

Mid-term Exam (Graph Mining – Spring 2024)

Full Name:

Student ID:

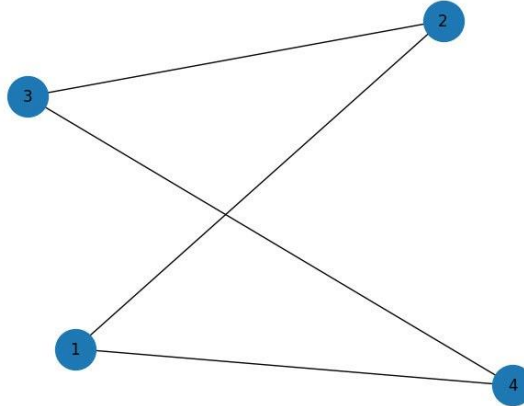
- The formula and solution process should be presented with the answer.
- All the codes must include detail comments in English.

1. Consider an undirected graph G of four nodes given in the following figure, calculate betweenness and closeness centrality of node 1 (5pt)

Equation betweenness centrality: $B(v_i) = \sum_{s,t \in V} \frac{\sigma(s,t|v_i)}{\sigma(s,t)}$, where $\sigma(s,t)$ is the number of shortest paths from node s to node t, $\sigma(s,t|v_i)$ is the number of shortest paths from node s to node t that passing through node v_i .

Normalized betweenness centrality: $\bar{B}(v_i) = \frac{B(v_i)}{(n-1)(n-2)/2}$ where n is number of nodes.

Equation closeness centrality: $C(v_i) = \frac{N-1}{\sum_{j=1}^{N-1} d(v_j, v_i)}$, where $d(v_j, v_i)$ is number of nodes in the shortest path between node v_j and node v_i , and N-1 is the number of nodes reachable from v_i .



Ans:

Betweenness centrality of node 1

	$\sigma(s,t)$	$\sigma(s,t i)$	$\sigma(s,t i)/\sigma(s,t)$
1,2	1	0	0
1,3	2	0	0
1,4	1	0	0
2,3	1	0	0
2,4	2	1	1/2
3,4	1	0	0

$$\bar{B}(v_i) = \frac{B(v_i)}{(n-1)(n-2)/2} = \frac{1/2}{(4-1)(4-2)/2} = \frac{1/2}{3 \cdot 2/2} = \frac{1}{6}$$

Closeness centrality of node 1

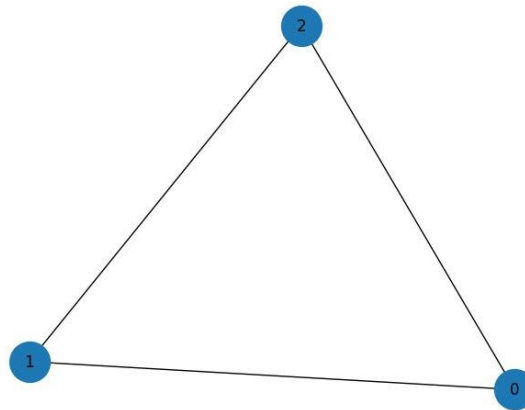
$$C(v_i) = \frac{N - 1}{\sum_{j=1}^{N-1} d(v_j v_i)} = \frac{4 - 1}{1 + 2 + 1} = \frac{3}{4}$$

2. Calculate Eigenvector, Katz and PageRank centrality(10pt)

- a. Consider an undirected graph G of three nodes given in the following figure, calculate Eigenvector, Katz centrality of node 2 with $\alpha = 0.1$, $\beta = 1$, $t = 1$ (6pt)

Equation Eigenvector: $x_i(t) = \sum_{v_j \in N(v_i)} A_{ij} x_j(t-1)$, where A is adjacency matrix, t is time, with the centrality at time t = 0 being $x_j(0) = 1 \forall j$

Equation Katz: $Katz(G) = \beta(I - \alpha A^T)^{-1} \cdot \mathbf{1}$, where α is damping factor, β is bias constant, I refers to the identity matrix, and $\mathbf{1}$ is a column vectors of ones. From Katz(G) results, write down the Katz centrality of node 2.



