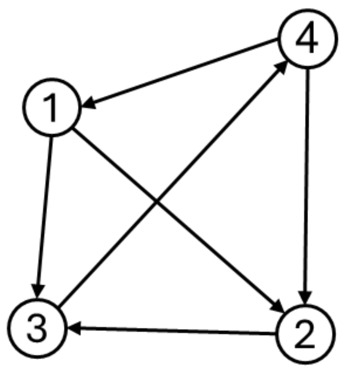
Final Exam (Graph Mining – Spring 2024)

Full Name:
Student ID:

- The formula and solution process should be presented with the answer.
- The answer should be written in English.

1. Consider a directed graph G of four nodes given in the following figure, calculate PageRank centrality of all nodes, with $x_0 = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$ and β = 0.85. (10pt)



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

Ans:

a. Betweenness centrality of node 1:

	$\sigma(j,k)$	$\sigma(j,k\|i)$	$\sigma(j,k\|i)/\sigma(j,k)$
1,2	1	0	0
1,3	2	0	0
1,4	1	0	0
1,5	2	0	0
2,3	1	0	0
2,4	2	1	1/2
2,5	1	0	0
3,4	2	0	0
4,5	1	0	0

$$\bar{B}(v_i) = \frac{\mathrm{B}(v_i)}{(\mathrm{n}-1)\,(\mathrm{n}-2)/2} = \frac{1/2}{(5-1)\,(5-2)/2} = \frac{1/2}{4*3/2} = \frac{1}{12} \approx 0.0833$$

- **Closeness centrality of node 1:**

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

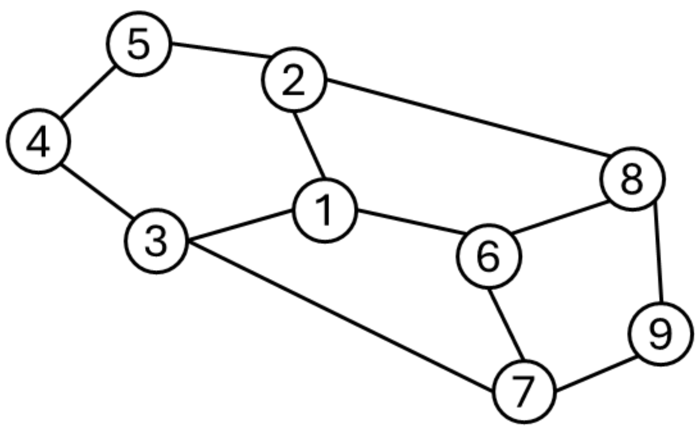
b. PageRank score equally 4 pages

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

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

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

2. Consider an undirected graph G of nine nodes given in the following figure. There are two communities in the graph: A = {1, 2, 3, 4, 5} and B = {6, 7, 8, 9}. Calculate the Normalized-cut measurement and conductance of A and B. The conductance is referred to in Equation (1). (5pt)



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

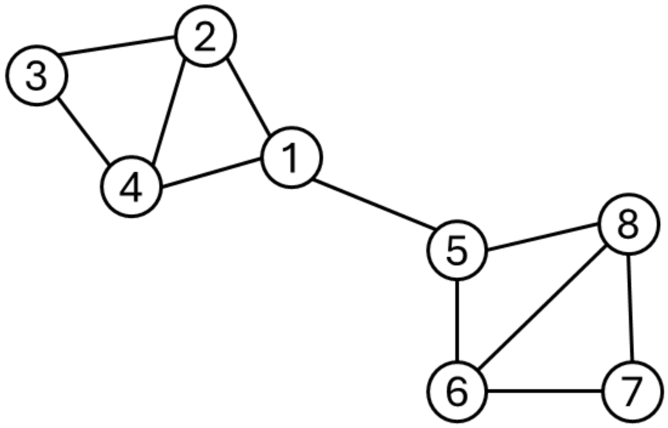
where $assoc(A, V)$ and $assoc(B, V)$ is the total connection from nodes in A and B to all nodes in the graph, respectively. $cut(A, B)$ is the number of cuts between 2 communities A and B.

