Mid-term Exam (Graph Mining – Spring 2025)

Full Name:
Student ID:

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

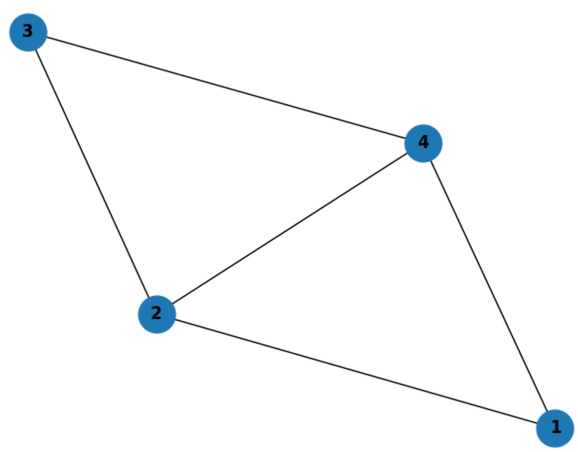
1. (Centrality Measures) Consider an undirected graph G of four nodes (15pt)
   a. Calculate betweenness centrality of node 4

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.

   b. Calculate closeness centrality of node 4

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

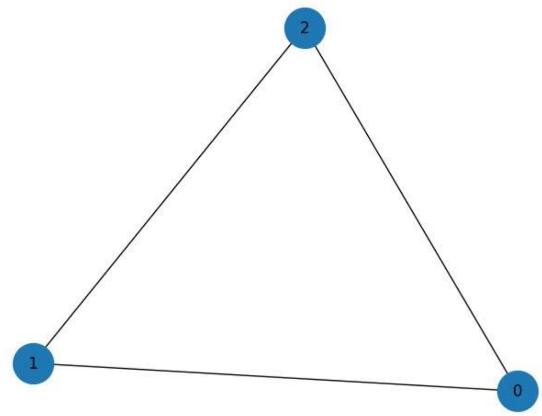


   c. The betweenness centrality and closeness centrality are based on the shortest paths between nodes. What are the roles of the shortest paths in the two centralities, respectively.

2. (Centrality Measures) Consider an undirected graph G of three nodes given in the following figure (15pt)
   a. Calculate Eigenvector, Katz centrality of node 1 with $\alpha = 1$, $\beta = 2$, t = 1

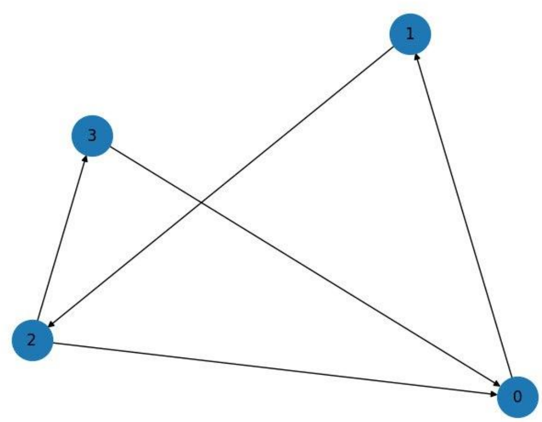
   Equation Eigenvector: x_i (t) = $\sum_{v_j \in N(v_i)} A_{ij} x_j(t-1)$, where A is adjacency matrix, t is time, with the centrality at time t = 0 being $x_j(0) = 1\ \forall j$

Equation Katz: $Katz(G) = \alpha \sum_j A_{ij} x_j + \beta$, where α is damping factor, β is bias constant.



b. Consider a directed graph G of four nodes given in the following figure, calculate PageRank centrality of all nodes, with β = 0.75, given pagerank of (0) = 1, pagerank of (2) = 2, pagerank of (3) = 3

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



c. In Katz centrality, what does the damping factor mean? What is the difference between Eigenvector centrality and Katz centrality are made by the damping factor?

d. In PageRank centrality, what is the purpose of the inverse of source node degree? Then, what is the difference between the static damping factor in Katz centrality and the inverse of degree in PageRank?

