Alexandre Zajac
zajac.alexandre@gmail.com

Lab session # 4
ALTEGRAD 2020

12/11/20

1 Question 1

Here we consider a graph G with 2 connected components, and we want to know the number of edges and triangles:

- The first component is a complete graph of $n = 100$ vertices. The number of edges is thus:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n \cdot (n-1)}{2} = 4\ 950 \tag{1}$$

  since we can reach (n-1) edges from every of the n edges, without repetition. Since every triangle is closed in a complete graph, the number of triangles is:

$$\frac{1}{3} * \binom{n}{3} = \frac{161\ 700}{3} = 53\ 900 \tag{2}$$

- The second component is a bipartite graph of $n = 50$ vertices. The number of edges simplifies to $n * n = 2500$. Since it's a bipartite graph, there are no closed triangles, so the count of triangles here is 0.

To sum up the number of edges is $\mathbf{4\ 950 + 2\ 500 = 7\ 450}$ and the number of triangles is **53 900**.

2 Question 2

The global clustering coefficient is:

$$C = \frac{\text{number of closed triplets}}{\text{number of total triplets (open and closed)}} \tag{3}$$

So we can see that the maximum value for C is **1**, and this happens only when the number of open triplet is 0: this is the case in a **complete graph**.

3 Question 3

Here we consider a connected graph with a single connected component. Since the matrix $L = D - W$ is symmetric, positive and semi-definite, its smallest trivial eigenvalue is **0 associated to the unit eigenvector 1**.

With the **Rayleigh-Ritz Theorem**, we know that the eigenvector corresponding to smallest eigenvalue (0) offers no useful information when solving the Two-Way Cut from the Laplacian, and so removing it before applying spectral clustering doesn't affect the results.

4 Question 4

Let's see the algorithm for spectral clustering:
Here we can see clearly that step 1, 2 and 3 are purely deterministic since they are mathematical computation with no approximation or random process involved. However, when applying the K-means algorithm, we are doing a stochastic step. Even if the results of the K-means step and thus clustering algorithm should attain a local optimum every time we run it, running k-means with different random seeds could give different solutions to the minimization problem. At that level, the Spectral clustering algorithm's output is **stochastic**.

5 Question 5

Here is the formula for calculating the modularity of a graph clustering:

$$Q = \sum^{n_c} \left[\frac{l_c}{m} - (\frac{d_c}{2m})^2 \right] \tag{4}$$

Now let's compute the modularities for each of the 2 scenarios.

Algorithm 1 Spectral Clustering

Input: Graph $G = (V, E)$ and parameter k

Output: Clusters $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$ (i.e., cluster assignments of each node of the graph)

1: Let $\mathbf{A}$ be the adjacency matrix of the graph

2: Compute the Laplacian matrix $\mathbf{L_{rw}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$. Matrix $\mathbf{D}$ corresponds to the diagonal degree matrix of graph G (i.e., degree of each node v (= number of neighbors) in the main diagonal)

3: Apply eigenvalue decomposition to the Laplacian matrix $\mathbf{L_{rw}}$ and compute the eigenvectors that correspond to d smallest eigenvalues. Let $\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\ldots|\mathbf{u}_d] \in \mathbb{R}^{m \times d}$ be the matrix containing these eigenvectors as columns

4: For $i = 1, \ldots, m$, let $y_i \in \mathbb{R}^d$ be the vector corresponding to the i-th row of $\mathbf{U}$. Apply k-means to the points $(y_i)_{i=1,\ldots,m}$ (i.e., the rows of $\mathbf{U}$) and find clusters $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$

Figure 1: Algorithm for Spectral Clustering

- Scenario a): For both green and blue community, we have $l_c = 6$ and $d_c = 13$ and $m = 13$, so:

$$Q = 2 * \left(\frac{6}{13} - \left(\frac{13}{26} \right)^2 \right) \approx 0.423 \tag{5}$$

- Scenario b): For the green community we have $l_c = 2$, $d_c = 11$ and $m = 13$. For the blue community, we have $l_c = 4$ and $d_c = 15$, so we get:

$$Q = \left(\frac{2}{13} - \left(\frac{11}{26} \right)^2 \right) + \left(\frac{4}{13} - \left(\frac{15}{26} \right)^2 \right) \approx -0.050 \tag{6}$$

6 Question 6

Let's recall the formula for calculating the shortest path kernel between two Floyd-transformed graphs G1 and G2:

$$k(G_1, G_2) = \sum_{e_1 \in E_1} \sum_{e_2 \in E_2} k_{edge}(e_1, e_2) \tag{7}$$

So let's see the two graphs we have, P_n which is a path of n vertices and C_n, a cycle of n vertices (here $n = 4$):

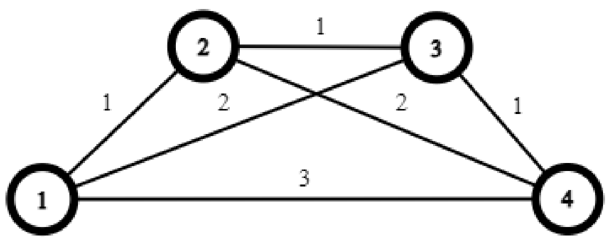

Figure 2: A P_n graph representation

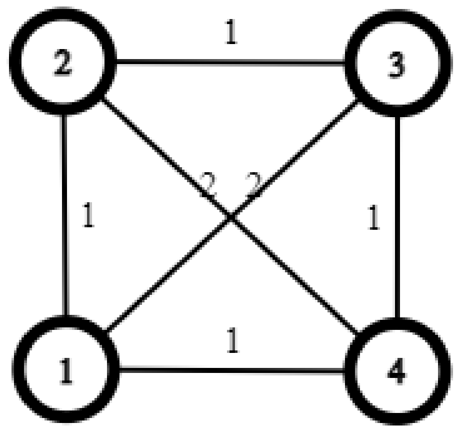

Figure 3: A C_n graph representation

We can see that the row vector associated with P_n is $p = (3, 2, 1)$, and the one associated with C_n is $c = (4, 2, 0)$. If we denote Sk_1 the shortest path kernel for (C_4, C_4), Sk_2 the shortest path kernel for (C_4, P_4) and Sk_3 the shortest path kernel for (P_4, P_4), we finally get:

$$Sk_1 = c \cdot c = 20 \tag{8}$$

$$Sk_2 = c \cdot p = 16 \tag{9}$$

$$Sk_3 = p \cdot p = 14 \tag{10}$$