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 set of nodes in a network which have interactions more frequently with each other than they do with the rest of the network. The nodes form groups and can be represented on a graph through adjacent edges. Members of a community are likely to have the same features, making predictions possible.
Link prediction: using community detection tools, the likelihood that two nodes will form a connection is determined. This is used when given an incomplete set of social links between a complete set of social network members. By observing the set of links at one point in time, the set of links at a future point in time can be predicted.
Linear threshold influence model: the activity of a node is influenced by its neighbors, and a node will become active once a threshold number (or weighted value) of its neighbors are active. This is one example of a herd behavior: once enough neighbors of a group member take on a trait, that group member is also likely to take on the trait.
Independent Cascade influence model: when a single node becomes active, it has a one-time chance to also activate any of its neighbors, who then, in turn, may activate their neighbors - a cascade. This is possible because each node has some influence on its neighbors, and has a probability of activating any neighbor when it itself activates. If enough nodes nodes activate at each transition, the network can change states very quickly.
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): the basic method is that Facebook groups users by traits – such as family members, schools, workplaces, geographic areas, interests – and then looks for likely connections in those groups. Other sources of potential connections are contact lists on mobile devices, applications, and services linked to Facebook. Interacting with another person's page or searching for their name is also a way to create a friend suggestion. More controversial methods Facebook may use to discover connections are cookies or trackers embedded into websites which identify users without their knowledge or consent, and report their online activity back to Facebook. One such example would be canvas fingerprinting, where the HTML5 canvas is used to create tracking data.
Sources:

"Canvas Fingerprinting." Wikipedia. Wikimedia Foundation, n.d. Web. 16 Mar. 2016. <https://en.wikipedia.org/wiki/Canvas_fingerprinting>.
"Facebook "Suggested Friends" Is Creepier than I Ever Could Have Imagined. • /r/OkCupid." Reddit. N.p., 11 July 2015. Web. 16 Mar. 2016. <https://www.reddit.com/r/OkCupid/comments/3cy24i/facebook_suggested_friends_is_creepier_than_i/>.
Leona, Cate. "How Does Facebook Suggest Friends?" Udemy Blog. N.p., n.d. Web. 16 Mar. 2016. <https://blog.udemy.com/how-does-facebook-suggest-friends/>.
Solanki, Kartik. "How Does Facebook Suggest Friends?" Quora. N.p., 14 Apr. 2015. Web. 16 Mar. 2016. <https://www.quora.com/How-does-Facebook-suggest-friends>.

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	3.7	4.5	5
Bill	2	3.5	4	3.5	2
Chan	5	1	1	3	5
Dan	3	4	4.5		3

a. Solution in the image below, with final result added to the matrix above in (Angelica, Item 3).

$$Sim(I_1, I_3) = cosine(I_1, I_3) = cosine((2,5,3),(4,1,4.5))$$
$$= \frac{2\cdot 4 + 5\cdot 1 + 3\cdot 4.5}{\sqrt{2^2+5^2+3^2}\cdot\sqrt{4^2+1^2+4.5^2}} = \frac{26.5}{\sqrt{38}\cdot\sqrt{37.25}} = 0.7044$$

$$Sim(I_2, I_3) = cosine(I_2, I_3) = cosine((3.5,1,4),(4,1,4.5))$$
$$= \frac{3.5\cdot 4 + 1\cdot 1 + 4\cdot 4.5}{\sqrt{3.5^2+1^2+4^2}\cdot\sqrt{4^2+1^2+4.5^2}} = \frac{33}{\sqrt{29.25}\cdot\sqrt{37.25}} = 0.9997$$

$$Sim(I_4, I_3) = cosine(I_4, I_3) = cosine((3),(1))$$
$$= \frac{3\cdot 1}{\sqrt{3^2}\cdot\sqrt{1^2}} = \frac{3}{3} = 1$$

$$Sim(I_5, I_3) = cosine(I_5, I_3) = cosine((2,5,3),(4,1,4.5))$$
$$= \frac{2\cdot 4 + 5\cdot 1 + 3\cdot 4.5}{\sqrt{2^2+5^2+1^2}\cdot\sqrt{4^2+1^2+4.5^2}} = \frac{26.5}{\sqrt{38}\cdot\sqrt{37.25}} = 0.7044$$

$$r(U_1, I_3) = \left(\frac{1}{0.7044+0.9997+1+0.7044}\right)(3.5\cdot 0.7044 + 2\cdot 0.9997 + 4.5\cdot 1 + 5\cdot 0.7044)$$
$$= \left(\frac{1}{3.4085}\right)(12.4868) = 3.6634 \approx 3.7$$

b. Solution in the image below, with final result added to the matrix above in (Bill, Item 4).

