CAP 6315 Social Networks and Big Data Analytics

Homework 3 Community Detection and Link Prediction

10 points, Due: March 27 (Sunday)

Question 1 [0.5 pt/each]: Please use your own language to briefly explain the following concepts, or answer the question:

Network community: A network community is a substructure of the network where nodes inside the community are more densely connected (i.e, more edges), compared to the connections between nodes inside the community and nodes outside the communities.

Link prediction: Link prediction aims to predict whether two nodes in a network may form a linkage or not. It can be used to predict whether two individuals in a social network are likely to establish a friendship connection or not.

Linear threshold influence model: Linear threshold influence model is a special type of computational model for infusion diffusion in social networks. It assumes that each node maintains a threshold values and the node may be influenced by its active friends with respect to certain weight values. An node become active and pass the influence to all its friends, if the accumulated influence values of the node is greater than its threshold.

Independent Cascade influence model: Linear threshold influence model is a special type of computational model for infusion diffusion in social networks. It assumes that each node has a probability value *p* to influence its neighbors, independently. Once a node become active, it will indecently try to influence each of its neighbors, with probability p of success.

Please use necessary sources (such as research papers and internet) to briefly explain how Facebook suggests friends for users (i.e., “Friends you might know” feature):

Question 2 [2.0 pts]:

The following table shows a toy rating matrix where each row denotes a user, and each column represents one item (it could be a book, a movie etc.). The number in each field (if not empty) indicates the user’s rating on that particular item, and an empty field means that the user has no rating on the given item.

a. Please use item-based collaborative filtering to calculate Angelica’s rating on item 3 (using Cosine distance). Please show your solution and the final matrix [1.0 pt]
b. Please use user-based collaborative filtering to calculate Bill’s rating on Item 4 (using Cosine distance). Please show your solution and the final matrix [1.0 pt]

Users	Item 1	Item 2	Item 3	Item 4	Item 5
Angelica	3.5	2		4.5	5
Bill	2	3.5	4		2
Chan	5	1	1	3	5
Dan	3	4	4.5		3

Solutions

Item based solutions:

$P(U_1, I_3) = \alpha[Sim(I_1, I_3)\times P(U_1, I_1) + Sim(I_2, I_3)\times P(U_1, I_2) + Sim(I_4, I_3)\times P(U_1, I_4) + Sim(I_5, I_3)\times P(U_1, I_5)]$

$Sim(I_1, I_3) = (2*4 + 5*1 + 3*4.5)/(sqrt(2^2+5^2+3^2)* sqrt(4^2+1^2+4.5^2)) = 26.5/(6.2 *6.1) = 0.70$

$Sim(I_2, I_3) = 33/(5.4*6.1) = 1.00$

$Sim(I_4, I_3) = 3/3 = 1.00$

$Sim(I_5, I_3) = 37.5/(5.4*7.07) = 0.70$

$P(U_1, I_3) = \frac{0.7\times3.5+1\times2+1\times4.5+0.70\times5}{0.7+1+1+0.7} = \frac{12.45}{3.4} = 3.66$

(2) Predict $P(U_2, I_4)$

$P(U_2, I_4) = \alpha[Sim(I_1, I_4)\times P(U_2, I_1) + Sim(I_2, I_4)\times P(U_2, I_2) + Sim(I_3, I_4)\times P(U_2, I_3) + Sim(I_5, I_4)\times P(U_2, I_5)]$

$Sim(I_1, I_4) = 30.75/(6.1 *5.41) = 0.93$

$Sim(I_2, I_4) = 33/(5.4*6.1) = 0.99$

$Sim(I_3, I_4) = 3/3 = 1.00$

$Sim(I_5, I_4) = 37.5/(5.4*7.07) = 0.98$

$$P(U_2, I_4) = \frac{0.93\times2+0.99\times3.5+1\times4+0.98\times2}{0.93+0.99+1+0.98} = \frac{11.285}{3.81} = 2.89$$

(3) Predict $P(U_4, I_4)$

$P(U_4, I_4) = \alpha[Sim(I_1, I_4)\times P(U_4, I_1) + Sim(I_2, I_4)\times P(U_4, I_2) + Sim(I_3, I_4)\times P(U_4, I_3) + Sim(I_5, I_4)\times P(U_4, I_5)]$

$Sim(I_1, I_4) = 30.75/(6.1 *5.41) = 0.93$

$Sim(I_2, I_4) = 33/(5.4*6.1) = 0.99$

$Sim(I_3, I_4) = 3/3 = 1.00$

$Sim(I_5, I_4) = 37.5/(5.4*7.07) = 0.98$

$$P(U_4, I_4) = \frac{0.93\times3+0.99\times4+1\times4.5+0.98\times3}{0.93+0.99+1+0.98} = \frac{14.19}{3.81} = 3.64$$

