Adriano Del Gallo
adriano.delgallo@telecom-paristech.fr





# 1 Question 1

The maximum number of edges in an undirected graph with $n$ edges is $\binom{n}{2}$ since we only count edge once. The number of triangles is an undirected graph is at most $\binom{n}{3}$ since there are $n(n-1)(n-2)$ ways to chose the vertices and 6 ways to create the same undirected triangle from a directed one.

# 2 Question 2

We see that nodes with lower degree are far more probable than ones with a high degree. By drawing the log plot, we obtain a straight line. Which suggests that we have an exponential distribution for the frequency of the degree. Further more, we observe that the degree zero is very uncommon so our exponential law is truncated at 1.

# 3 Question 3

The decomposition focuses on small values because subsets of connected components can be identified by the eigenspaces of the laplacian associated to the eigenvalue zero. More precisely, a fully connected component will have eigenvalue 0 for the indicator vector of this subset. Thus our goal is the find the eigenvectors associated to value zero or values close to zero, which will give us the projectors onto the connected components.
The eigenvalue problem is an approximation of the minimum ratio cut problem which is why we use k-means to project the eigenvalue solutions onto the admissible solution space. The formulation of the problem is that we look for partitions which minimise the weights going from one set to another, i.e. clusters.

# 4 Question 4

In this graph $m = 10$ and $n_c = 3$
For the green component : $l_c = 1$ and $d_c = 2$ so we get

$$\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2 = \frac{1}{10} - \frac{1}{100} = 0.090 \tag{1}$$

For the blue component : $l_c = 3$ and $d_c = 7$ so we get

$$\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2 = \frac{3}{10} - \frac{49}{400} = 0.177 \tag{2}$$

For the grey component : $l_c = 5$ and $d_c = 11$ so we get

$$\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2 = \frac{5}{10} - \frac{121}{400} = 0.197 \tag{3}$$

By summing over the three communities we get $Q = 0.465$

# 5 Question 5

The graphs in Figure 1 are non-isomorphic however their embedding from the shortest path kernel is the same $f_{G_1} = f_{G_2} = [4, 2, 0, 0, -]$

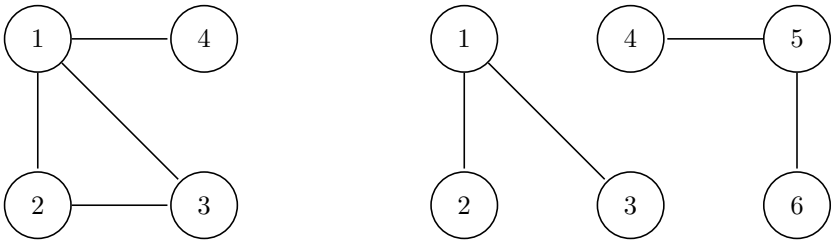


Figure 1: Two non-isomorphic graphs with same shortest path kernel embedding

# 6 Question 6

In our tests the accuracy for shortest path kernel is 1.0 and the accuracy for graphlet kernel is 0.45.
The graphlet kernel performed less well than a random classifier which means its performances are really bad. Because of the way we generated our graphs (pathgraphs and cycles) it is very likely that when we extract three random nodes, the sub-graph will be isomorphic to graphlet $G_4$. So except in the rare case where we get $G_2, G_3$ or $G_1$, the graphlet kernel will always identify $G_4$. So the graphs in the dataset will have very close embeddings in the kernel space, which makes it hard to perform classification. We also notice that each element of this dataset gets a unique embedding in the shortest path kernel space and that the embedding vectors have a specific structure for each class. This will produce high quality embeddings which accounts for the perfect separability observed.