$\bar{V}_2 = \frac{1}{4}(2+3.5+4+2) = 2.875$

$P_{2,4} = 2.875 + K \sum_{i=\{1,3,4\}} \omega(2,i)(V_{i,4} - \bar{V}_i)$ let $K=1 \Rightarrow$

$P_{2,4} = 2.875 + \sum_{i=\{1,3,4\}} sim(2,i)(V_{i,4} - \bar{V}_i)$

$sim(U_2, U_1) = cosine(U_2, U_1) = cosine((2, 3.5, 2), (3.5, 2, 5))$

$= \frac{2 \cdot 3.5 + 3.5 \cdot 2 + 2 \cdot 5}{\sqrt{2^2+3.5^2+2^2} \cdot \sqrt{3.5^2+2^2+5^2}} = \frac{24}{\sqrt{20.25} \cdot \sqrt{41.25}} = 0.8304$

$sim(U_2, U_3) = cosine(U_2, U_1) = cosine((2, 3.5, 4, 2), (5, 1, 1, 5))$

$= \frac{2 \cdot 5 + 3.5 \cdot 1 + 4 \cdot 1 + 2 \cdot 5}{\sqrt{2^2+3.5^2+4^2+2^2} \cdot \sqrt{5^2+1^2+1^2+5^2}} = \frac{27.5}{\sqrt{36.25} \cdot \sqrt{52}} = 0.6334$

$sim(U_2, U_4) = cosine(U_2, U_4) = cosine((2, 3.5, 4, 2), (3, 4, 4.5, 3))$

$= \frac{2 \cdot 3 + 3.5 \cdot 4 + 4 \cdot 4.5 + 2 \cdot 3}{\sqrt{2^2+3.5^2+4^2+2^2} \cdot \sqrt{3^2+4^2+4.5^2+3^2}} = \frac{44}{\sqrt{36.25} \cdot \sqrt{54.25}} = 0.9922$

$\bar{V}_1 = \frac{1}{4}(3.5+2+4.5+5) = 3.75$ $\bar{V}_3 = \frac{1}{5}(5+1+1+3+5) = 3$

$\bar{V}_4 = \frac{1}{4}(3+4+4.5+3) = 3.625$

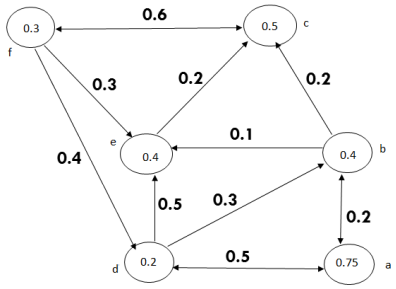
$P(U_2, I_4) = 2.875 + 0.8304(4.5-3.75) + 0.6334(3-3) +$

