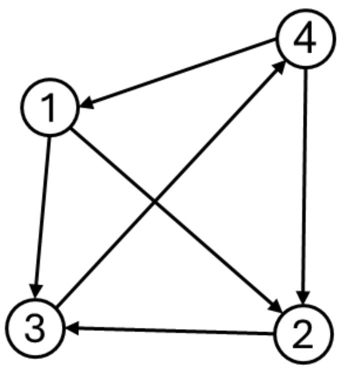
Final Exam (Graph Mining – Spring 2024)

Full Name:
Student ID:

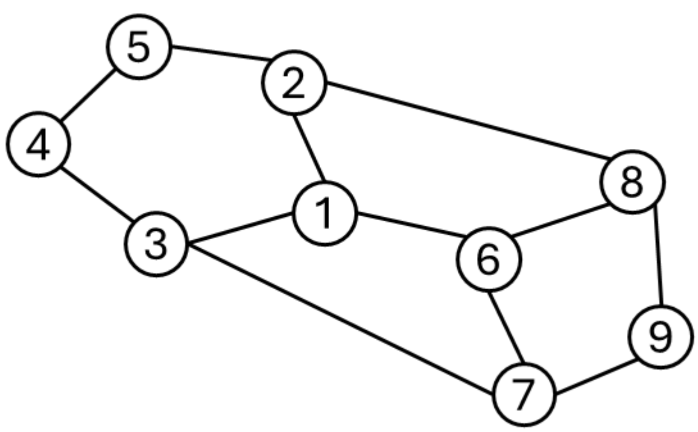
- The formula and solution process should be presented with the answer.
- The answer should be written in English.

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.

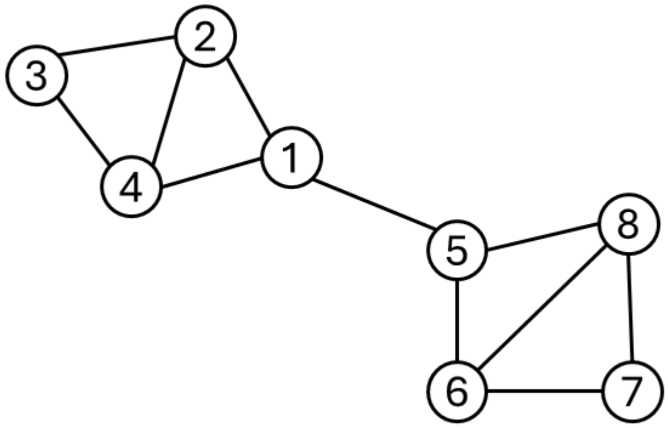
2. Consider an undirected graph G of nine nodes given in the following figure. There are two communities in the graph: A = {1, 2, 3, 4, 5} and B = {6, 7, 8, 9}. Calculate the Normalized-cut measurement and conductance of A and B. The conductance is referred to in Equation (1). (5pt)



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.

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)

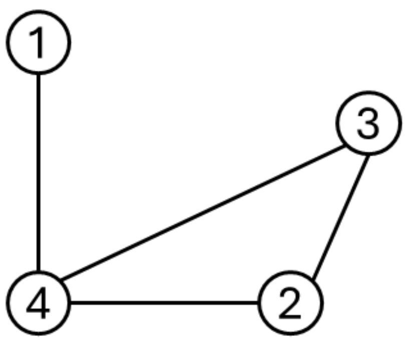


Equation (1): $Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \cdot \delta(v_i, v_j)$

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

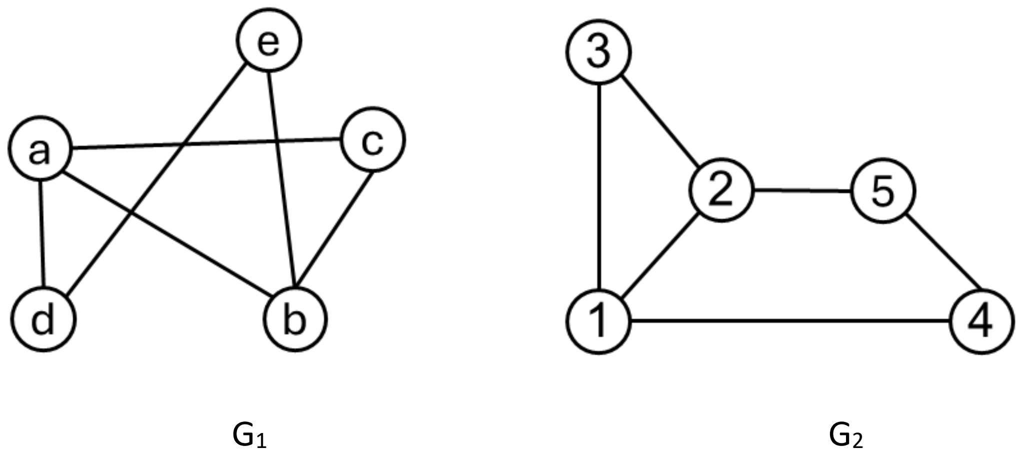
where m is the number of edges, A is the adjacency matrix of G, d_i is the degree of node v_i.

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.

