HW4 - Data Mining - due by 10 am 01 April 2019

Homework Policy:

1. You can discuss your HW with at most one other group. However, discussions are restricted to oral only. Written similarity is regarded as plagiarism. If you group discusses with other groups, you must acknowledge the discussion in your written solution.
2. Homework submission: one person in the group submitting the homework is ok, but all the names must appear on the paper of the written homework.
3. The written homework must be submitted by the due date. Please turn in a hard copy in class. The hard copy will be graded.
4. This list may be updated further.

Problem 1 (10 points) Suppose that we have a database with 4 users and 4 items as follows:

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{pmatrix} 1 & & & 1 \\ & 1 & 1 & \\ 1 & 1 & & 1 \\ & 1 & & 1 \end{pmatrix} \end{array}$$

Value 1 is interpreted as the user likes the item. We will use the similarity based method to do recommendation. We represent each user as the set of items the user likes. For example A will be represented as $\{I_1, I_4\}$.

a (**5 points**) Construct the similarity matrix S where $S[i, j]$ is the Jaccard similarity between user i and user j.

b (**5 points**) For each user U and the each item I where $M[U, I]$ is unknown, we define the preference of U to I is:

$$Pref(U, I) = \sum_{j \in R(I)} S[U, j] \tag{1}$$

where $R(I)$ is the set of users that like item I. Compute preferences of A to I_2 and to I_3 and preferences of D to I_1 and to I_3.

Problem 2 (10 points) Given the utility matrix:

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{pmatrix} 2 & & 4 & 1 \\ & 1 & 3 & \\ 2 & 4 & & 3 \\ 1 & 5 & & \end{pmatrix} \end{array}$$

We will factorize the matrix M into the product of two matrices U, V of dimensions 4×2 and 2×4, respectively. Suppose that we start with the following matrices:

$$U_0 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 1 \\ 2 & 0 \end{bmatrix} \quad V_0 = \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \end{bmatrix} \tag{2}$$

(a) (**3 points**) What is the RMSE of the factorization by U_0, V_0

(b) (**4 points**) Find the first column of U, assuming other columns and V are given. That is, find x_1, x_3, x_3, x_4, so that the factorization

$$\begin{bmatrix} x_1 & 2 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \end{bmatrix} \tag{3}$$

has minimum RMSE.

(c) (**4 points**) Find the second row of V, assuming other rows and U are given. That is, find y_1, y_2, y_3, y_4, so that the factorization

$$\begin{bmatrix} x_1 & 2 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ y_1 & y_2 & y_3 & y_4 \end{bmatrix} \tag{4}$$

has minimum RMSE with x_1, x_2, x_3, x_4 from part (b).

Problem 3 (10 points) (Exercise 10.2.1 and 10.2.2 MMDS book) Given a graph in Figure 1. Using the Girvan-Newman algorithm:

(a) (3 points) Compute the contribution of shortest paths from node (A) to betweeness of each edge.

(b) (3 points) Compute the contribution of shortest path from node (B) to each edge.

(c) (4 points) Calculate the betweeness of every edge. Hints: make use of calculation in part (a) and (b) and the symmetry of the graph.

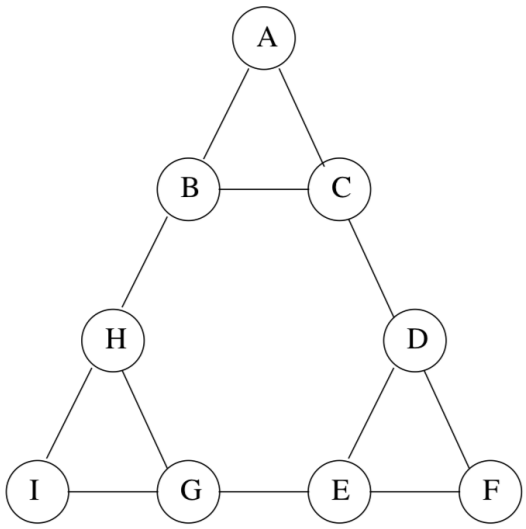

Figure 1: The graph for Problem 3.

Problem 4 (5 points) Suppose that each edge in the network of 5 nodes is generated with probability p. Find p so that the probability of observing the network in Figure 2 is maximum.

Problem 5 (5 points) How many triangles are there in the graph in Figure 3.

Problem 6 (10 points) : **To be updated**

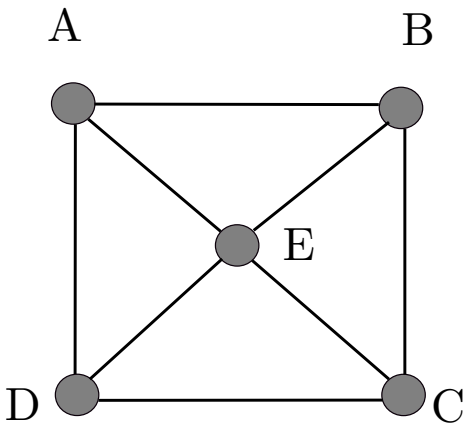

Figure 2: The graph for Problem 4.

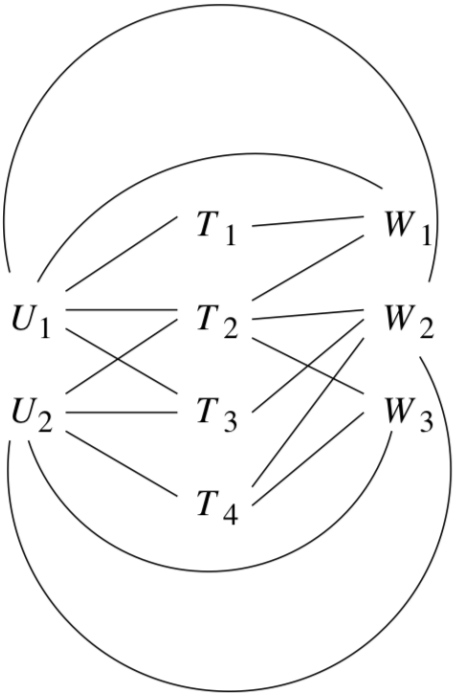

Figure 3: The graph for Problem 5.