Ans:

Eigenvector: $x_i(t) = \sum_{v_j \in N(v_i)} A_{ij} x_j(t-1)$

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$x_j(0) = 1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$x_2(1) = \sum_{v_j} A_{2j} x_j(0) = (1 * 1) + (0 * 1) + (1 * 1) = 2$$

Explanation:

The first element $1*1$ represents the presence of a connection from node 2 to node 0

The second element $1*0$ represents the presence of a connection from node 2 to node 1

The third element $1*1$ represents the presence of a connection from node 2 to node 2

$$Katz(G) = \beta(I - \alpha A^T)^{-1} \cdot \mathbf{1}$$

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

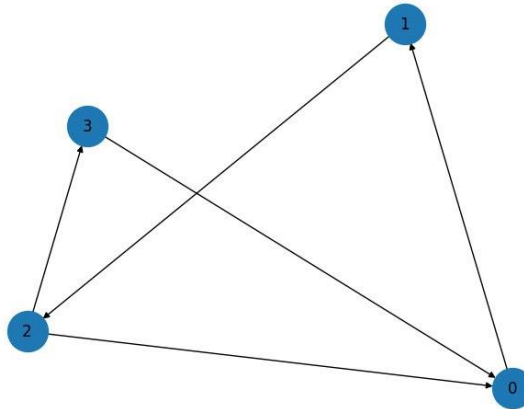
$$A^T = A$$

$$\begin{aligned} Katz(G) &= \mathbf{1} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \right)^{-1} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -0.1 & -0.1 \\ -0.1 & 1 & -0.1 \\ -0.1 & -0.1 & 1 \end{bmatrix}^{-1} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{\det} \begin{bmatrix} 1 & -0.1 & -0.1 \\ -0.1 & 1 & -0.1 \\ -0.1 & -0.1 & 1 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{1 + (-0.1)^3 + (-0.1)^3 - (-0.1)^2 - (-0.1)^2 - (-0.1)^2} \begin{bmatrix} 0.99 & 0.11 & 0.11 \\ 0.11 & 0.99 & 0.11 \\ 0.11 & 0.11 & 0.99 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{1 - 0.001 - 0.001 - 0.01 - 0.01 - 0.01} \begin{bmatrix} 0.99 & 0.11 & 0.11 \\ 0.11 & 0.99 & 0.11 \\ 0.11 & 0.11 & 0.99 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{0.968} \begin{bmatrix} 0.99 & 0.11 & 0.11 \\ 0.11 & 0.99 & 0.11 \\ 0.11 & 0.11 & 0.99 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{0.99}{0.968} & \frac{0.11}{0.968} & \frac{0.11}{0.968} \\ \frac{0.11}{0.968} & \frac{0.99}{0.968} & \frac{0.11}{0.968} \\ \frac{0.11}{0.968} & \frac{0.11}{0.968} & \frac{0.99}{0.968} \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{aligned}$$

$$Katz(\text{node } 2) = \frac{0.11}{0.968} + \frac{0.99}{0.968} + \frac{0.11}{0.968} = \frac{1.21}{0.968} = 1.25$$

- b. Consider a directed graph G of four nodes given in the following figure, calculate PageRank centrality of all nodes, with $\beta = 0.85$ (4pt)

Equation PageRank centrality of node i: $x_i = \sum_{(j,i) \in E} x_j + \beta$, where x_j is PageRank score of all pages j that point to page i



Ans:

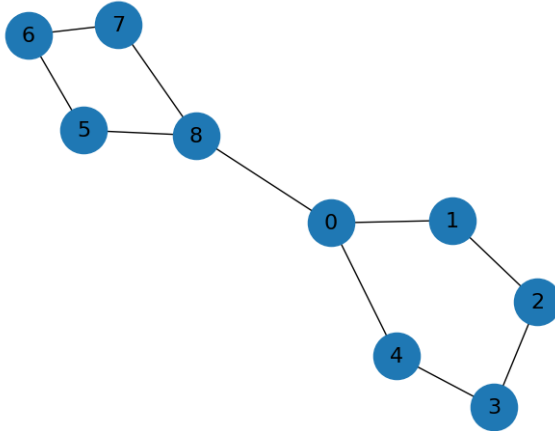
PageRank score equally 4 pages

$$x_0 = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$$

$$E = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$x_i = \sum_{(j,i) \in E} x_j + \beta = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} + 0.85 = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} + 0.85 = \begin{pmatrix} 1.35 \\ 1.1 \\ 1.1 \\ 1.1 \end{pmatrix}$$

3. Consider an undirected graph G of nine nodes given in the following figure. There are two communities in the graph: A = {0, 1, 2, 3, 4} and B = {5, 6, 7, 8}. (10pt)



- a. Calculate Min-cut and Normalized cut measurements of A and B.

Ans:

$$\text{Min_cut}(A, B) = 1$$

$$\text{N_cut}(A, B) = \frac{1}{1+5} + \frac{1}{1+4} = \frac{11}{30}$$

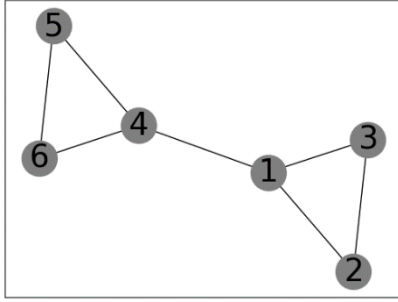
- b. Calculate conductance of A and B using the equation (1).

$$\text{conductance}(A, B) = \frac{\text{cut}(A, B)}{\min(\text{assoc}(A, V), \text{assoc}(B, V))} \quad (1)$$

where $\text{assoc}(A, V)$ and $\text{assoc}(B, V)$ is the total connection from nodes in A and B to all nodes in the graph, respectively. $\text{cut}(A, B)$ is the number of cut between 2 communities A and B.

Ans: $\text{Conductance}(A, B) = \frac{1}{\min(6, 5)} = \frac{1}{5} = 0.2$

4. Consider an undirected graph G of six nodes given in the following figure. Apply the Equation (1) to calculate the clustering C_i of each node i and Equation (2) to calculate the average clustering in the graph G. (10pt)



$$\text{Equation (1): } C_i = \frac{2L_i}{d_i(d_i-1)} \quad (1)$$

where d_i is the degree of node i and L_i is number of edges between neighbors of node i .

$$\text{Equation (2): } \langle C \rangle = \frac{1}{N} \sum_{i=0}^N C_i \quad (2)$$

Ans:

$$\text{Clustering node 1: } C_1 = \frac{2L_1}{d_1(d_1-1)} = \frac{2.1}{3.(3-1)} = 0.333$$

$$\text{Clustering node 2: } C_2 = \frac{2L_2}{d_2(d_2-1)} = \frac{2.1}{2.(2-1)} = 1$$

$$\text{Clustering node 3: } C_3 = \frac{2L_3}{d_3(d_3-1)} = \frac{2.1}{2.(2-1)} = 1$$

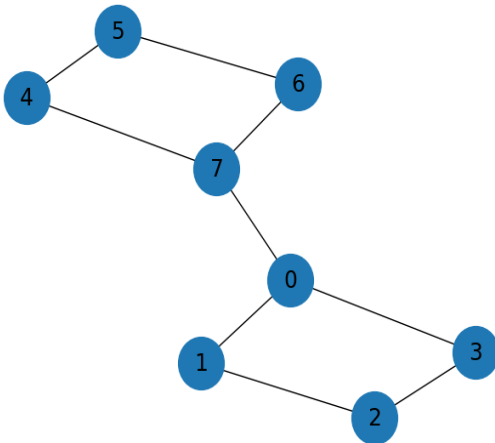
$$\text{Clustering node 4: } C_4 = \frac{2L_4}{d_4(d_4-1)} = \frac{2.1}{3.(3-1)} = 0.33$$

$$\text{Clustering node 5: } C_5 = \frac{2L_5}{d_5(d_5-1)} = \frac{2.1}{2.(2-1)} = 1$$

$$\text{Clustering node 6: } C_6 = \frac{2L_6}{d_6(d_6-1)} = \frac{2.1}{2.(2-1)} = 1$$

Average Clustering $\langle C \rangle = 0.7778$.