$0.9922(\underbrace{\text{no rating}}_{0} - 3.625)$

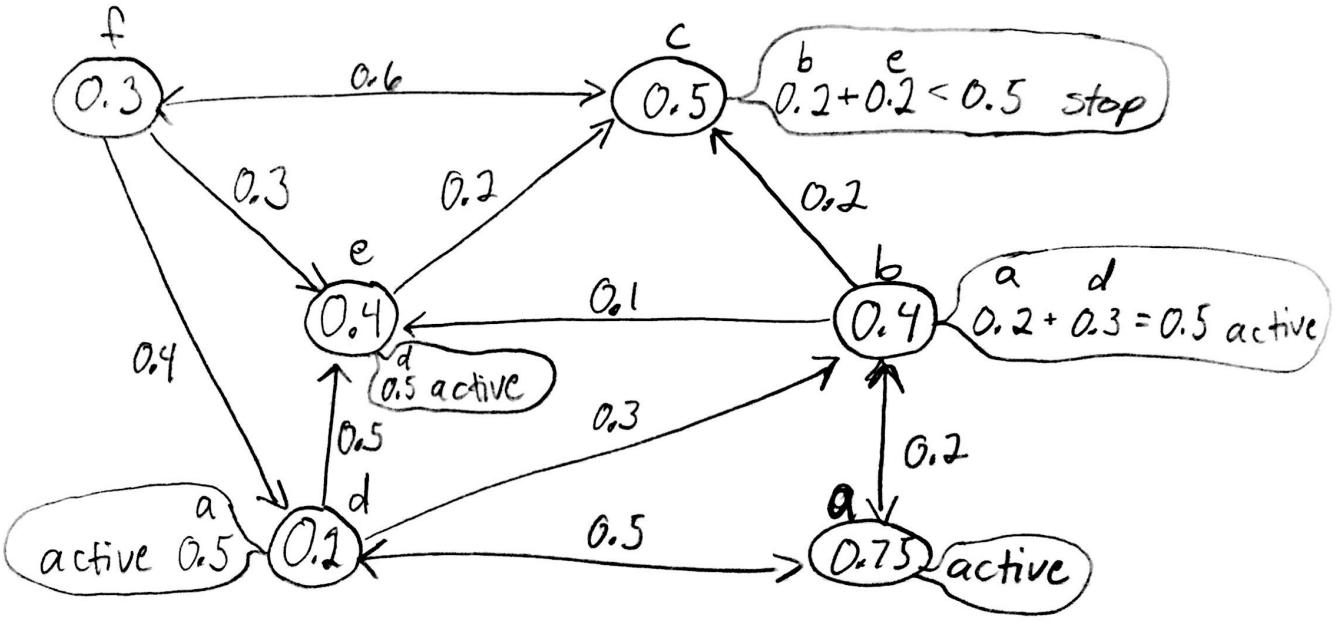
$= 2.875 + 0.6228 + 0 = 3.4978 \approx 3.5$

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].



1. Solution in the image below.

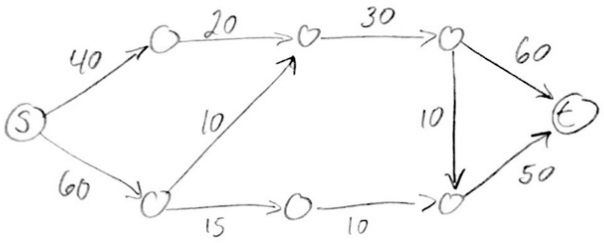


influence order: a → b, d → e, c
activation order: a → d → e, b

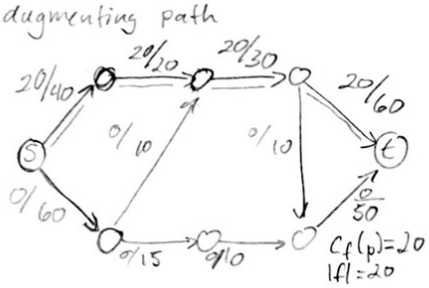
2. The least influential node is e. Nodes a, c, d, and f will always activate at least one other node on activation. Node b does not activate any other nodes by itself, but it does influence 3 other nodes, compared to e influencing only 1. Additionally, let P denote the influential power a node has on the network, and define $P_{node} = \Sigma$ (outgoing edge weights). From this definition, $P_a = (0.2 + 0.5) = 0.7$, $P_b = (0.2 + 0.2 + 0.1) = 0.5$, $P_c = 0.6$, $P_d = (0.5 + 0.3 + 0.5) = 1.3$, $Pe = 0.2$, $P_f = (0.4 + 0.3 + 0.6) = 1.3$. Node e has the lowest network influential power, $P_e = 0.2$.

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.

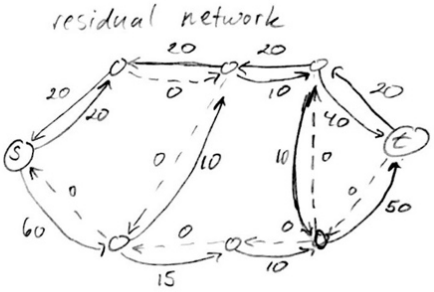
Solution in the image below.



