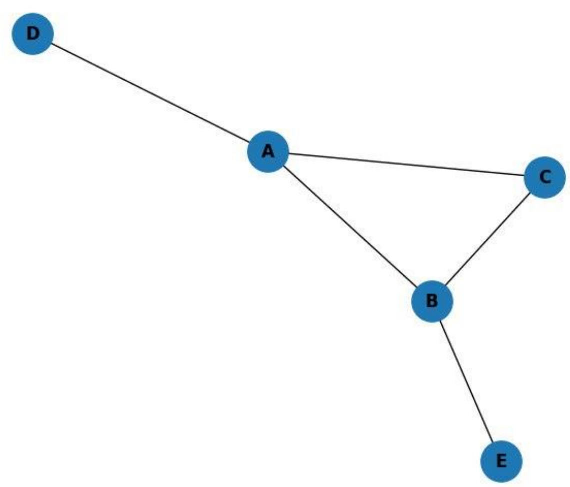
Final Exam (Graph Mining – Spring 2025)

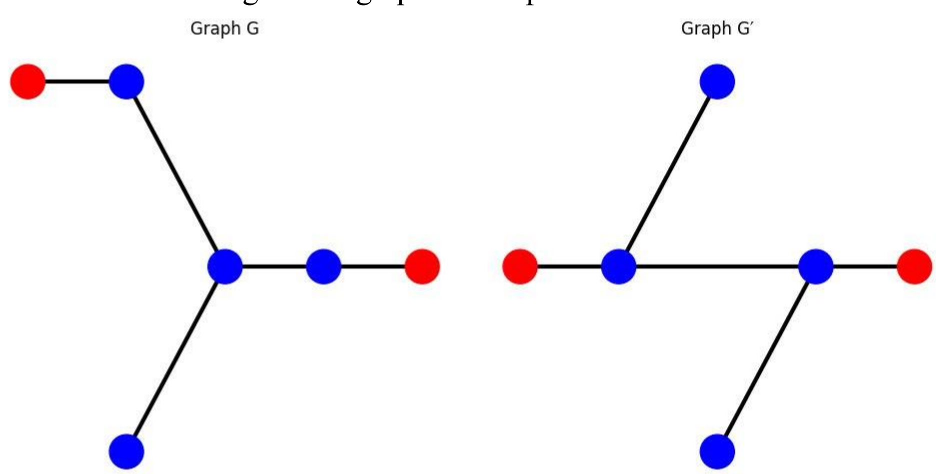
Full Name:
Student ID:

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

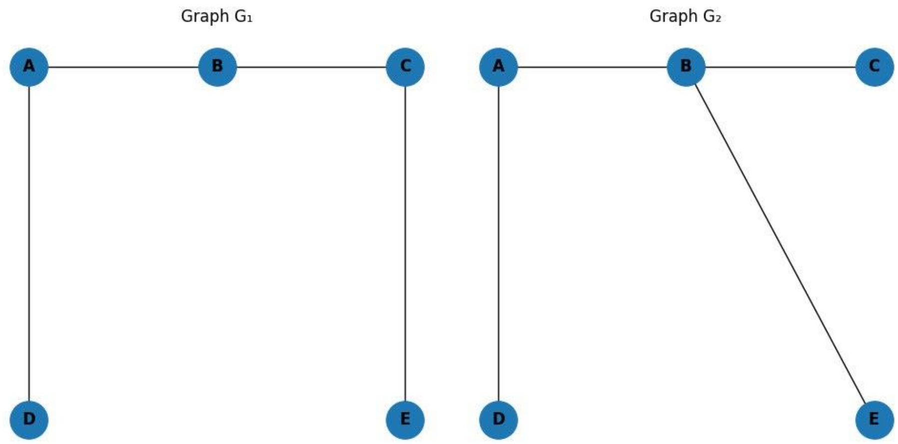
1. (Graphlet Kerel) (5pt)
   a) Given the example graph G with 5 nodes, compute the graphlet count vector f_G for graphlets of size k=3
   b) What is its main principle, what computational challenge does it face, and how does it address graph isomorphism when dealing with unlabeled versus labeled graphs?