5. Consider an undirected graph G of eight nodes given in the following figure with two communities: $A = \{0, 1, 2, 3\}$ and $B = \{4, 5, 6, 7\}$. Apply the Equation (1) to calculate the modularity Q of the two communities. (10pt)



$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \cdot \delta(v_i, v_j) \quad (1)$$

$$\delta(v_i, v_j) = \begin{cases} 1 & \text{if } v_i \text{ and } v_j \text{ are in the same community.} \\ 0 & \text{otherwise.} \end{cases}$$

where m is the number of edges, A is the adjacency matrix of G , d_i is the degree of node v_i

Ans:

$$Q = \frac{1}{2 \times m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \cdot \delta(v_i, v_j)$$

$$Q = \frac{1}{2 \times 9} \left[\left(1 - \frac{d_0 d_1}{18}\right) + \left(1 - \frac{d_1 d_2}{18}\right) + \left(1 - \frac{d_2 d_3}{18}\right) + \left(1 - \frac{d_3 d_0}{18}\right) + \left(1 - \frac{d_7 d_4}{18}\right) + \left(1 - \frac{d_4 d_5}{18}\right) + \left(1 - \frac{d_5 d_6}{18}\right) + \left(1 - \frac{d_6 d_7}{18}\right) \right]$$

$$Q = \frac{1}{2 \times 9} \left[\left(1 - \frac{3 \times 2}{18}\right) + \left(1 - \frac{2 \times 2}{18}\right) + \left(1 - \frac{2 \times 2}{18}\right) + \left(1 - \frac{2 \times 3}{18}\right) + \left(1 - \frac{3 \times 2}{18}\right) + \left(1 - \frac{2 \times 2}{18}\right) + \left(1 - \frac{2 \times 2}{18}\right) + \left(1 - \frac{2 \times 3}{18}\right) \right]$$

$$Q = \frac{1}{2 \times 9} \left[4 \left(1 - \frac{3 \times 2}{18}\right) + 4 \left(1 - \frac{2 \times 2}{18}\right) \right] = 0.3209$$

6. Consider an undirected graph G of eight nodes given in the following figure, calculate Jaccard's coefficient (JC), Adamic-Adar (AA) index of node 2 and node 6 (10pt)

Equation JC: score (x, y) = $\frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$, where N(x), N(y) are neighbor nodes of node x, y respectively

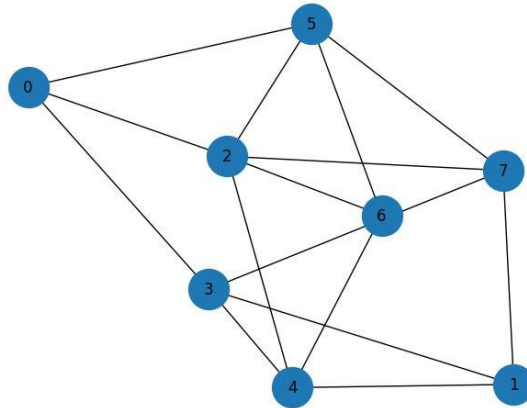
Ans:

$$JC(2, 6) = \frac{|\{5, 7, 4\}|}{|\{5, 7, 4, 0, 3\}|} = \frac{3}{5 + 5 - 3} = \frac{3}{7}$$

Equation AA: score (x, y) = $\sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$, with $\log(4) \approx 0.6$

Ans:

$$AA(2, 6) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|} = \frac{1}{\log |N(4)|} + \frac{1}{\log |N(5)|} + \frac{1}{\log |N(7)|} = \frac{3}{\log(4)} = \frac{3}{0.6} = 5$$



