Altegrad Lab 4
MVA

Matthieu Boyer

Introduction

In this report, I have added the answers to the theoretical questions and the figures and numerical results from certain tasks. The tasks are completed in the different files of the `/code` folder of the archive, based on the given template.

Task 1

The graph has 9 877 nodes and 25 998 edges.

Task 2

The graph has 429 connected components.

The largest has 8 638 nodes and 24 827 edges, representing 87.46% of total nodes and 95.50% of total edges.

Question 1

G can be seen as the graph coproduct of the graphs K_{20} and $K_{10,10}$. The complement of $K_{10,10}$ is (considering only its vertices in G) the coproduct of K_{10} and K_{10} (as all vertices in each component become connected and we remove all edges between components). The complement of K_{20} is the graph $\bar{K}_{20}$ with 20 vertices and no edges. There are four possible cases when considering three vertices in the complement of G

1. At least two vertices are in $\bar{K}_{20}$, and there are no edges and thus no triangles.
2. There is at least one vertex in each copy of K_{10}, there are no edges between the two and there are no triangles.
3. The three vertices are in the same copy of K_{10}: there are $2 \times \binom{3}{10}$ such possibilities which all yield to a different triangle.
4. Two vertices are in the same copy of K_{10} and the last in in K_{20}: there are $2 \times \binom{2}{10} \times 20 = 1800$ such possibilities.

In the end, there are $\boxed{2040}$ triangles in the complement of G.

Question 2

Taking the gradient with respect to x the expression defining the Rayleigh quotient of G with matrix A, we get:

$$\nabla_x R(A, x) = \frac{\|x\|^2 \nabla \langle x, Ax \rangle - \langle x, Ax \rangle \nabla \|x\|^2}{\|x\|^4} = \frac{2 \|x\|^2 Ax - 2(\langle x, Ax \rangle)x}{\|x\|^4}$$

since A is symmetric as G is undirected. The gradient cancels if and only if $\langle x, x \rangle Ax = \langle x, Ax \rangle x$. This means that the gradient vanishes precisely when x is an eigenvector of A and thus, a non-zero vector x is a stationary point of $R(A, \cdot)$ if and only if x is an eigenvector of A.

Task 4

If we try for 50 clusters in the giant connected component, 8 638 out of 8 638 nodes are assigned, with a minimum cluster size of 5, a maximum cluster size of 7 619 and an average cluster size of 172.76.

Question 3

Let's start by computing modularities for each cluster arrangement. The graph has $m = 13$ edges.

1. For the graph on the left, the blue cluster has 5 nodes with $l_c = 7$ inner edges and $d_c = 15$ for its degree sum. The orange cluster has 4 nodes with $l_c = 5$ inner edges and $d_c = 11$ for its degree sum. As such,

$$Q_1 = \left[\frac{7}{13} - \left(\frac{15}{26}\right)^2\right] + \left[\frac{5}{13} - \left(\frac{11}{26}\right)^2\right] = 0.41$$

2. For the graph on the right, the blue cluster has 3 nodes with $l_c = 2$ inner edges and $d_c = 8$ for its degree sum. The orange cluster has 6 nodes with $l_c = 7$ and 18 for its degree sum. As such,

$$Q_2 = \left[\frac{2}{13} - \left(\frac{8}{26}\right)^2\right] + \left[\frac{7}{13} - \left(\frac{18}{26}\right)^2\right] = 0.12$$

As the modularity is much larger in the first clustering (with $Q_2 \simeq \frac{Q_1}{4}$), it would seem that the first clustering has a better clustering structure.

Task 6

We ran the experiment 50 times for random clustering, the value below is the mean of the modularities. The modularity obtained by spectral clustering with $k = 50$ clusters is 0.2094. The modularity obtained by random clustering with $k = 50$ clusters is -0.0004. Spectral clustering thus produces significantly better community structure than random assignment.

Question 4

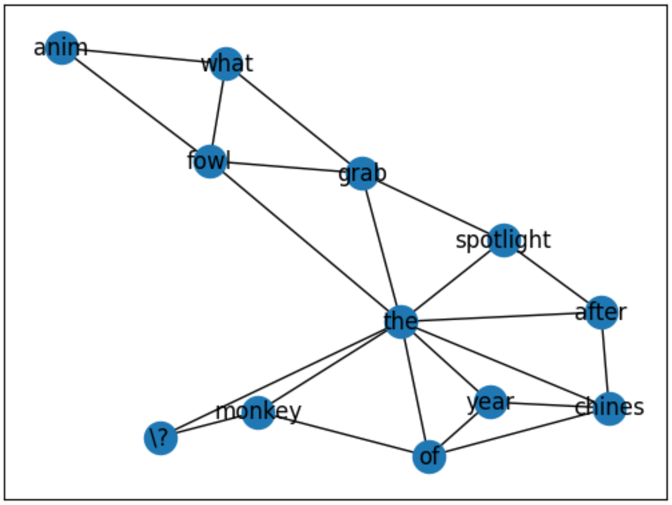
Let's consider the Shrikandhe graph and the lattice graph $L_2(4)$ (or 4×4 rook graph) (see . Both graphs have the same distance maps (from their strong regularity with parameters $(16, 6, 2, 2)$) and thus the same shortest-paths kernel graphs. However, it can be shown that the two graphs are non-isomorphic. Indeed, $L_2(4)$ has 36 4-cycles whereas Shrikhande has 120, and thus cannot be isomorphic. See `https://www.sciencedirect.com/science/article/pii/S0022000020300386` for another proof (and reference to the idea).