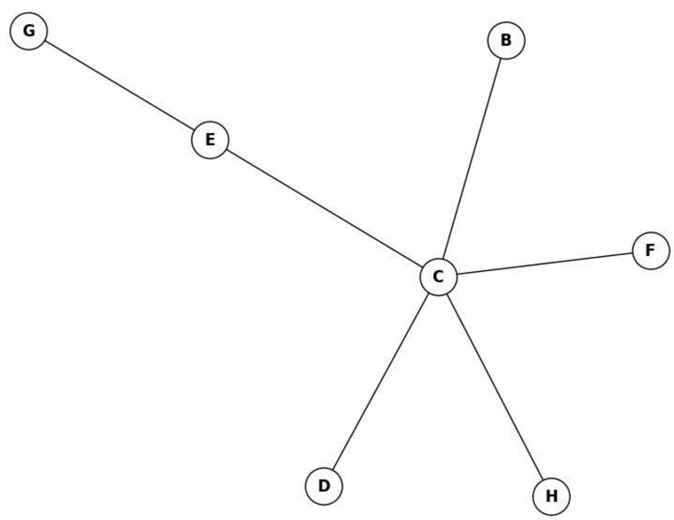
2. (Shortest Path Kernel) (5pt)
   a) Given the two graphs, apply the shortest path kernel. (label R as red node and B as blue node)
   b) Why is it faster than checking for subgraph isomorphisms?



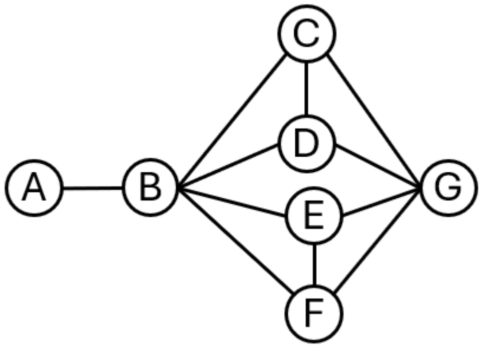
3. (WL Kernel) (5pt)
   a) Given two graphs, draw whether the graphs are isomorphic
   b) Explain the underlying idea of the Weisfeiler-Lehman kernel. How does the iterative relabeling process help capture graph structure?



4. (Node2Vec) (10pt)
   a) 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.
   b) Describe how Node2Vec uses two hyperparameters — the return parameter p and the in-out parameter q — to interpolate between Breadth-First Search (BFS) and Depth-First Search (DFS). How does each parameter influence the walk behavior?



5. (LINE) 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)

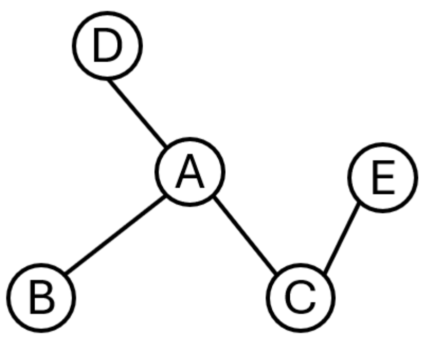


$$x_i = (w_{i1}, w_{i2}, \ \dots, \ w_{i|V|}) \quad (1)$$

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, D) 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).
c) How does LINE separately model and preserve both proximities in its objective functions?

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



Equation (1):

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

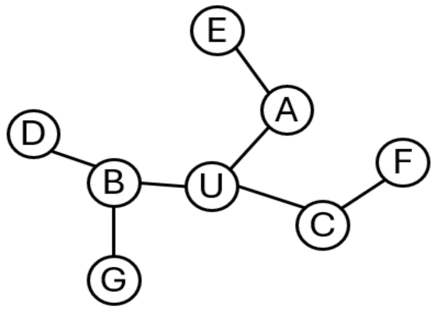
$$M_g = I - \beta \cdot A,$$

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

where *I* refers to the Identity matrix.

a) 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 $\beta = 1$.
b) What specific challenge does HOPE aim to solve that is not adequately addressed by methods like LINE or DeepWalk?