Ans:

Min_cut (A, B) = $\frac{3}{1+5} + \frac{3}{1+4} = \frac{33}{30} = \frac{11}{10} = 1.1$

Conductance (A, B) = $\frac{3}{\min\,(6,5)} = \frac{3}{5} = 0.6$

3. Consider an undirected graph G of eight nodes given in the following figure with two communities: B = {1, 2, 3, 4} and C = {5, 6, 7, 8}. Apply the Equation (1) to calculate the modularity Q of the two communities. (10pt)



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

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

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

<u>Ans</u>:

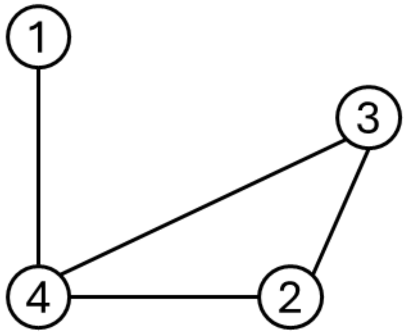
$$Q = \frac{1}{2 \times m}\sum_{i,j}\left(A_{ij} - \frac{d_i d_j}{2m}\right) \cdot \delta(v_i, v_j)$$

$$Q = \frac{1}{2 \times 11}\left[\left(1 - \frac{d_1 d_2}{22}\right) + \left(1 - \frac{d_2 d_3}{22}\right) + \left(1 - \frac{d_3 d_4}{22}\right) + \left(1 - \frac{d_2 d_4}{22}\right) + \left(1 - \frac{d_4 d_1}{22}\right) + \left(1 - \frac{d_8 d_5}{22}\right) + \left(1 - \frac{d_5 d_6}{22}\right) + \left(1 - \frac{d_6 d_7}{22}\right) + (1 - \frac{d_6 d_8}{22}) + \left(1 - \frac{d_7 d_8}{22}\right)\right]$$

$$Q = \frac{1}{2 \times 11}\left[\left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{3 \times 2}{22}\right) + \left(1 - \frac{2 \times 3}{22}\right) + \left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{3 \times 2}{22}\right) + \left(1 - \frac{3 \times 3}{22}\right) + \left(1 - \frac{2 \times 3}{22}\right)\right]$$

$$Q = \frac{1}{2 \times 11}\left[6\left(1 - \frac{3 \times 3}{22}\right) + 4\left(1 - \frac{3 \times 2}{22}\right)\right] = \frac{71}{242} \approx 0.2934$$

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



Katz index equation: score $(x, y) = \sum_{l=1}^{L} \beta^l \left|paths_{xy}^{(l)}\right| = \beta A_{xy} + \beta^2 A_{xy}^2 + \cdots + \beta^L A_{xy}^L$, where $A^2 = A * A$ and A is adjacency matrix of graph G.

<u>Ans:</u>

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

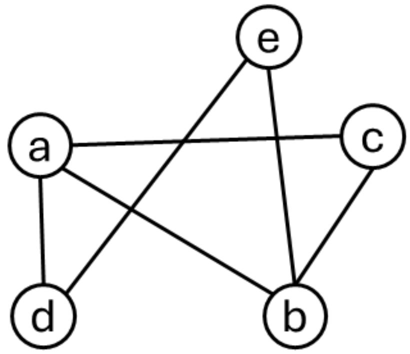
$$(x, y) = \sum_{l=1}^{L} \beta^l \left|paths_{xy}^{(l)}\right| = \beta A_{xy} + \beta^2 A_{xy}^2 = 0.5 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} + 0.5^2 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}^2$$