7. Consider an undirected graph G of four nodes given in the following figure, calculate Katz Index with L = 2, $\beta = 0.5$ (5pt)

Equation: score (x, y) = $\sum_{l=1}^L \beta^l |paths_{xy}^{(l)}| = \beta A_{xy} + \beta^2 A_{xy}^2 + \dots + \beta^L A_{xy}^L$, where $A^2 = A * A$, which A is adjacency matrix

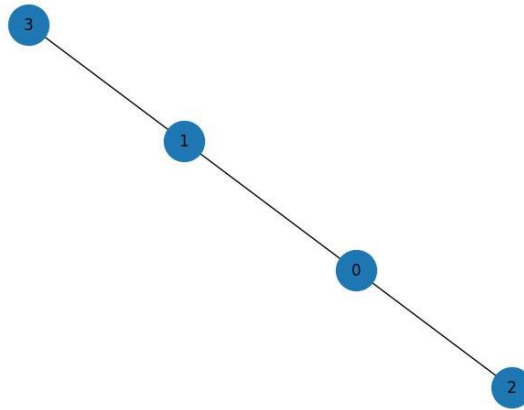
Ans:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$(x, y) = \sum_{l=1}^L \beta^l |paths_{xy}^{(l)}| = \beta A_{xy} + \beta^2 A_{xy}^2 = 0.5 \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} +$$

$$0.5^2 \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \end{pmatrix} + 0.25 \begin{pmatrix} 2 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0.5 & 0 & 0 & 0.25 \\ 0 & 0.5 & 0.25 & 0 \\ 0 & 0.25 & 0.25 & 0 \\ 0.25 & 0 & 0 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.25 \\ 0.5 & 0.5 & 0.25 & 0.5 \\ 0.5 & 0.25 & 0.25 & 0 \\ 0.25 & 0.5 & 0 & 0.25 \end{pmatrix}$$



8. Consider an undirected graph G of three nodes given in the following figure, calculate Hitting time of node 1 and node 2 (5pt).

Equation Hitting time: score $(x, y) = -H_{x,y} = -\frac{1}{|N(x)|} \sum_k (1 + H_{k,y})$,

where $H(k, y) = 1 + \sum_m p_{mj} H(m, y)$ when $k \neq y$, otherwise $H(k, y) = 0$, p_{mj} is the element in the row m-th and column j-th of the matrix, $P = AD^{-1}$, which P is a transition matrix, A is adjacency matrix and D is degree matrix.

Ans:

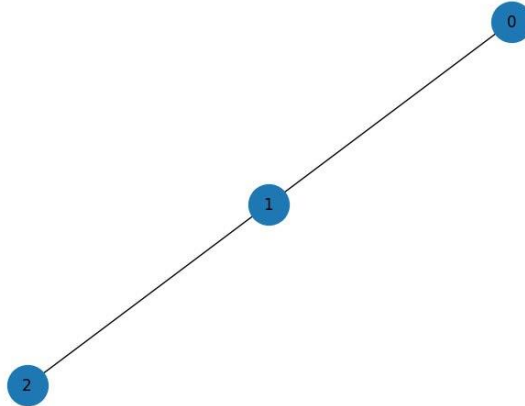
$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

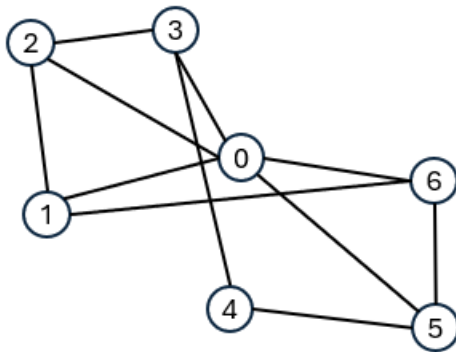
$$P = AD^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} * \frac{1}{\det D} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} * \frac{1}{2} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}$$