7. (Struc2Vec) 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)

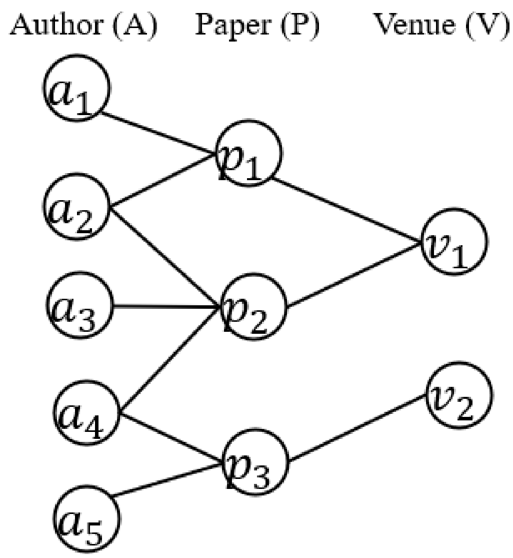


$$f_k(u, v) = f_{k-1}(u, v) + g\big(S(R_k(u)), S(R_k(v))\big) \quad (1)$$

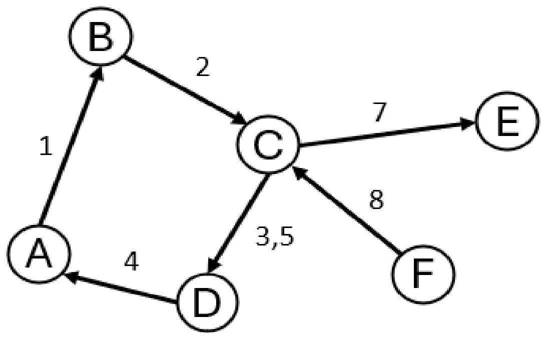
where *g(.)* measures the distance between the ordered degree sequences, which is based on the Manhattan Distance ($g(x, y) = \sum_{i=1}^{n} |x_i - y_i|$, with n is the number of dimensions).
$f_0(u, v) = 0$

a) Calculate $R_0(B)$, $R_1(B)$, $S(R_0(B))$, and $S(R_1(B))$.
b) Calculate the structural distance $f_1(E, G)$ between two nodes E and G.
c) Why is structural identity important in certain graph learning tasks?

8. (metapath2vec) 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). (10pt)
   a) List all the meta-path APA and APVPA.
   b) How does meta-path-guided sampling help capture semantic and structural correlations between heterogeneous node types?



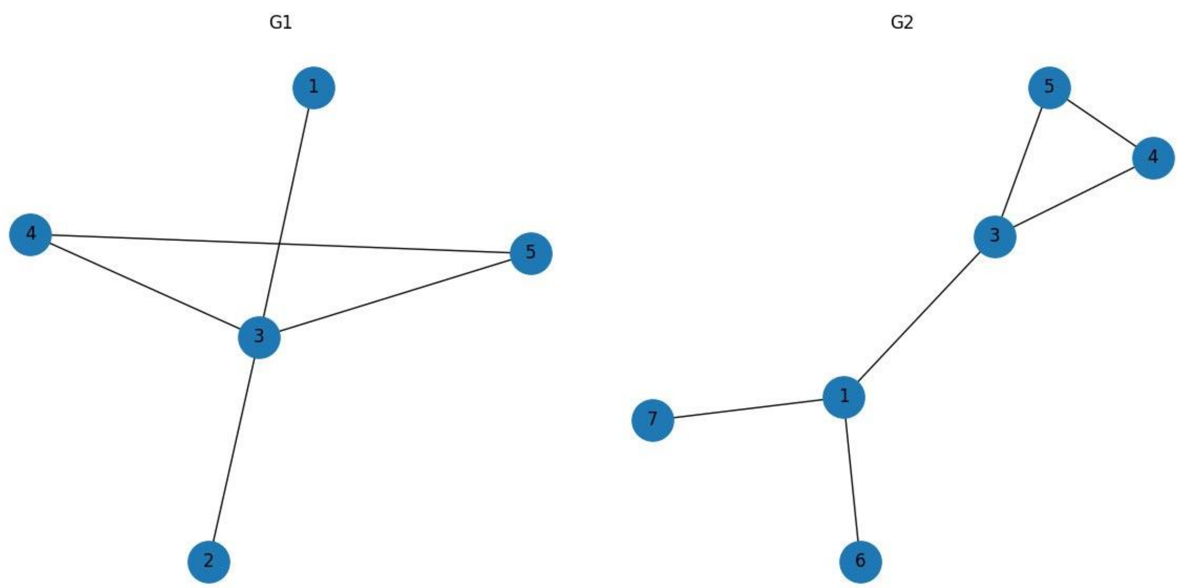
9. (CTDNE) Consider a dynamic graph given in the following figure. The edges are labeled by time. (10pt)