Users	Item 1	Item 2	Item 3	Item 4	Item 5
Angelica	3.5	2	**3.656**	4.5	5
Bill	2	3.5	4	**2.89**	2
Chan	5	1	1	3	5
Dan	3	4	4.5	**3.64**	3

User based:

Average (U_1) = 3.75

Average (U_2) = 2.875

Average (U_3) = 3

Average (U_4) = 3.625

$$p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i)$$

U_1 = a, j = 3

(1) $P_{1,3}$

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

Sim(U_1, U_2) = (3.5*2 + 2*3.5 + 5*2)/(sqrt(3.5^2+2^2+5^2)* sqrt(2^2+3.5^2+2^2)) = 0.83

Sim(U_1, U_3) = (3.5*5 + 2*1 + 4.5*3+ 5*5)/(sqrt(3.5^2+2^2+4.5^2+5^2)* sqrt(5^2+1^2+3^2+5^2)) = 48/(7.8*7.7) = 0.95

Sim(U_1, U_4) = (3.5*3 + 2*4 + 5*3)/(sqrt(3.5^2+2^2+5^2)* sqrt(3^2+4^2+3^2)) = 0.89

$P_{1,3}$ = 3.75+ (0.83*(4-2.875) + 0.95*(1-3) + 0.89*(4.5-3.625)) = 3.56

(2)$P_{2,4}$

Sim(U_2, U_1) = (3.5*2 + 2*3.5 + 5*2)/(sqrt(3.5^2+2^2+5^2)* sqrt(2^2+3.5^2+2^2)) = 0.83

Sim(U_2, U_3) = (2*5 + 3.5*1 + 4*1+ 2*5)/(sqrt(2^2+3.5^2+4^2+2^2)* sqrt(5^2+1^2+1^2+5^2)) = 48/(7.8*7.7) = 0.63

Sim(U_2, U_4) = (2*3 + 3.5*4 +4*4.5 + 2*3)/(sqrt(2^2+3.5^2+4^2+2^2)* sqrt(3^2+4^2+4.5^2+3^2)) = 0.99

$P_{2,4}$ = 2.875+ (0.83*(4.5-3.75)+0.63*(3-3)) = 3.50

(3)$P_{4,4}$

Sim(U_4, U_1) = (3.5*3 + 2*4 + 5*3)/(sqrt(3.5^2+2^2+5^2)* sqrt(3^2+4^2+3^2)) = 0.89

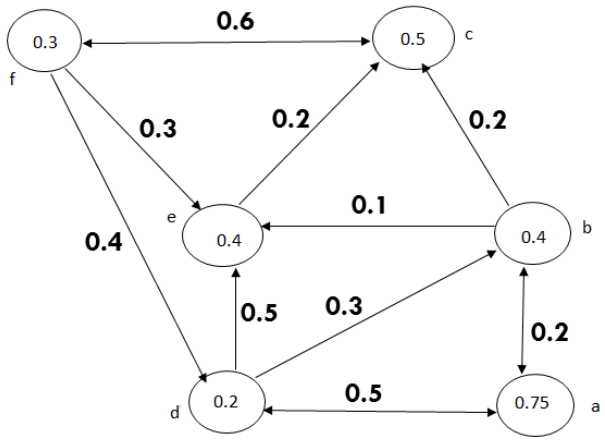
Sim(U_4, U_3) = (5*3 + 1*4 +1*4.5 + 5*3)/(sqrt(5^2+1^2+1^2+5^2)* sqrt(3^2+4^2+4.5^2+3^2)) = 0.72

$P_{4,4}$ = 3.625+ (0.89*(4.5-3.75)+0.72*(3-3)) = 4.29

Users	Item 1	Item 2	Item 3	Item 4	Item 5
Angelica	3.5	2	**3.557**	4.5	5
Bill	2	3.5	4	**3.498**	2
Chan	5	1	1	3	5
Dan	3	4	4.5	**4.296**	3

Question 3 [2.0 pts] In the following directed social network, each value on the path indicates the influence value from a node u to node v, and the value inside each node denotes the threshold of the node.

1. Assume node "a" is initially selected to be activated, please use Linear Threshold influence model to show (1) the order of the nodes which will be influenced [0.5 pt], and the order of the nodes which will be activated [0.5 pt].
2. Please determine which node has the least influential power, and explain why [1 pt]



Solutions:

A -> B and D.

D -> E and B.