3. (Graph Visualization) Given undirected graph with 4 nodes A (0, 0), B (1, 0), C (1, 1), and D (0, 1) in the following figure. (10pt)

a. Calculate displacement vector for node A, where spring constant = 1, repulsive force constant = 2 and ideal edge length l = 2, log(1/2)= -0.3, $\log(\sqrt{2}/2) = -0.15$

Repulsive force

$$f_{rep}(u,v) = \frac{c_{rep}}{\|p_v - p_u\|^2}\overrightarrow{p_u p_v}$$

Attractive force

$$f_{spring}(u,v) = c_{spring}\log\left(\frac{\|p_v - p_u\|}{l}\right)\overrightarrow{p_u p_v}$$

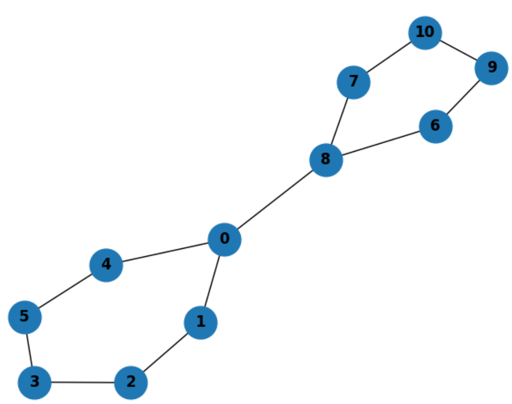
$$f_{attr}(u,v) = f_{spring}(u,v) - f_{rep}(u,v)$$

Resulting displacement vector

$$F_u = \sum_{v\in V} f_{rep}(v,u) + \sum_{uv\in V} f_{attr}(v,u)$$

b. The attractive force affects only adjacent nodes. Then, why the repulsive force affects both adjacent and non-adjacent nodes? Also, in this regard, please explain which graph visualization is a good visualization.

4. (Community Detection) Consider an undirected graph G of eleven nodes given in the following figure. There are two communities in the graph: A = {0, 1, 2, 3, 4, 5} and B = {6, 7, 8, 9, 10}. (10pt).



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

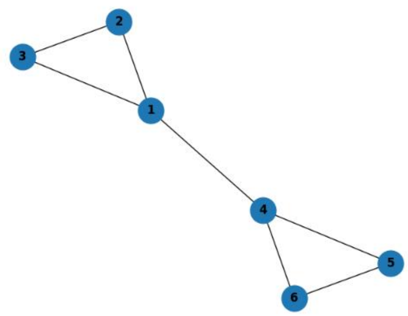
b. Calculate conductance of A and B using the equation (1).

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

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.

c. What is the difference between Min-cut and Normalized cut? Please explain in terms of how the two measures consider intra-compactness and inter-connectivity of communities.

5. (Community Detection) Consider an undirected graph G of six nodes given in the following figure. (10pt)



a. Calculate the global clustering coefficient C_i

Equation (1): $C_i = \frac{\#of\ closed\ triplets}{\#of\ connected\ triplets}$

b. Calculate the average clustering coefficient $\langle C \rangle$ in the graph G

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

c. What is the difference between global clustering and average clustering coefficients? Please explain in terms of how the two metrics measure density of edges within clusters or an entire graph.

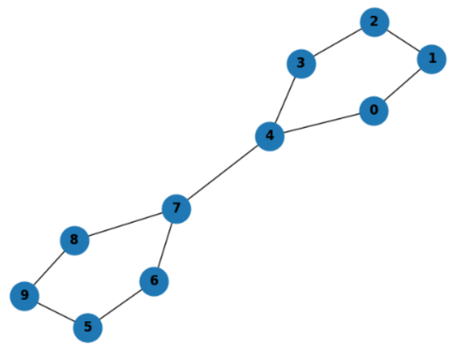
6. (Community Detection) Consider an undirected graph G of ten nodes given in the following figure with two communities: A = {0, 1, 2, 3, 4} and B = {5, 6, 7, 8, 9}. (5pt)

a. The community structure will be strong or weak if there are many pairs (i, j) with $A_{ij} > \frac{d_i d_j}{2m}$.

b. Apply the Equation (1) to calculate the modularity Q of the two communities.

c. Which edge is more or less valuable according to Modularity? What kinds of community structures are supposed by Modularity? Explain in terms of correlations of Modularity with scale-free graphs.