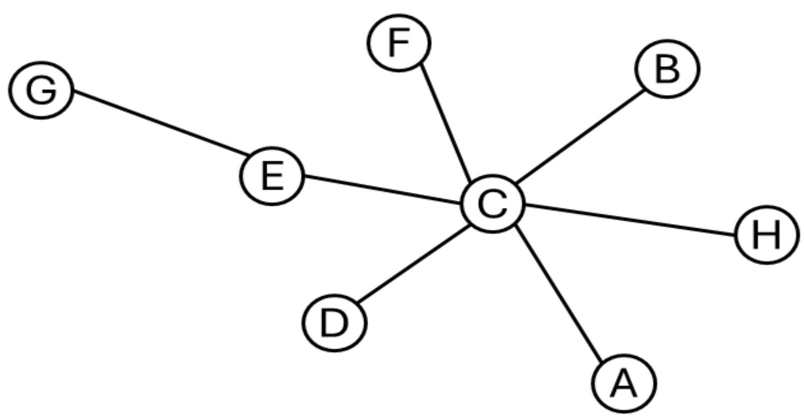
5. Consider two undirected graphs G_1 and G_2 in the following figure. (10pt)
   a) Conduct Weisfeiler-Lehman (WL) relabeling process with the maximum degree 3. Initial labels of every node are “1”.
   b) Calculate the Cosine similarity of graph G_1 and G_2 using Equation (1) by feature vectors based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



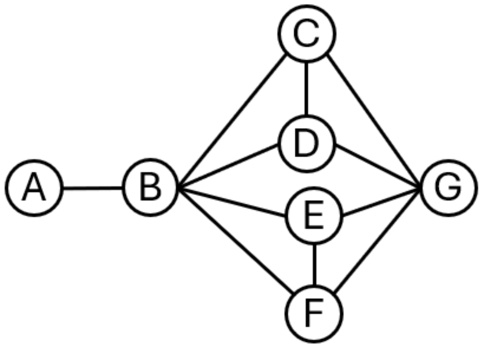
Cosine Similarity equation 1: $cosine(WL_{G_1}, WL_{G_2}) = \frac{WL_{G_1}.WL_{G_2}}{\|WL_{G_1}\|\|WL_{G_2}\|}$.

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.

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)



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)

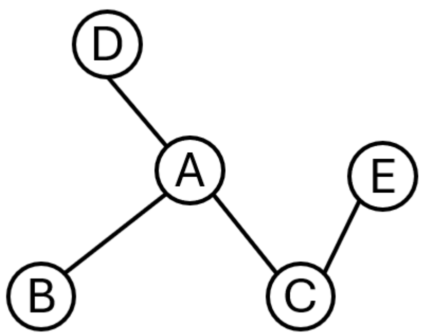


$$x_i = \left(w_{i1}, w_{i2}, \ \dots, \ w_{i|V|}\right) \quad (1)$$

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).

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



Equation (1):

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

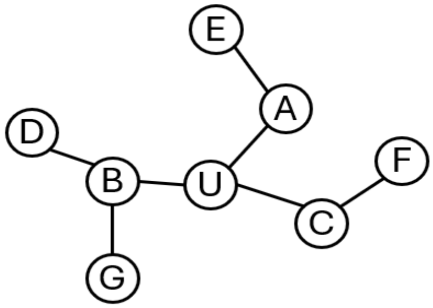
$$M_g = I - \beta \cdot A,$$

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

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 $\beta = 1$.

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)



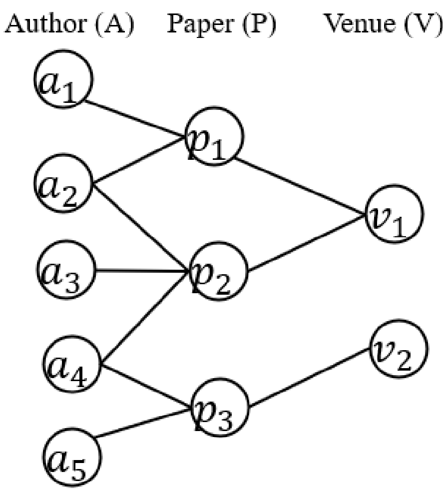
$$f_k(u, v) = f_{k-1}(u, v) + g\big(S(R_k(u)), S(R_k(v))\big) \quad (1)$$

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).

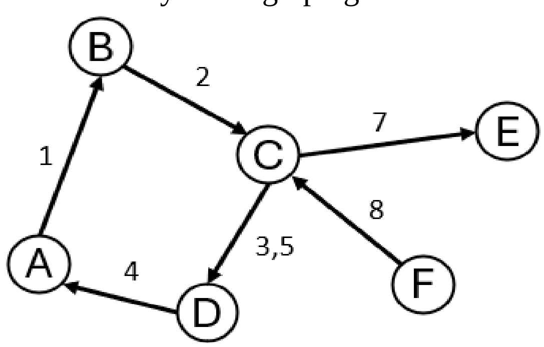
$f_0(u, v) = 0$

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.

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)



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

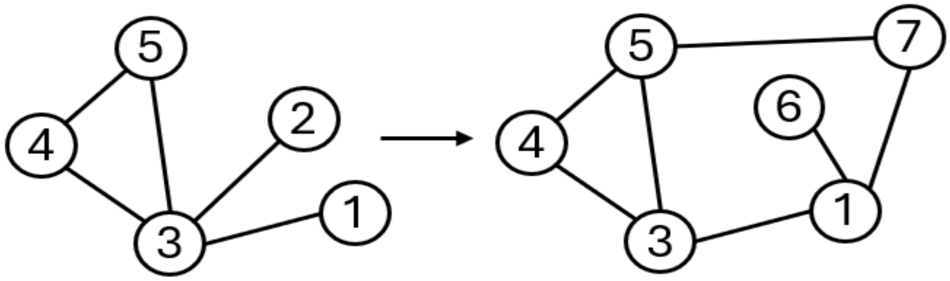


Equation (1):
$N_t(v) = \{(u, t')| e = (v, u, t') \in E_T \wedge T(e) > t\}$,
where *T(e)* refers to the timestamp of the edge *e*

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.

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} at timestamp *t=2*.
   b) Calculate ΔV_2.



G_1 G_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.