Course: CSC591/791, Graph Data Mining: Theory, Algorithms, and Applications
Final Exam: Comprehensive, home-take exam.
STUDENT's ID: ____________________

This exam is intended to be an **individual** effort. You are allowed to use and reference the materials posted on Moodle, cited in the exam, your homework assignments, your notes, the class slides, and the lecture assignments. You are **NOT** allowed to use the other sources, such as your classmates and the internet at large.

1. Good luck!
2. You get BONUS 5 points if your solution is in LaTeX (both source and PDF)
3. Otherwise, a PDF (even a scanned hand-written solution) must be uploaded into Moodle
4. The Total Exam Score by the TA: ______ out of **(100 points)**

From the NCSU Code of Student Conduct:

DEFINITIONS OF ACADEMIC DISHONESTY

1. Academic dishonesty is the giving, taking, or presenting of information or material by a student that unethically or fraudulently aids oneself or another on any work which is to be considered in the determination of a grade or the completion of academic requirements or the enhancement of that student's record or academic career.
2. A student shall be guilty of a violation of academic integrity if he or she:
   - represents the work of others as his or her own;
   - obtains assistance in any academic work from another individual in a situation in which the student is expected to perform independently;
   - gives assistance to another individual in a situation in which that individual is expected to perform independently;
   - offers false data in support of laboratory or field work.

3. The act of submitting work for evaluation or to meet a requirement is regarded as assurance that the work is the result of the student's own thought and study, produced without assistance, and stated in that student's own words, except as quotation marks, references, or footnotes acknowledge the use of other sources. Submission of work used previously must first be approved by the instructor.

4. Regulations regarding academic dishonesty are set forth in writing in order to give students general notice of prohibited conduct. They should be read broadly and are not designed to define academic dishonesty in exhaustive terms.

5. If a student is in doubt regarding any matter relating to the standards of academic integrity in a given course or on a given assignment, that student shall consult with the faculty member responsible for the course before presenting the work.

By signing this exam, the student acknowledges the above terms and agrees to abide by NCSU policies on academic integrity.

Agreed ________________ (specify your ID) Date ____________

1. ______ out of **(15 points)**: Given an infectious disease following the SIS virus propagation model with transmission probability $\beta = 0.3$ and healing probability $\delta = 0.5$, in a contact network with 10 nodes and the following edges: {(0, 1), (0, 3), (1, 4), (2, 3), (2, 6), (5, 6), (5, 7), (6, 7), (7, 9), (8, 9)}, answer:

   (a) ______ out of **(1 point)**: What is the spectral radius of the network?

   (b) ______ out of **(1 point)**: What is the effective strength of the virus?

   (c) ______ out of **(1 point)**: According to the theorem studied in class [1], can the infection result in an epidemic?

   (d) ______ out of **(9 points)**: Using the NetShield algorithm [2], which 3 nodes should be immunized to minimize the spread of the infection?

   (e) ______ out of **(3 points)**: After immunizing the 3 nodes selected in (4), can the infection result in an epidemic?

   References:

   1 B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, C. Faloutsos. Got the Flu (or Mumps)? Check the Eigenvalue! arXiv:1004.0060 [physics.soc-ph], 2010.

   2 H. Tong, B. A. Prakash, C. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, D. H. Chau. On the Vulnerability of Large Graphs. In ICDM, 2010.

   Answers:
   (a) The spectral radius (i.e., the largest eigenvalue of the adjacency matrix) of the network is $\lambda_1 = 2.3728$.
   (b) The effective strength of the virus is $s = \lambda_1 * \frac{\beta}{\delta} = 2.2143 * \frac{0.3}{0.5} = 1.4237$.
   (c) $s > 1$. Therefore, the infection can result in an epidemic.
   (d) The 3 nodes selected for immunization by the NetShield algorithm are $\{6, 7, 3\}$.
   (e) The effective strength of the virus after immunizing the nodes is $s' = \lambda_1' * \frac{\beta}{\delta} = 1.4142 * \frac{0.3}{0.5} = 0.8485$. Since $s' < 1$ after immunizing the nodes, the infection cannot result in an epidemic.

2. ______ out of **(15 points)**: Given random variables x, y, w, z, **prove or disprove** the following statements about their conditional independence ($\perp$) relationships (see the lecture on D-separation and example below; if you decide to disprove the stmt, then a counter-example of a DAG will suffice):

   (a) ______ out of **(0 points)**: Example: $(x \perp y, w | z) \Longrightarrow (x \perp y | z)$
   Proof:

$$p(x,y|z) = \sum_w p(x,y,w|z) = \sum_w p(x|z)p(y,w|z) =$$
$$p(x|z) \sum_w p(y,w|z) = p(x|z)p(y|z)$$

(b) ______ out of **(6 points)**: $(x \perp y, z) \& (x, y \perp w | z) \Longrightarrow (x \perp w | z)$
Proof:

$$p(x, w|z) = \sum_y p(x, y, w|z) = \sum_y p(x, y|z) p(w|z) =$$
$$p(x|z) p(w|z) \sum_y p(y|z) = p(x|z) p(w|z)$$

(c) ______ out of **(9 points)**: $(x \perp y, z) \& (x \perp y | w) \Longrightarrow (x \perp y | z, w)$
Proof by counter-example: Consider the DAG, $x \rightarrow z \leftarrow u \rightarrow w \leftarrow y$. Using D-deparation,

$$(x \perp y | z)$$
$$(x \perp y | w)$$
$$\neg (x \perp y | z, w)$$

since the path between x and y is activated when both z and w are observed.

