Problem 1. Frequent Graph Pattern Mining (10 points)

Solve Q2 of 11.5 EXERCISES in the textbook (page 297). Just draw all frequent subgraphs with at least one edge (i.e., \( \text{minsup} = 1 \)). Don’t need to show canonical codes; instead, show the frequency (the times that a frequent subgraph appears) of each subgraph.

Problem 2. Proof for Pagerank (15 points)

In the Pagerank algorithm, an initial probability distribution \( x_0 \in \mathbb{R}^n \) (note that we use a column vector to denote any probability distribution over \( n \) graph nodes) is iteratively updated by the formula \( x_{t+1}^\top = x_t^\top P \) \((t = 0, 1, 2, \cdots)\). \( P \in \mathbb{R}^{n \times n} \) \((P1 = 1)\) is the transition matrix.

1) Show that once \( 1^\top x_0 = 1 \) then \( 1^\top x_t = 1 \) for all subsequent \( t \geq 1 \).

2) If the underlying graph is well-behaved, show that \( 1^\top v_i = 0 \) holds for any left eigenvector \( v_i \) except the largest left eigenvector \( v_0 \) (i.e., the unique stationary distribution) of the the transition matrix \( P \).

(Hint: To prove 2), you should utilize the Perron Frobenius Theorem which tells that all left eigenvalues other than the largest left eigenvalue of \( P \) are strictly less than 1.)

Problem 3. Undirected vs. Directed Graphs (15 points)

We intend to construct a neighborhood graph over \( n \) data points \( X = \{x_1, x_2, \cdots, x_n\} \). Here we have two graph construction strategies: a) we link an edge between two data points \( x_i \) and \( x_j \) if and only if \( x_j \) is among \( k \) nearest neighbors of \( x_i \) according to Euclidean distance; b) we link an edge between two data points \( x_i \) and \( x_j \) if and only if \( x_j \) is among \( k \) nearest neighbors of \( x_i \) or \( x_i \) is among \( k \) nearest neighbors of \( x_j \). We denote the constructed graphs using the two strategies by \( G_a \) and \( G_b \), respectively. Please answer:

1) Of \( G_a \) and \( G_b \), which one is an undirected graph and which one is a directed graph?

2) If we know the adjacency matrix \( A \) of the graph \( G_a \), what is the adjacency matrix of the graph \( G_b \)?

Problem 4. Undirected Graph (25 points)

In class, we describe commute time and cosine similarity of Euclidean embeddings in undirected graphs. Now we want to compute these measures for a given undirected graph displayed in Fig. 1. Please answer:

1) What is a stationary probability distribution of random walks over the graph? Try the simplest method.

2) Besides the node \( v_5 \), what is the most relevant node to be recommended for the node \( v_0 \)? Try to use the commute time as the distance measure.

3) Besides the node \( v_5 \), what is the most relevant node to be recommended for the node \( v_0 \)? Try to use the cosine similarity of Euclidean embeddings as the similarity measure.
Figure 1. An undirected graph of six nodes, in which the weights of the respective edges are given.

Figure 2. A directed graph of ten nodes, in which the weight of each edge is 1.

(Hint: You should first write the graph adjacency matrix $A$, and then compute the graph Laplacian matrix $L = \text{diag}(A1) - A$. The pseudo inverse $L^+$ is computed from the eigendecomposition of $L$, that is, $L = V \text{diag}(0, \cdots, 0, \sigma_i, \cdots, \sigma_n) V^T$. Inverting the nonzero eigenvalues and keeping the zero eigenvalues of $L$ give rise to the pseudo inverse $L^+ = V \text{diag}(0, \cdots, 0, 1/\sigma_i, \cdots, 1/\sigma_n) V^T$. You are suggested to use the function $\text{eig}$ of Matlab to obtain eigendecomposition, and also handle the numerical issue by treating the eigenvalues of small absolute values, such as $|\sigma| \leq 10^{-6}$, as zero eigenvalues.)

Problem 5. Directed Graph (25 points)

We want to implement the Pagerank method we studied in class. Given a directed graph shown in Fig. 2, please answer:

1) What is the rank of all nodes with the importance being from high to low? Try the Pagerank method with 0.1 teleporting probability.

2) With respect to the node $v_2$, what is the rank of all other nodes with the relevance being from high to low? Try the Personalized Pagerank method with 0.5 teleporting probability.

3) If the node $v_1$ is from category $a$ and the node $v_5$ is from category $b$, what is the category ($a$ or $b$) that any
other node mostly belongs to? Try the Personalized Pagerank method with 0.5 teleporting probability respectively biased to nodes $v_1$ and $v_5$. 