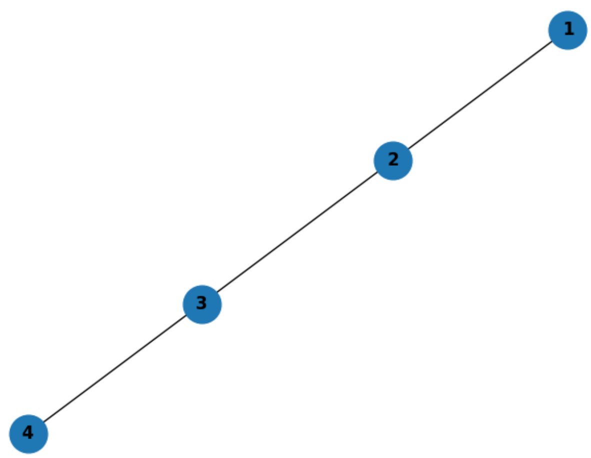
$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{d_i d_j}{2m}\right) \cdot \delta(v_i, v_j) \qquad (1)$$



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

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

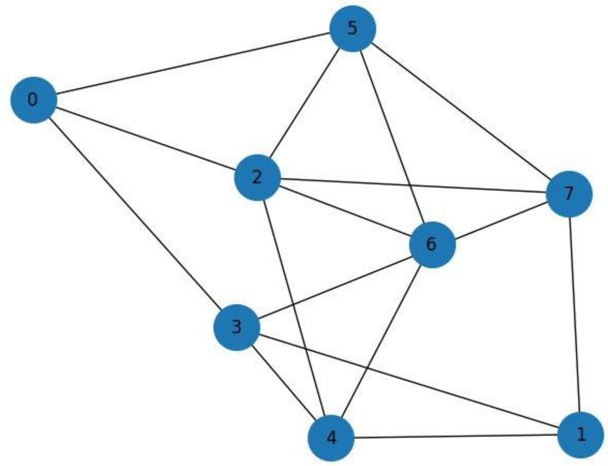
7. (Community Detection) Consider an undirected graph G of four nodes given in the following figure (5pt)
    a. Calculate k-means clustering with K =2 in one interaction and initial centroids are node 1 and node 4, with Euclidean distance
    b. What is the advantage of spectral clustering based on decomposition of adjacency matrices compared to applying k-means clustering directly to adjacency matrices as the question 7.a?



8. (Link Prediction) Consider an undirected graph G of eight nodes given in the following figure. (10pt)
    a. Calculate Jaccard's coefficient (JC), Adamic-Adar (AA) index of node 4 and node 6
    b. JC and AA are commonly modifications of the common neighborhoods by applying weights to the number of common neighborhoods. Please explain their differences in terms of their methods for applying the weights and their underlying rationales.

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

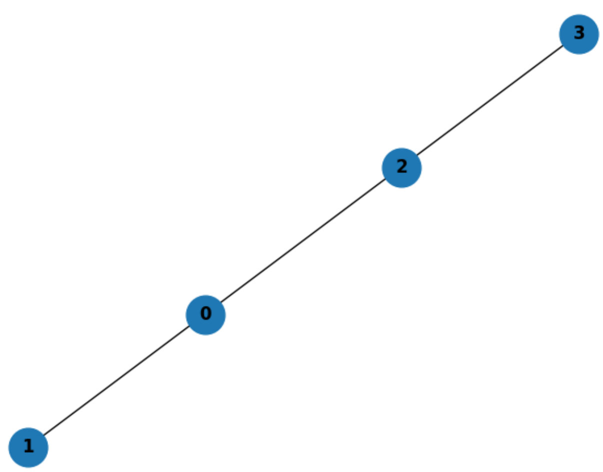
Equation AA: score (x, y) = $\sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$, with log(1) ≈ 0, log(2) ≈ 0.3, log(3) ≈ 0.4, log(4) ≈ 0.6, log(5) ≈ 0.7



9. (Link Prediction) Consider an undirected graph G of four nodes given in the following figure. (10pt)
   a. Calculate Katz Index with L = 2, β = 0.5

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

   b. Different from JC and AA, the Katz index is a global link prediction method. Why do we say it is global?



10. (Link Prediction) Consider an undirected graph G of three nodes given in the following figure. (10pt)

a. Calculate Hitting time of node 1 and node 2

b. Please explain the common points and differences between the Hitting time and PageRank centrality. Also, what are the correlations between information propagation between nodes, transition probabilities, and the Hitting time.

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

where H(k, y) = 1 + $\sum_m \mathrm{p}_{mj} H(m, y)$ when k≠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.