

Name	
ASU ID Number	

CSE 472: Social Media Mining

Homework II - Network Models and Data Mining

Prof. Huan Liu
Due at 2021 Sept 28, 11:59 PM

This is an *individual* homework assignment. Please submit a digital copy of this homework to **Grade-scope**. For your solutions, even when not explicitly asked you are supposed to concisely justify your answers.

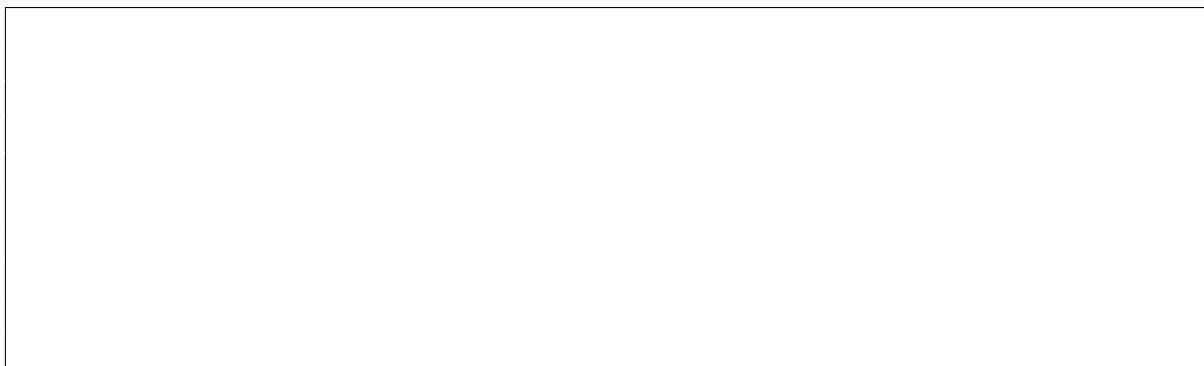
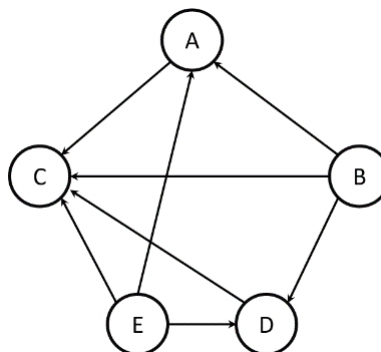
1. [Network Models]

- a. Assuming that we are interested in a sparse random graph, what should we choose as the value of p ? How does the value of p affect the sparseness? Where p defines the probability of forming edges.

- b. We can make a simple random graph model of a network with clustering or transitivity as follows. We take n vertices and go through each distinct trio of three vertices, of which there are $\binom{n}{3}$, and with independent probability p we connect the members of the trio together using three edges to form a triangle, where $p = \frac{c}{\binom{n-1}{2}}$ with c constant. Show that the mean degree of a vertex in this network is $2c$.

- c. In a friendship network, each node represents a user, and an edge exists between two nodes if one user befriends another. When a new user arrives (new node added to the graph), that user befriends other users with probability proportional to their indegree, hence, following the preferential attachment model. Suppose a new user is going to join the following friendship network, what would be the probability of this user getting connected to the previous ones based on the preferential attachment model? where n is the total number of nodes. A node attaches to another node v_i with probability give in equation (1), where $v_i \in \{A, B, C, D, E\}$

$$P(v_i) = d_i^{in} / \sum_{j=1}^n d_j^{in} \quad (1)$$



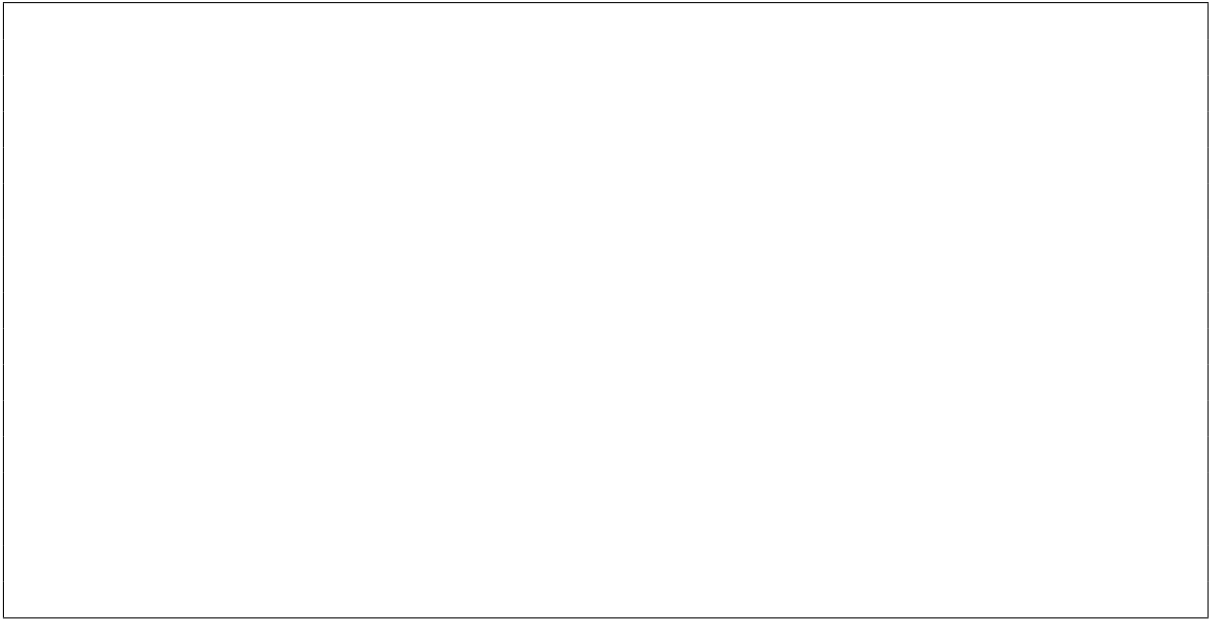
2. **[Data Mining]** Consider the given dataset from an employee database. For a given entry row, *count* column represents the number of data tuples having the values for *department*, *salary*, and *status* given in that row. For example, there are 15 instances with values of (department = sales, salary = high, status=senior). Let *status* be the class label attribute, answer the following questions.

Department	Salary	Status	Count
Sales	High	Senior	15
Sales	Low	Junior	20
Systems	Medium	Junior	20
Marketing	High	Junior	10
Marketing	High	Senior	5

- a. What is the value for the $H(Status)$? Where $H(x)$ defines the entropy of x ?

- b. Based on the Information Gain values, which feature is the most probable to be the root node of the decision tree? Show all your work.

- c. Draw the final decision tree. An example of how to draw the tree on the text box:



- d. Given a data instance having the values “*Sales*” and “*High*” for the attributes *department* and *salary*, respectively, what would a Naive Bayesian classification of the class attribute *Status* for the instance be? Detail all your calculations.