$$H(1,2) = 1 + \sum_m p_{mj} H(m,y) = 1 + 0 = 1 \quad (m = y)$$

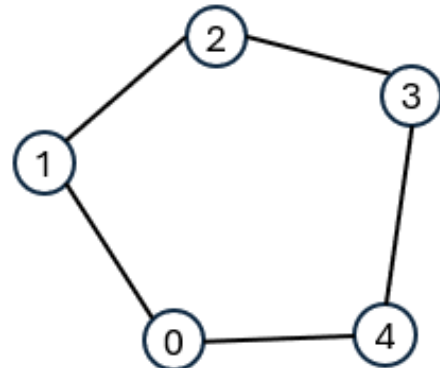
$$\text{score}(1,2) = -H_{x,y} = -\frac{1}{|N(x)|} \sum_k (1 + H_{k,y}) = -\frac{1}{2} (1 + 1) = -1$$



9. Consider two undirected graphs G_1 and G_2 below, calculate the graph edit distances from G_1 to G_2 . The set of elementary operations: vertex insertion, vertex deletion, edge insertion, and edge deletion. In addition, the cost of insertion and deletion is 2 and 1, respectively. (5pt)



G_1



G_2

Ans:

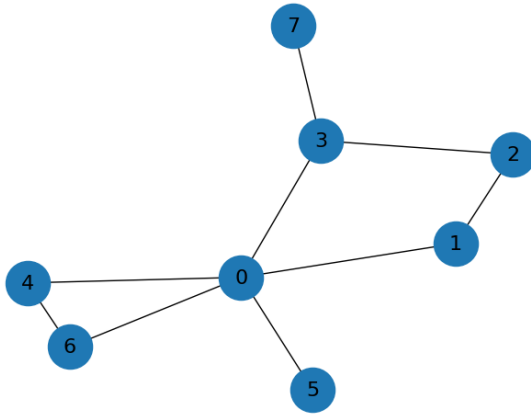
Remove node 1 and 2: cost 2.

Delete Edges (0, 1), (0, 2), (0, 5), (1, 2), (1, 6), and (2, 3): cost 6.

Hence, the cost distance = 2+6 = 8.

10. Consider an undirected graph G of eight nodes in the left side and four graphlets kernel g_1, g_2, g_3, g_4 in the right side of following figure. Answer two question below: (10pt)

- Count the number of the kernel sub-graphs of limited size 3.
- Make a feature vector for graph G based on these graphlets kernel.



Ans:

- Count subgraph:

$$N(g_1)=1:$$

(0, 4, 6)

$$N(g_2)=11:$$

(0, 1, 2), (0, 1, 3), (0, 1, 6),

(0, 3, 2), (0, 3, 7),

(0, 4, 5), (0, 5, 1), (0, 5, 6), (0, 6, 3),

(2, 1, 3),

(3, 2, 7)

$$N(g_3)=18:$$

(0, 1, 7),

(0, 4, 2), (0, 4, 7),

(0, 5, 2), (0, 5, 7),

(0, 6, 2), (0, 6, 7),

(1, 2, 4), (1, 2, 5), (1, 2, 6), (1, 2, 7),

(2, 3, 5), (2, 3, 4), (2, 3, 6),

(3, 7, 1), (3, 7, 4), (3, 7, 5), (3, 7, 6),

(4, 6, 5), (4, 6, 1), (4, 6, 2), (4, 6, 3), (4, 6, 7)

$$N(g_4)=15$$

(0, 2, 7),

(1, 3, 4), (1, 3, 5), (1, 3, 6),

(1, 5, 6), (1, 5, 7),

(2, 4, 7), (2, 5, 7), (2, 5, 6), (2, 6, 7),

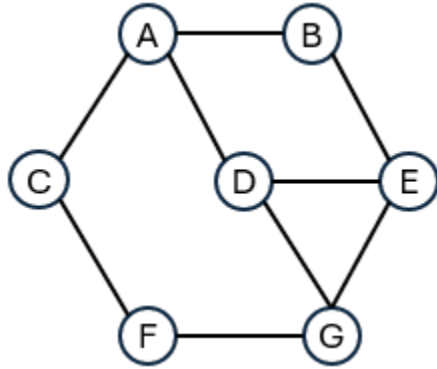
(3, 4, 5), (3, 5, 6),

(4, 1, 5), (4, 2, 5), (4, 5, 7)