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 D at time *t* = 0.
b) List all the temporal random walks from node D to other nodes with length 4.
c) What are the main limitations of static embedding methods like DeepWalk, Node2Vec, or LINE when applied to temporal (time-evolving) networks?

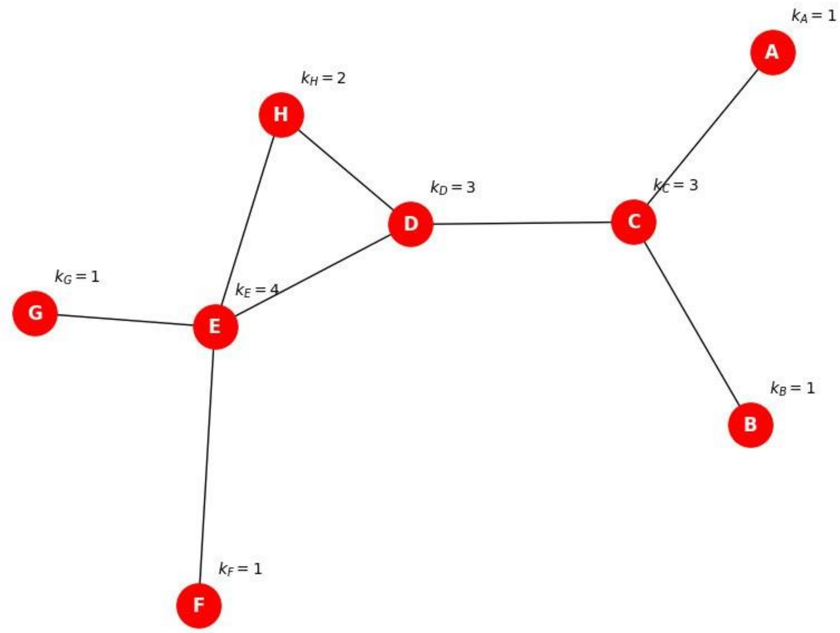
10. (dynnode2vec) Consider two snapshots of a dynamic graph with structural evolution from time *t=1* to *t=2*, as shown in the following figure. The evolving nodes in the timestamp *t* are defined as in Eq. 1 based on the Dynnode2vec method. (10pt)
    a) Calculate $V_{add}, E_{add}, V_{del}$, and E_{del} at timestamp *t=2*.
    b) Calculate ΔV_2.
    c) Describe key challenges in dynamic networks such as temporal evolution, node/edge addition or removal, and the need for temporal consistency in embeddings.



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.

11. (Div2Vec) (5pt)
    a) Given graph, use Div2Vec to perform a random walk of length 3 starting from node A.