E -> C

B -> C

++

the order of the nodes which will be activated

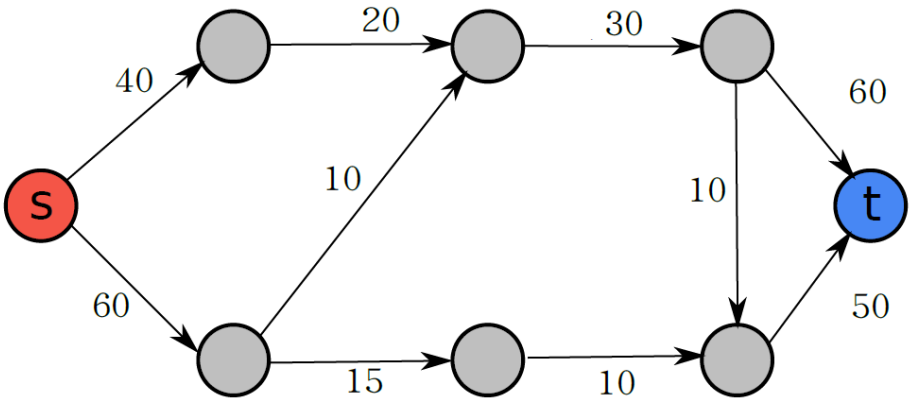
A -> D -> E and B.

++

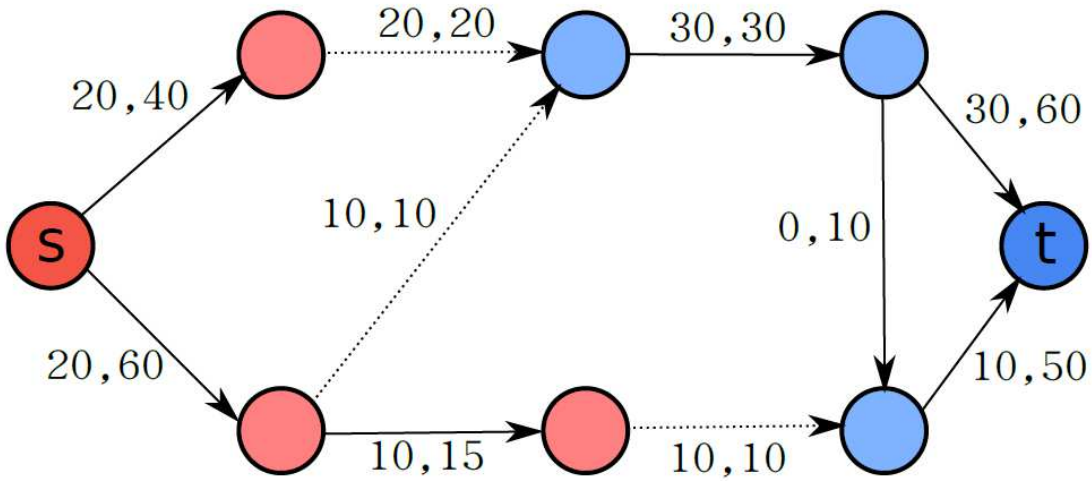
Please determine which node has the least influential power, and explain why

Node E has least influence, because once E is activated, it only influence node C, then the influence propagation will stop. So it will only influence two nodes (including node E itself). For any other node, it will influence at least three nodes.

Question 4 [2.0 pts]: In the following network, assume "S" and "t" denote source and sink node, respectively, and the value on each edge denotes the weight/capacity of each edge. Please use Ford Fulkerson algorithm to find the min-cut which separate the network into two community (one includes "s" and the other includes "t"). Please show your solution.



Solution



Question 5 [2 pts]: Programming Task

Zachary's karate club (http://networkdata.ics.uci.edu/data.php?id=105) dataset is an undirected social network recording friendships between 34 members of a karate club at a US university in the 1970s. Some background information about this network can be found from the following URL: https://sites.google.com/site/ucinetsoftware/datasets/zacharykarateclub

Please download the social network and implement influence modeling as follows:

a. Assume the influence value for all edges from member u to member v, i.e., (u,v), are equal to α, and the probability value for each member to influence his/her neighbor is *p*. We are now trying to select one representative from the network to broadcast the message, please use Independent Cascade model to determine which member should be selected such that the number of numbers can be reached from the selected representative is maximized (please show your solution and report the results with respect to different parameter settings including *p*={0.05, 0.2, 0.4}). Because Independent Cascade model uses a random process, please repeat the experiments for 10 times and calculate the average results for each node. (2 pts) [1 pt for program and 1 point for the report].