Task 10

For the shortest path kernel we find an accuracy of 0.8947, whereas we get an accuracy of 0.6842 for the graphlet kernel. The difference in accuracy is 0.2105 or about 25% in relative error.

Task 11

Here is an example of a graph generated by the `create_graphs_of_words` function:



Task 13

The SVM classifier we trained gives an accuracy of 0.938.

Question 5

Initial Setup

Graph G: Nodes with labels $\{2, 1, 5, 3, 1, 4\}$
Graph G': Nodes with labels $\{2, 1, 5, 4, 3, 4\}$

Step 1: Collect Initial Labels

For each node, we collect its initial label and the multiset of its neighbors' labels.
Graph G:

- Node with label 2: neighbors $= \{1, 5\} \rightarrow$ multiset $\{1, 5\}$
- Node with label 1 (left): neighbors $= \{2, 3\} \rightarrow$ multiset $\{2, 3\}$
- Node with label 5: neighbors $= \{2, 3\} \rightarrow$ multiset $\{2, 3\}$
- Node with label 3: neighbors $= \{1, 5, 1, 4\} \rightarrow$ multiset $\{1, 1, 4, 5\}$
- Node with label 1 (right): neighbors $= \{3\} \rightarrow$ multiset $\{3\}$

- Node with label 4: neighbors $= \{3\} \to$ multiset $\{3\}$

Graph G':

- Node with label 2: neighbors $= \{1, 5\} \to$ multiset $\{1, 5\}$
- Node with label 1: neighbors $= \{2, 4\} \to$ multiset $\{2, 4\}$
- Node with label 5: neighbors $= \{2, 3\} \to$ multiset $\{2, 3\}$
- Node with label 4 (left): neighbors $= \{1, 3, 4\} \to$ multiset $\{1, 3, 4\}$
- Node with label 3: neighbors $= \{5, 4\} \to$ multiset $\{4, 5\}$
- Node with label 4 (right): neighbors $= \{4\} \to$ multiset $\{4\}$

Step 2: Create New Labels

For each node, we concatenate its current label with the sorted multiset of neighbors' labels.

Graph G:

- 2: $(2, [1, 5]) \to$ new label "2-1-5"
- 1 (left): $(1, [2, 3]) \to$ new label "1-2-3"
- 5: $(5, [2, 3]) \to$ new label "5-2-3"
- 3: $(3, [1, 1, 4, 5]) \to$ new label "3-1-1-4-5"
- 1 (right): $(1, [3]) \to$ new label "1-3"
- 4: $(4, [3]) \to$ new label "4-3"

Graph G':

- 2: $(2, [1, 5]) \to$ new label "2-1-5"
- 1: $(1, [2, 4]) \to$ new label "1-2-4"
- 5: $(5, [2, 3]) \to$ new label "5-2-3"
- 4 (left): $(4, [1, 3, 4]) \to$ new label "4-1-3-4"
- 3: $(3, [4, 5]) \to$ new label "3-4-5"
- 4 (right): $(4, [4]) \to$ new label "4-4"

Step 3: Count Label Occurrences

Graph G label histogram:

- "2-1-5": 1
- "1-2-3": 1
- "5-2-3": 1
- "3-1-1-4-5": 1

- “1-3”: 1
- “4-3”: 1

Graph G’ label histogram:

- “2-1-5”: 1
- “1-2-4”: 1
- “5-2-3”: 1
- “4-1-3-4”: 1
- “3-4-5”: 1
- “4-4”: 1

Step 4: Compute Kernel Value

The WL kernel computes the inner product of the label histograms.

Common labels after 1 iteration:

- “2-1-5”: $1 \times 1 = 1$
- “5-2-3”: $1 \times 1 = 1$

Kernel value: $k(G, G') = 2$

The kernel value of 2 indicates **limited structural similarity** between graphs G and G' after one WL iteration. Out of 6 refined labels in each graph, only 2 labels match, meaning that most nodes have different local neighborhood structures. The matching labels (“2-1-5” and “5-2-3”) indicate that both graphs share some common structural patterns in terms of how certain labeled nodes connect to their neighbors, but the majority of the local structures differ significantly. A higher kernel value would indicate greater structural similarity.

Task 14

I could not run the code due to time constraints on my computer.