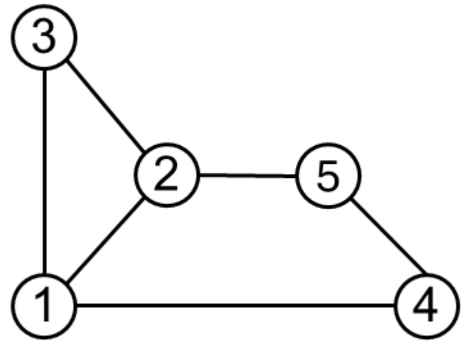
$$= \begin{pmatrix} 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0 \end{pmatrix} + 0.25 \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0 \end{pmatrix} + \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0 \\ 0.25 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0.25 \\ 0 & 0.25 & 0.25 & 0.75 \end{pmatrix}$$

$$= \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.75 & 0.75 \\ 0.25 & 0.75 & 0.5 & 0.75 \\ 0.5 & 0.75 & 0.75 & 0.75 \end{pmatrix}$$

5. Consider two undirected graphs G_1 and G_2 in the following figure. (10pt)
   a) Conduct Weisfeiler-Lehman (WL) relabeling process with the maximum degree 3. Initial labels of every node are "1".
   b) Calculate the Cosine similarity of graph G_1 and G_2 using Equation (1) by feature vectors based on frequency of the WL subgraphs from the result of question a.





G1 G2

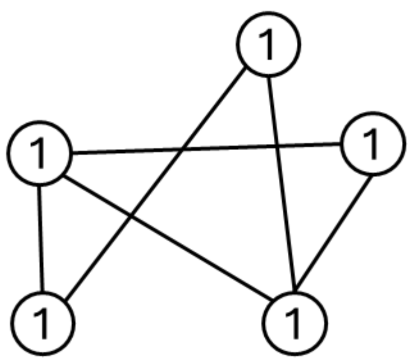
Cosine Similarity equation 1: $cosine(WL_{G_1}, WL_{G_2}) = \frac{WL_{G_1}.WL_{G_2}}{\|WL_{G_1}\|\|WL_{G_2}\|}$.

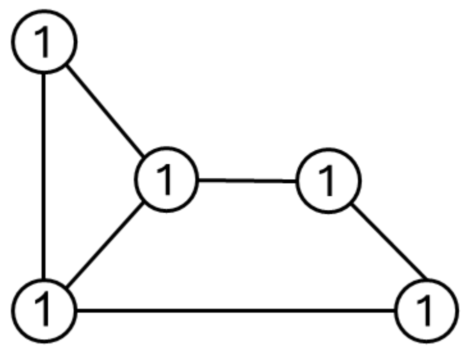
where WL_{G_1} and WL_{G_2} is feature vectors of WL subgraph G_1 and G_2. "." denotes the dot product and "‖ ‖" denotes the Euclidean norm.

<u>Ans</u>:

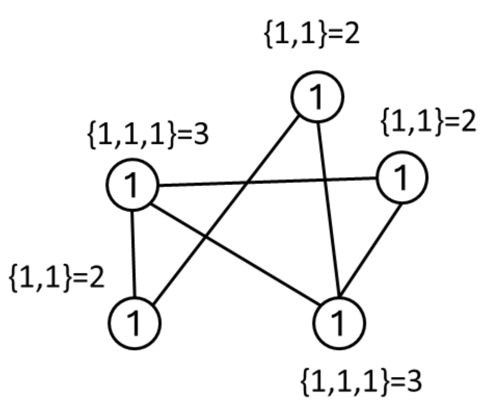
a. Relabeling process:

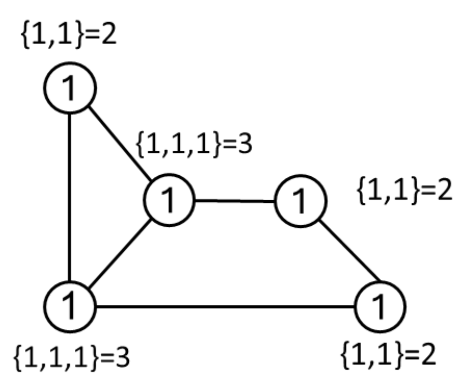
- Step 1:

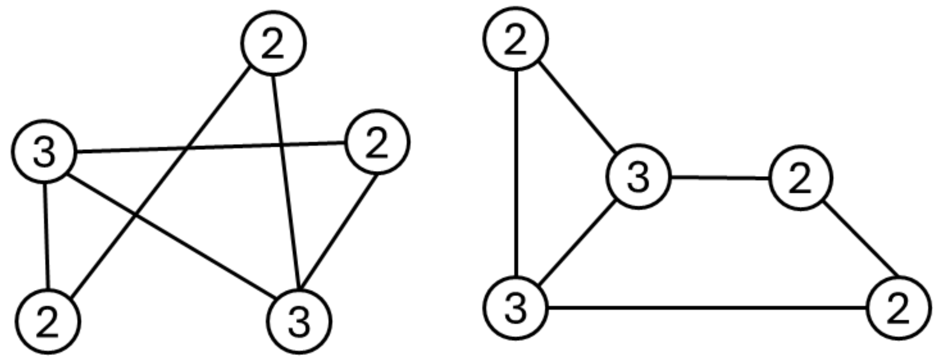




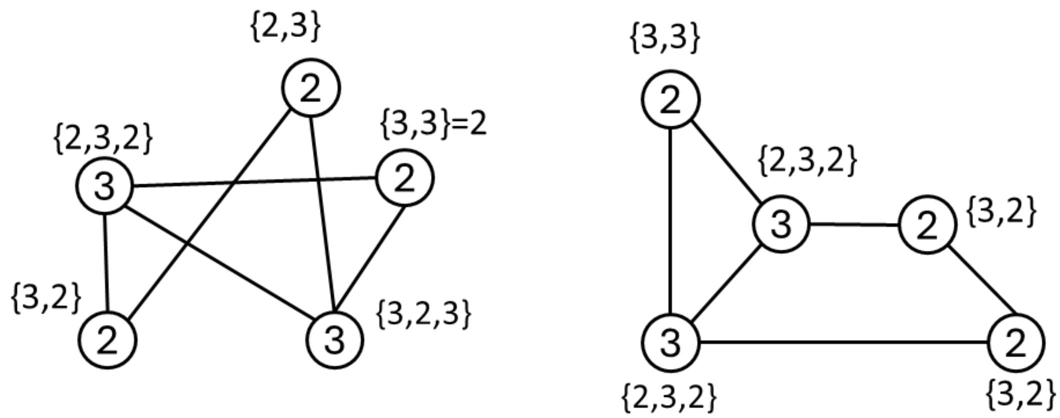
- Step 2:

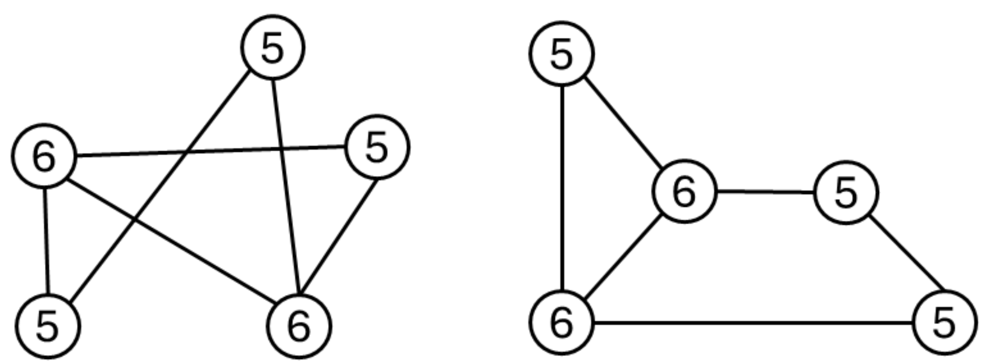




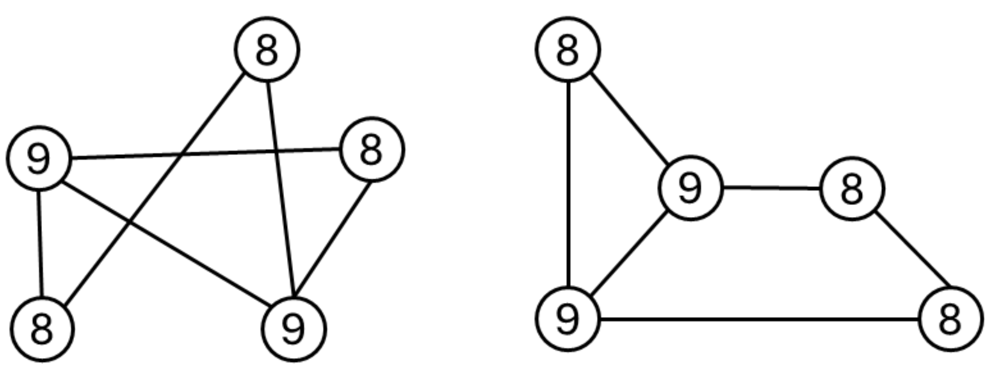


- Step 3:





- Step 4:



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

The number of label "a": 9

The number of label "b": 9

The number of label "c": 8

The number of label "d": 8

The number of label "e": 8

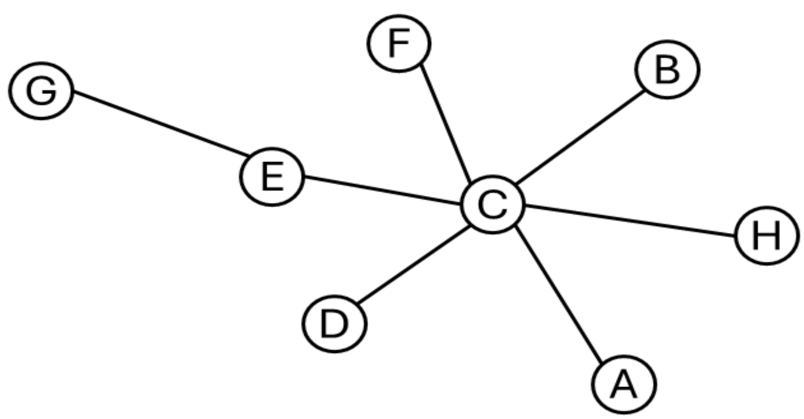
Therefore, the feature vector for G_1 is: (9, 9, 8, 8, 8)

Similarly, the feature vector for G_2 is: (9, 9, 8, 8, 8)

The similarity of G_1 and G_2 is

$$\frac{9*9+9*9+8*8+8*8+8*8}{\sqrt{9^2+9^2+8^2+8^2+8^2}.\sqrt{9^2+9^2+8^2+8^2+8^2}} = 1$$

6. Consider an undirected graph with eight nodes in the following figure. A biased random walk (Node2Vec algorithm) has the return parameter $p = 0.5$ and the in-out parameter $q = 0.5$. Assume that all edge weights of the graph are 1 and the walker is currently on node C by departing from node E. Calculate transition probabilities from node C to its neighbors. (10pt)



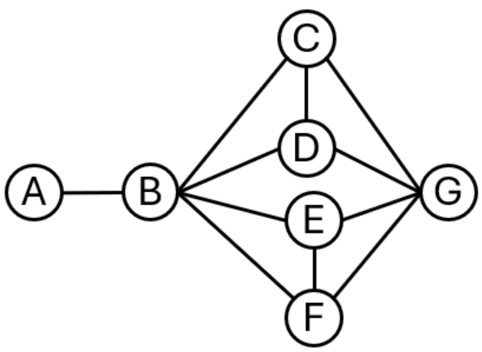
Ans:

There are four neighbors of node C: A, B, D, E, F, and H. Since the walker starts from E to C, the transition probabilities from C to its neighbors can be calculated as follows:

$$P_{C\to A} = 1 \times \frac{1}{q} = \frac{1}{0.5} = 2$$
$$P_{C\to G} = 1 \times 1 = 1$$
$$P_{C\to F} = 1 \times \frac{1}{p} = \frac{1}{0.5} = 2$$
$$P_{C\to D} = 1 \times \frac{1}{p} = \frac{1}{0.5} = 2$$

$$P_{C\to H} = 1 \times \frac{1}{p} = \frac{1}{0.5} = 2$$
$$P_{C\to A} = 1 \times \frac{1}{p} = \frac{1}{0.5} = 2$$
$$P_{C\to B} = 1 \times \frac{1}{p} = \frac{1}{0.5} = 2$$

7. Consider an undirected graph G of seven nodes A, B, C, D, E, F, and G given in the following figure. Let x_i is the initial vector representations of a node i, as shown in Eq. 1. (10pt)



$$x_i = \left(w_{i1}, w_{i2}, \ \dots, \ w_{i|V|}\right) \quad (1)$$

where $w_{ik} = \begin{cases} 1 & \text{if } (i,k) \in E, \\ 0 & \text{otherwise} \end{cases}$,
|V| denotes the number of nodes in the graph.

a) Calculate the initial vectors of all the nodes in graph G based on Eq. 1.
b) Calculate the second-order proximity between pairs of nodes (A, C) and (B, G) based on Manhattan Distance (the distance between two data points is computed as $D_{(x,y)} = \sum_{i=1}^{n}|x_i - y_i|$, where n is the number of dimensions).

Ans:

a)

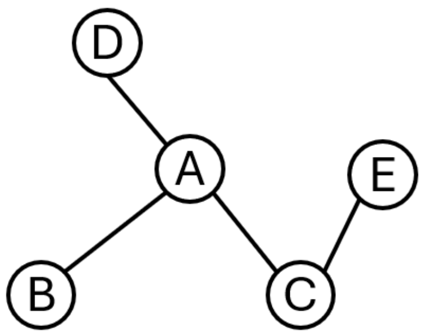
$$x_A = (0,1,0,0,0,0,0)$$
$$x_B = (1,0,1,1,1,1,0)$$
$$x_C = (0,1,0,1,0,0,1)$$
$$x_D = (0,1,1,0,0,0,1)$$
$$x_E = (0,1,0,0,0,1,1)$$
$$x_F = (0,1,0,0,1,0,1)$$
$$x_G = (0,0,1,1,1,1,0)$$

b)

$$D_{AC} = \sum_{i=1}^{n}\left|x_{A,i} - x_{C,i}\right|$$

$$= |(0,1,0,0,0,0,0) - (0,1,0,1,0,0,1)|$$
$$= 2$$
$$D_{BG} = \sum_{i=1}^{n} |x_{B,i} - x_{G,i}|$$
$$= |(1,0,1,1,1,1,0) - (0,0,1,1,1,1,0)|$$
$$= 1$$

8. Consider an undirected graph G of five nodes A, B, C, D, and E given in the following figure. (10pt)



Equation (1):

$$S = (M_g)^{\mathrm{T}} \cdot M_l,$$

$$M_g = I - \beta \cdot A,$$

$$M_l = \beta \cdot A\ ,$$

where I refers to the Identity matrix.

From the HOPE method (Asymmetric Transitivity Preserving Graph Embedding), a high-order proximity matrix S is defined in Eq. (1). Calculate the S matrix based on the Katz proximity measurement with $\beta = 1$.

