Khaoula Belahsen
khaoula.belahsen@yahoo.fr

1 Question 1

- Let's suppose we have an undirected graph G = (V,E) with n nodes, and no self-loops. Each edge is defined by his 2 end points. Therefore, the maximum number of edges corresponds to the different choices of these 2 endpoints from the n nodes or $C_n^2 = \frac{n(n-1)}{2}$.
  The maximum number of edges is therefore n(n-1)/2.
- Let A be adjacency matrix representation of graph. If we calculate A3, then the number of triangle in Undirected Graph is equal to $trace(A^3)/6$.

2 Question 2

In the following figure, I plotted the frequency density histogram for the degrees of the nodes in the graph.

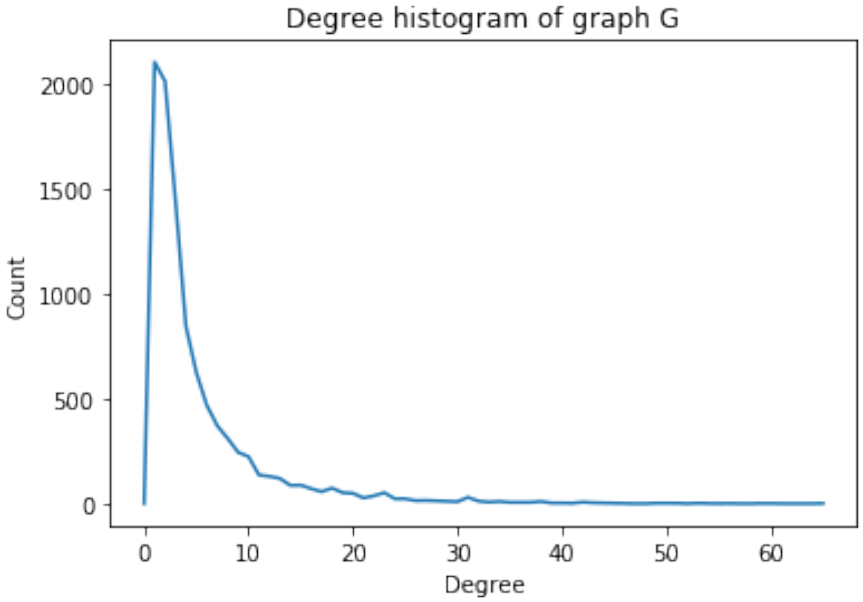

Figure 1:

We can see the highest count of degrees is for smaller degrees ¡ 10. The rest of degrees is poorly represented or not represented at all. The maximum degree is 65, with a mean of 5 and a median of 3.

We notice that the distribution of degrees follows a gamma law of probability.

3 Question 3

Why spectral clustering focuses on the smallest eigenvalues of the Laplacian matrix L?

With clustering, we want to separate points in different groups according to their similarities. For a similarity graph, we want to find a partition such that the edges between different groups have a very low weight and the edges within a group have high weight. [1]

Spectral clustering is an approximation to graph partitioning. If we have a similarity graph with its adjacency matrix, we construct a partition by solving the min cut problem.

To make sure that we have largely represented clusters, we use objective functions such as RatioCut and Ncut.[1].

$$RatioCut(A_1, \cdots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{|A_i|}$$

$$Ncut(A_1, \cdots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$$

.
Von Luxburg shows that minimizing the previous two objective functions is equivalent to the following problem

$$min_{A_1, \cdots, A_k} Tr(H'LH)$$

subject to

$$HH' = I$$

Where H is a matrix defined in the paper, and L the laplacian of the graph. Von Luxburg shows that by the Rayleigh-Ritz theorem, the solution is given by choosing H as the matrix of the k smallest eigenvectors of L as columns, for a spectral clustering of k clusters.

What is the problem that is optimized by the eigenvalue decomposition?

Using the eigenvalues decomposition allows to compute calculations for NxN data in a linear time of N.

4 Question 4

Let's compute the modularity of the graph in figure 1 of the lab sheet.

Modularity is defined as :

$$Q = \sum^{n_c} \frac{l_c}{m} - (\frac{d_c}{2m})^2$$

The graph of figure 1 was clustered into 3 clusters that I will index by their colors b (blue), g (green) and v (violet).

We have that :

- $m = |E| = 10$: the number of edges
- $n_c = 3$: the number of communities
- For the blue cluster : $l_b = 3$, $d_b = 7$
- For the green cluster : $l_g = 1$, $d_g = 2$
- For the violet cluster : $l_v = 5$, $d_b = 11$

The following step is the compute the quantity $\frac{l_c}{m} - (\frac{d_c}{2m})^2$ for the three clusters, we have that :

$$\frac{l_b}{m} - (\frac{d_b}{2m})^2 = \frac{3}{10} - (\frac{2}{20})^2 = \frac{9}{100}$$

$$\frac{l_g}{m} - (\frac{d_g}{2m})^2 = \frac{1}{10} - (\frac{7}{20})^2 = \frac{71}{400}$$

$$\frac{l_v}{m} - (\frac{d_v}{2m})^2 = \frac{5}{10} - (\frac{11}{20})^2 = \frac{79}{400}$$

And finally, we sum the quantities to obtain Q :

$$Q = \frac{9}{100} + \frac{71}{400} + \frac{79}{400} = 0.465$$

The modularity usually ranged between -1 and 1, so the value of modularity for the graph of figure 1 shows a good community structure.

5 Question 5

Figure 2: Example of 2 non-isomorphic graphs with the same shortest path representation. sp = [0,3,2,0,..,0]

6 Question 6

In this part, we computed the embeddings using 2 different kernels : the shortest path kernel and the graphlet kernel.

When computing the accuracies for the same SVM model with the different embeddings, we notice that the shortest path kernels achieves a high accuracy of 0.95, while the graphlet kernel achieve a lower accuracy of 0.45.

We observe that graphlet kernel doesn't achieve a high accuracy. This may be due to the fact that we used 3-nodes graphlet kernels.

References

[1] Ulrike von Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007.