze the data. Clustering needs lots of experiments and it is always possible to cluster even if there is no real correlation. It’s useful for learning about data, but doesn’t always give correct biological insight. To get better results, one could filter/adjust the data