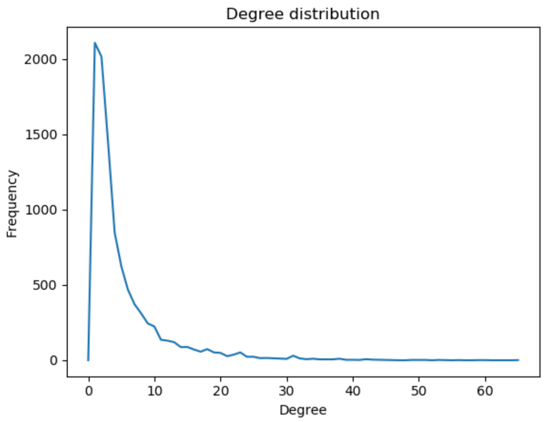
VU Thi Hai Yen
haiyen96.hp@gmail.com

1 Question 1

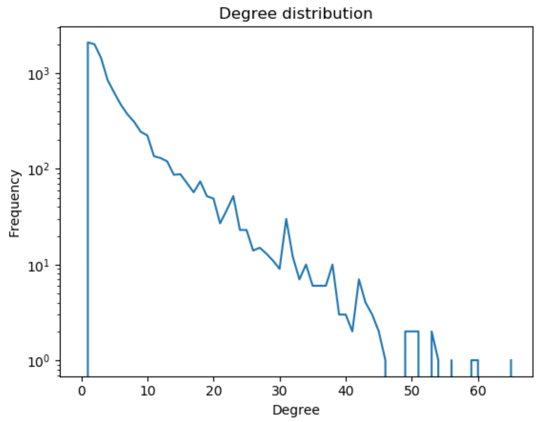
Consider an undirected graph G of n nodes without self-loops. Then for each node there are maximum $n-1$ edges passing through it. Thus the maximum number of edges in the graph is: $\frac{n(n-1)}{2}$ (as each edge is counted twice). The maximum number of triangles in the graph is equal to the number of 3-subset of n nodes, which is $\binom{n}{3} = \frac{n(n-1)(n-2)}{6}$. This happens when the graph is complete.

2 Question 2

We obtain the following graph:



We observe that the frequency of the nodes decreases as the degree increases. We can see that this follows an exponential distribution. To verify this, we plot the same graph with the y log-scale:



We see indeed that this graph is quasi linear.

3 Question 3

We have $\mathbf{L} = \mathbf{D} - \mathbf{A}$, thus for $\mathbf{x} = (x_1 \dots x_n)^\top$:

$$\mathbf{x}^\top \mathbf{L} \mathbf{x} = \sum_{i=1}^{n} D_{ii} x_i^2 - \sum_{i \neq j} A_{ij} x_i x_j = \frac{1}{2} \sum_{i \neq j} A_{ij} (x_i - x_j)^2.$$

The aim of spectral clustering is to map the n nodes to a vector $\mathbf{x} \in \mathbb{R}^n$ such that if node i and node j are connected, then their representations x_i and x_j should be close, which means that we want to minimize $(x_i - x_j)^2$ when $A_{ij} = 1$. This is equivalent to minimizing $\sum_{i \neq j} A_{ij}(x_i - x_j)^2$. However, we have found that $\sum_{i \neq j} A_{ij}(x_i - x_j)^2 = \mathbf{x}^\top \mathbf{L} \mathbf{x}$, thus the spectral clustering is aiming to find the vectors $\mathbf{x}$ that minimize $\mathbf{x}^\top \mathbf{L} \mathbf{x}$, which is indeed equivalent to finding the smallest eigenvalues of $\mathbf{L}$ as well as the corresponding eigenvectors.

4 Question 4

We have

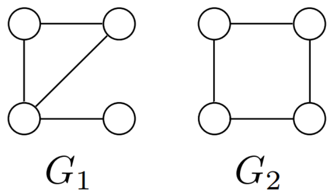
$$Q = \sum^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right].$$

In our case we have $m = 10$, we have 3 communities $(\{1, 2\}, \{3, 4, 5\}, \{6, 7, 8, 9\})$ with $l_1 = 1, l_2 = 3, l_3 = 5$ and $d_1 = 1 + 1 = 2, d_2 = 2 + 2 + 3 = 7, d_3 = 3 + 2 + 3 + 3 = 11$. Thus the modularity equals to

$$Q = \left[\frac{1}{10} - \left(\frac{2}{20} \right)^2 \right] + \left[\frac{3}{10} - \left(\frac{7}{20} \right)^2 \right] + \left[\frac{5}{10} - \left(\frac{11}{20} \right)^2 \right] = 0.465.$$

5 Question 5

We consider the two following graphs:



We have $\phi(G_1) = \phi(G_2) = [4, 2, 0, \dots, 0]$ but they are not isomorphic.

6 Question 6

We obtain the following results:

```
Accuracy of shorest path kernel:  0.95
Accuracy of graphlet kernel:  0.2
```

We see clearly that the shortest path kernel gives a very accurate result while the graphlet kernel gives a very bad result. The reason is that the graphlet kernel uses sampling method to count the number of occurrences of the graphlets, hence it doesn't count all the graphlets in graph, but just a random (small) portion of them.