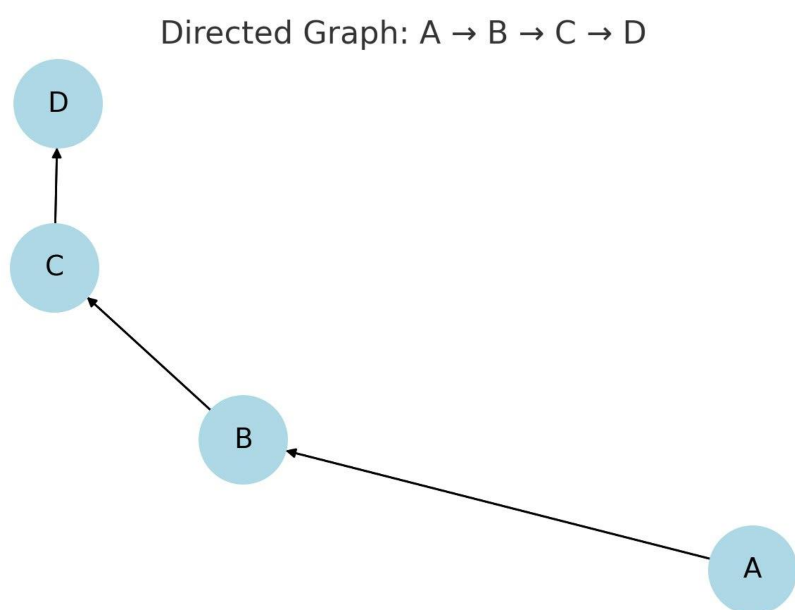
b) How Div2Vec balances accuracy vs. diversity using its sampling function?



12. (APP) (5pt)

a) Given directed graph with 4 nodes, apply APP method to simulate two sample random walks starting at node A, with restart probability α=0.3, and apply path sharing optimization.

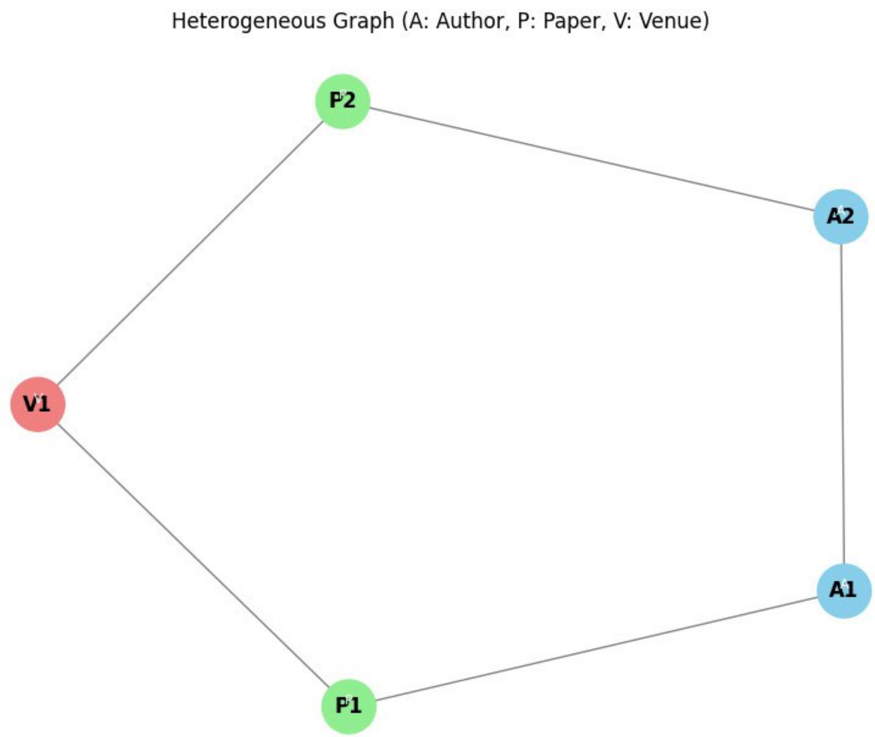
b) Compare how APP differs from Node2Vec in both walk structure and training objective.



13. (JUST) (5pt)

a) Simulate a 2-step walk starting from node A1 using the JUST algorithm with α=0.8, ℓ=1, with 0.5, 0.6 for walk 1, and walk 2, respectively.

b) How JUST's probabilistic transition strategy (jump/stay decision with decay functions) addresses the limitations of hand-crafted meta-paths.


Heterogeneous Graph (A: Author, P: Paper, V: Venue)
P2
A2
V1
A1
P1