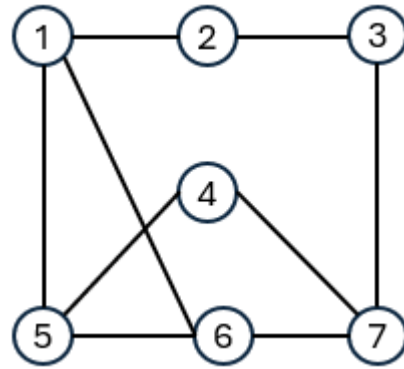
b. $f_G = (1, 11, 23, 15)$

11. Consider two undirected graphs in the following figure: G_1 on the left and G_2 on the right. (10pt)

- Conduct Weisfeiler-Lehman (WL) relabeling process with the maximum degree 3.
- Make feature vectors for the graphs based on frequency of the WL subgraphs in question a. Then calculate the similarity of graph G_1 and G_2 using equation (1).



G_1



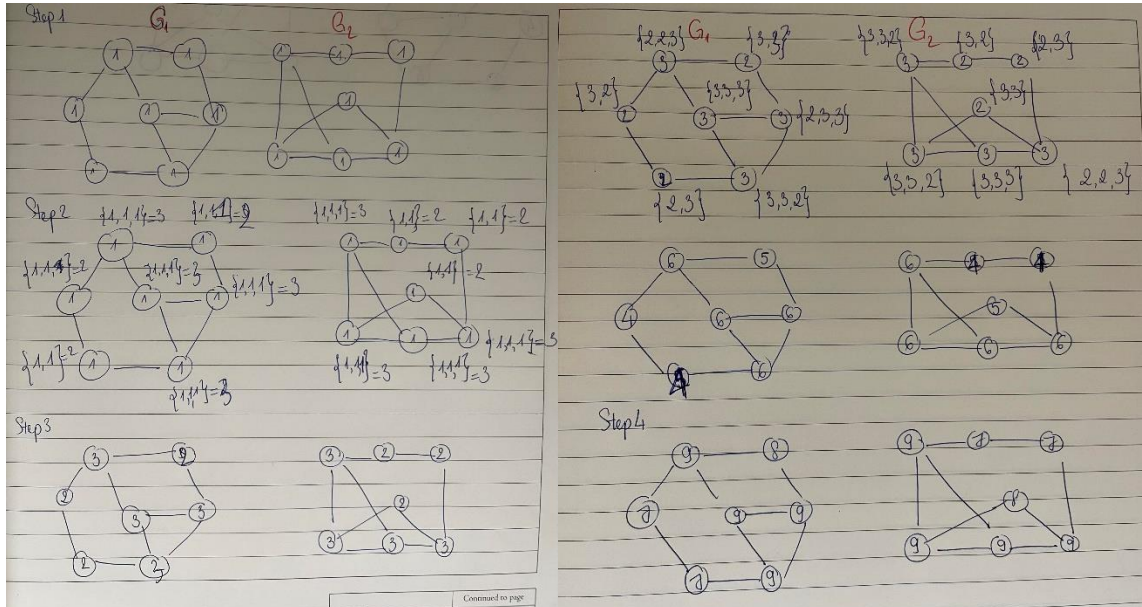
G_2

Cosine Similarity equation 1: $\text{cosine}(WL_{G_1}, WL_{G_2}) = \frac{WL_{G_1} \cdot WL_{G_2}}{\|WL_{G_1}\| \|WL_{G_2}\|}$.

where WL_{G_1} and WL_{G_2} is feature vectors of WL subgraph G_1 and G_2 . “.” denotes the dot product and “ $\| \cdot \|$ ” denotes the Euclidean norm.