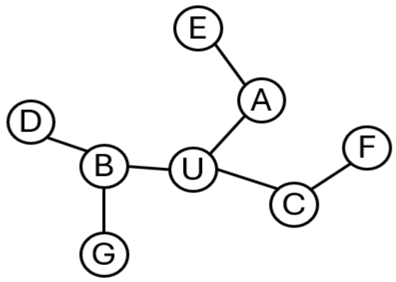
Ans:

$$S = \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \right) \left(\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \right)$$

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

$$= \begin{bmatrix} \square & \mathbf{-3} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{-1} \\ \square & \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} \\ \square & \mathbf{1} & \mathbf{-1} & \mathbf{-2} & \mathbf{-1} & \mathbf{1} \\ \square & \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} \\ \square & \mathbf{-1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{-1} \end{bmatrix}$$

9. Consider an undirected, unweighted graph given in the following figure. From the Struc2Vec method, let $R_k(U)$ denote the set of neighbor nodes within k-hop distance rooted at node U. Let S(v) denotes the ordered degree sequence of a node set $v \subset V$ (from the minimum to maximum values). Let $f_k(u,v)$ denotes the structural distance between u and v. (10pt)



$f_k(u,v) = f_{k-1}(u,v) + g\big(S(R_k(u)), S(R_k(v))\big)$ (1)

where *g(.)* measures the distance between the ordered degree sequences, which is based on the Manhattan Distance ($g(x,y) = \sum_{i=1}^{n} |x_i - y_i|$, with n is the number of dimensions).

$f_0(u,v) = -1$

a) Calculate $R_0(U)$, $R_1(U)$, $S(R_0(U))$, and $S(R_1(U))$.
b) Calculate the structural distance $f_1(E,D)$ between two nodes E and D

Ans:

a)

$R_0(U) = \{U\}$

$R_1(U) = \{A, B, C\}$

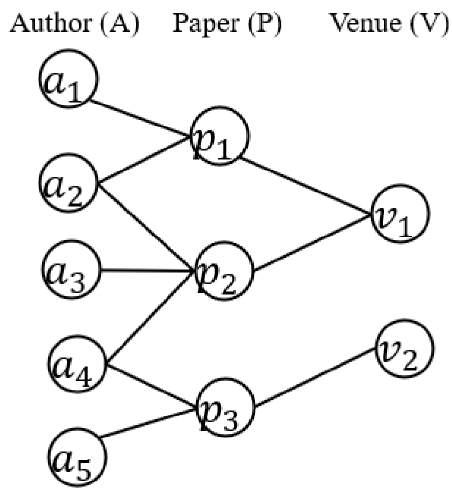
$S(R_0(U)) = \{3\}$

$S(R_1(U)) = \{2,2,3\}$

b)

$$f_1(E,D) = f_0(E,D) + g\Big(S\big(R_0(E)\big); S\big(R_0(D)\big)\Big)$$
$$= -1 + g(\{3\}, \{2\})$$
$$= -1 + 1$$
$$= 0$$

10. Consider a heterogeneous graph given in the following figure. There are three types of nodes in the academic network: *Author* (A), *Paper* (P), and *Venue* (V). List all the meta-path APA and APVPA. (10pt)



Ans:

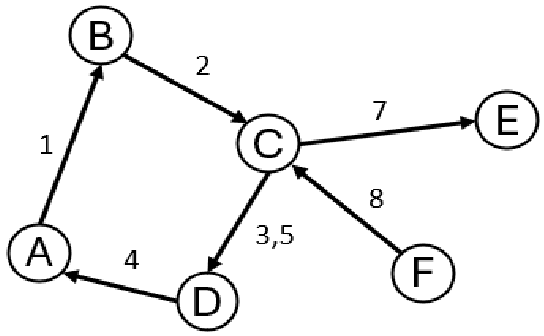
APA:

$$a_1 p_1 a_2$$
$$a_2 p_1 a_1$$
$$a_2 p_2 a_3$$
$$a_2 p_2 a_4$$
$$a_3 p_2 a_2$$
$$a_3 p_2 a_4$$
$$a_4 p_2 a_2$$
$$a_4 p_2 a_3$$
$$a_4 p_3 a_5$$
$$a_5 p_3 a_4$$

