**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 may 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 links in those groups. Other sources of these connections are phone numbers collected from contacts lists on mobile devices and applications 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. 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]:**

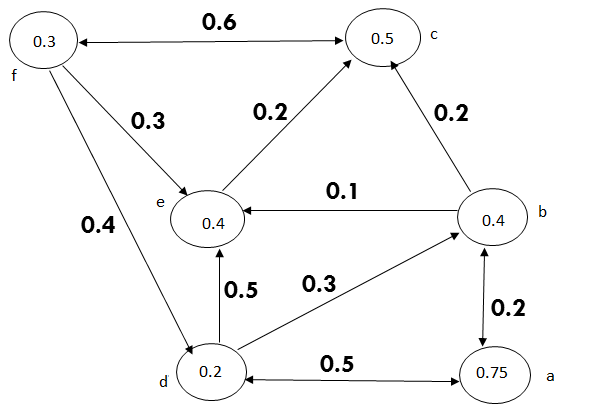
The following table shows a toy rating matrix where each row denotes a user, and each column represents one item (it could be a book, a movie etc.). The number in each field (if not empty) indicates the user’s rating on that particular item, and an empty field means that the user has no rating on the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

* 1. 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]
  2. 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]

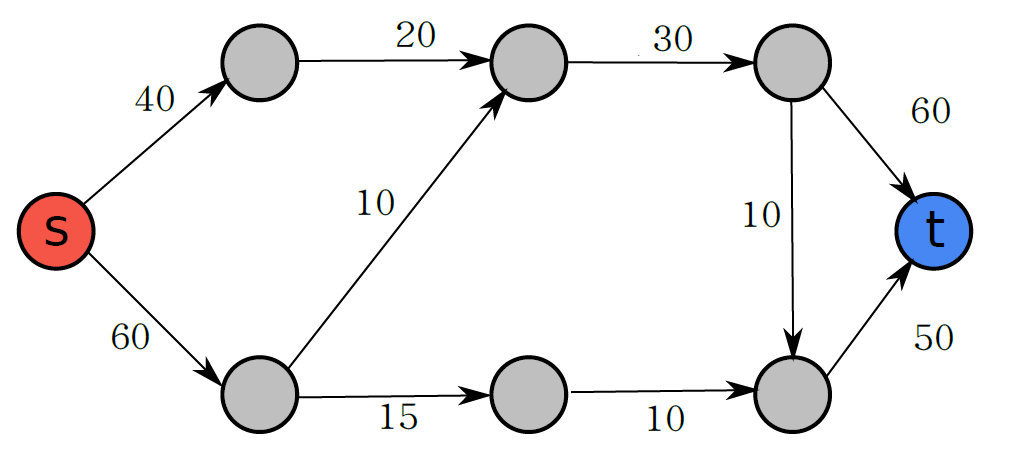
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Users | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
| Angelica | 3.5 | 2 |  | 4.5 | 5 |
| Bill | 2 | 3.5 | 4 |  | 2 |
| Chan | 5 | 1 | 1 | 3 | 5 |
| Dan | 3 | 4 | 4.5 |  | 3 |

**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]

****

**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.

****

**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:

* 1. 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].