Ans:

a.



b. Based on the WL relabeling process in (a), the number of WL subgraphs in G_1 is as follows:

The number of label “1”: 7

The number of label “2”: 3

The number of label “3”: 4

The number of label “4”: 2

The number of label “5”: 1

The number of label “6”: 4

The number of label “7”: 2

The number of label “8”: 1

The number of label “9”: 4

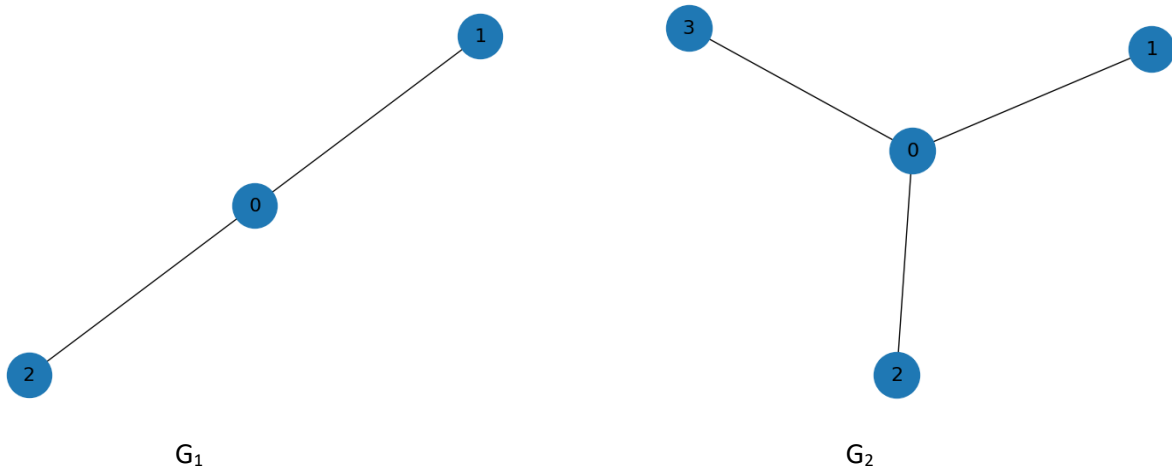
Therefore, the feature vector for G_1 is: (7,3,4,2,1,4,2,1,4)

Similarly, the feature vector for G_2 is: (7,3,4,2,1,4,2,1,4)

The similarity of G_1 and G_2 is

$$\frac{7 * 7 + 3 * 3 + 4 * 4 + 2 * 2 + 1 * 1 + 4 * 4 + 2 * 2 + 1 * 1 + 4 * 4}{\sqrt{7^2 + 3^2 + 4^2 + 2^2 + 1^2 + 4^2 + 2^2 + 1^2 + 4^2} \cdot \sqrt{7^2 + 3^2 + 4^2 + 2^2 + 1^2 + 4^2 + 2^2 + 1^2 + 4^2}} = 1$$

12. Consider an undirected graph G_1 and G_2 in the following figure. Make feature vectors of graphs G_1 and G_2 using the shortest path kernel and calculate similarity of graphs using the cosine similarity. (10pt)



Ans:

All shortest path Floyd-transformed S_1 of G_1 :

+ 0 – 1: 1.

+ 0 – 2: 1

+ 1 – 2: 2

⇒ Frequencies of length 1: 2

Frequencies of length 2: 1

All shortest path Floyd-transformed S_2 of G_2 :

+ 0 – 1: 1.

+ 0 – 2: 1

+ 0 – 3: 1

+ 1 – 2: 2

+ 1 – 3: 2

+ 2 – 3: 2

⇒ Frequencies of length 1: 3

Frequencies of length 2: 3

The feature vector of $S_1 = (2, 1)$

The feature vector of $S_2 = (3, 3)$

The similarity = $\frac{2*3+1*3}{\sqrt{4+1}.\sqrt{9+9}} = \frac{9}{\sqrt{5}\sqrt{18}} \approx 0.9847$