APVPA

$$a_1 p_1 v_1 p_2 a_2$$
$$a_1 p_1 v_1 p_2 a_3$$
$$a_1 p_1 v_1 p_2 a_4$$
$$a_2 p_1 v_1 p_1 a_1$$
$$a_2 p_1 v_1 p_2 a_3$$
$$a_2 p_1 v_1 p_2 a_4$$
$$a_2 p_1 v_1 p_2 a_4$$
$$a_3 p_2 v_1 p_1 a_1$$
$$a_3 p_2 v_1 p_1 a_2$$
$$a_3 p_2 v_1 p_2 a_2$$
$$a_3 p_2 v_1 p_2 a_4$$
$$a_4 p_2 v_1 p_1 a_1$$
$$a_4 p_2 v_1 p_1 a_2$$
$$a_4 p_2 v_1 p_2 a_2$$
$$a_4 p_2 v_1 p_2 a_2$$
$$a_4 p_3 v_2 p_3 a_5$$

11. Consider a dynamic graph given in the following figure. The edges are labeled by time. (5pt)



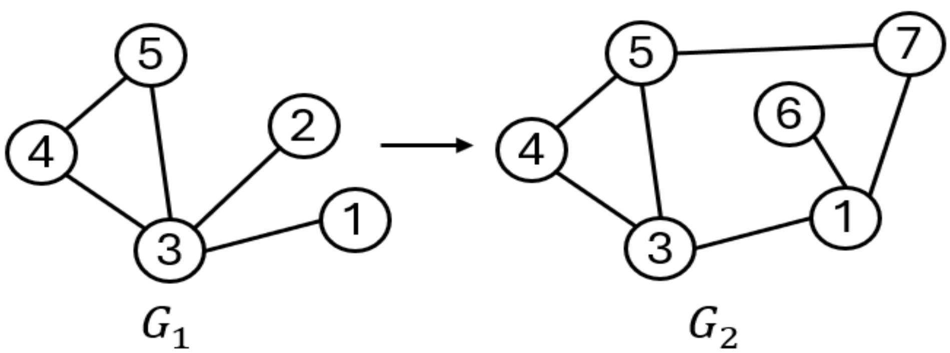
Equation (1):
$N_t(v) = \{(u, t')|e = (v, u, t') \in E_T \wedge T(e) > t\}$,
where *T(e)* refers to the timestamp of the edge *e*

a) From the CTDNE method, the temporal neighbors of a node *v* at time *t* can be computed as Eq. (1). Calculate the set of temporal neighbors of the node A at time *t* = 0.
b) List all the temporal random walks from node A to other nodes with length 3.

Ans:

a) $N_A^{t=0} = \{B\}$
b) ABCE; ABCD

12. Consider two snapshots of a dynamic graph with structural evolution from time *t=1* to *t=2*, as shown in the following figure. The evolving nodes in the timestamp *t* are defined as in Eq. 1 based on the Dynnode2vec method. (5pt)
   a) Calculate $V_{add}, E_{add}, V_{del},$ and E_{del} timestamp *t=2*.
   b) Calculate ΔV_2.



Equation (1): $\Delta V_t = V_{add} \cup \{v_i \in V_t | \exists e_i = (v_i, v_j) \in (E_{add} \cup E_{del})\}$, where V_{add} and E_{add} denote the sets of new nodes and edges that are added, respectively. V_{del} and E_{del} are the sets of new nodes and edges that are deleted, respectively.

Ans:

a)
$V_{add} = \{6,7\}$

$$E_{add} = \{e_{16}; e_{17}; e_{57}\}$$
$$V_{del} = \{2\}$$
$$E_{del} = \{e_{23}\}$$

b)

$$\mathit{\Delta}V_2 = \{6,7\} \cup \{1,5,6,7,3\}$$
$$= \{1,5,6,7,3\}$$