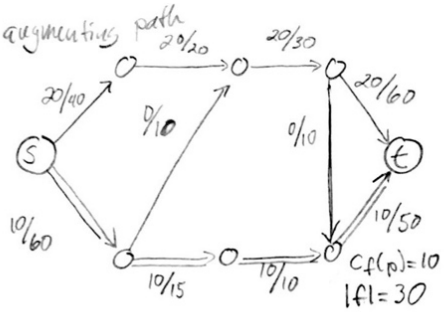
augmenting path



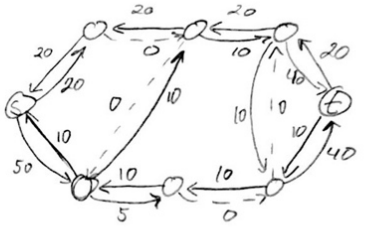
residual network



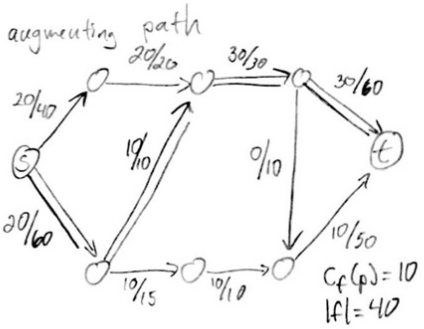
augmenting path



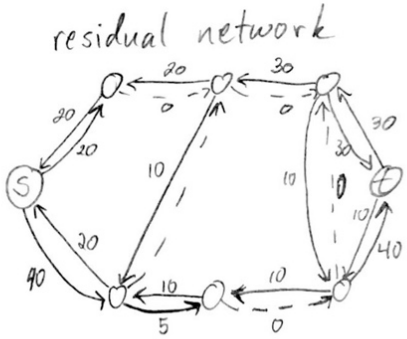
residual network



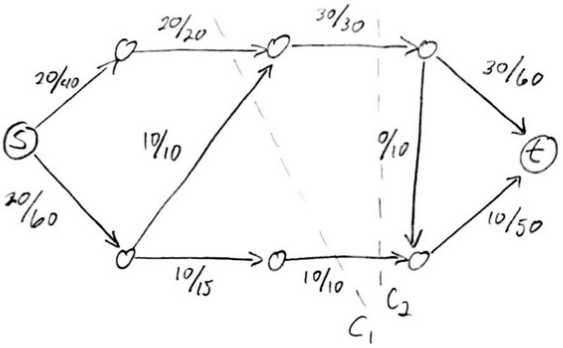
augmenting path



residual network



no more augmenting paths



2 min cuts – C_1 & C_2

C_1 : $f(S,T) = 20+10+10 = 40$

$c(S,T) = 20+10+10 = 40$

} $f(S,T) = c(S,T)$

C_2 : $f(S,T) = 30+10 = 40$

$c(S,T) = 30+10 = 40$

} $f(S,T) = c(S,T)$

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].

Report:

I implemented a program to read the karate graph .csv exported from Gephi and do an independent cascade model process on it. The graph is turned into a matrix and from each node, an independent cascade is run. The cascade is run 10 times from each node, and the average number of activated nodes after the cascade completes is recorded. This process is done 3 times with the parameter settings described in the problem. The end result is a text file with the adjacency matrix and the parameter value followed by the results for each of the 3 values. The total document is quite long, so it attached to the submission on blackboard, but will be summarized here. When programming this, I was unable to include the karate_edge.csv in a runnable .jar file, so the whole project is included in a single .zip file. The project runs in Eclipse, however.

Summary of the output:

With the influence parameter of 0.05, all nodes activated between 1 and 2.2 nodes after the cascade, with nodes 3 and 33 having the most activations.

With the influence parameter of 0.2, all nodes activated between 1.2 and 11.1 nodes after the cascade, with node 33 having the most activations.

With the influence parameter of 0.4, all nodes activated between 6.1 and 25.3 nodes after the cascade, with node 1 having the most activations.

These results do not point to a consistently most influential member, although 33 is a likely choice.