3. ______ out of (**15 points**): Suppose you need to design an efficient algorithm for analysizing random walks of an undirected simple graph G. Let A be its adjacency matrix with Singular Valude Decomposition (SVD) $A = U \Lambda U^t$.

(a) ______ out of **(5 points)**: Prove that the power matrix $A^k = U \Lambda^k U^t$.
Proof: To compute A^k, we observe that

$$A^k = \underbrace{(U \Lambda U^t)(U \Lambda U^t) ... (U \Lambda U^t)}_{k \text{times}}$$
$$U^t U = I$$
$$A^k = U \Lambda^k U^t$$

(b) ______ out of **(10 points)**: Devise an efficient algorithm to compute A^k for different values of k. What is its Big-O complexity?
Answer: An efficient algorithm to compute A^k:

i. Do SVD for A: $A = U \Lambda U^T$.
SVD has time complexity $O(min\{m^2 \, n, m \, n^2\})$.
ii. Calculate Λ^k: for each diagonal element λ_i in Λ, calculate λ_i^k.
This step takes $O(n \, log_2(k))$ by using exponentiating by squaring: $\lambda_i^k = \begin{cases} \lambda_i * (\lambda_i^2)^{\frac{k-1}{2}}, & \text{if } k \text{ is odd} \\ (\lambda_i^2)^{\frac{k}{2}}, & \text{if } k \text{ is even.} \end{cases}$
iii. Do the matrix multiplication $A^k = U \Lambda^k U^T$.
This last step takes $O(n^3)$ with a naive approach, but can be further reduced using other matrix multiplication algorithms like Strassen in $O(n^{\log_2 7}) \approx O(n^{2.807})$.

Thus, the Big-O complexity is $O(min\{m^2 \, n, m \, n^2\} + n \, log_2(k) + n^{2.807})$.

4. ______ out of (**10 points**): Let G be an undirected complete bi-partitie graph. Prove that if λ is an eigenvalue of its adjacency matrix then $-\lambda$ is also its eigenvalue. Hint: Assume that after the permutation of rows and columns, the adjacency matrux is of the form:

$$A = \begin{pmatrix} 0 & B \\ B^t & 0 \end{pmatrix}$$

Proof: If G is a bi-partite graph, then we can permute the rows and columns of its adjacency matrix A to obtain its block-diagonal form:

$$A = \begin{pmatrix} 0 & B \\ B^t & 0 \end{pmatrix}$$

Suppose that u is an eigenvector u of A with eigenvalue λ:

$$u = \begin{pmatrix} x \\ y \end{pmatrix}$$

Then

$$\begin{pmatrix} 0 & B \\ B^t & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

This implies that $B^t x = \lambda y$ and $By = \lambda x$. This implies that

$$\begin{pmatrix} 0 & B \\ B^t & 0 \end{pmatrix} \begin{pmatrix} x \\ -y \end{pmatrix} = \begin{pmatrix} -By \\ B^t x \end{pmatrix} = -\lambda \begin{pmatrix} x \\ -y \end{pmatrix}$$

Therefore, $\begin{pmatrix} x \\ -y \end{pmatrix}$ is an eigenvector of A with eigenvalue $-\lambda$.

5. ______ out of (**20 points**): Consider the following training examples:

Instance	x_1	x_2	x_3	class, y
1	T	T	T	+
2	T	T	T	+
3	F	F	T	+
4	F	T	F	-
5	T	F	T	-
6	T	F	F	-
7	F	F	F	-
8	F	T	F	-

(a) ______ out of **(10 points)**: You will be using a naïve Bayes classifier for this question. Given a test example with attributes $x_1 = T$, $x_2 = T$, and $x_3 = F$, which class will be assigned to this test example? Show your work clearly.

Answer:

$$p_+ = p(y=+|x_1=T, x_2=T, x_3=F) = \frac{p(x_1=T, x_2=T, x_3=F, y=+)}{p(x_1=T, x_2=T, x_3=F)}$$
$$= \frac{p(x_1=T, x_2=T, x_3=F|y=+)\,p(y=+)}{p(x_1=T, x_2=T, x_3=F)}$$

$$p_- = p(y=-|x_1=T, x_2=T, x_3=F) = \frac{p(x_1=T, x_2=T, x_3=F|y=-)\,p(y=-)}{p(x_1=T, x_2=T, x_3=F)}$$

Note that the denominator in both formulas above has the same constant value.

$$\begin{aligned}
&p(x_1=T, x_2=T, x_3=F|y=+)p(y=+)\\
&= p(x_1=T|y=+)p(x_2=T|y=+)p(x_3=F|y=+)p(y=+)\\
&= \frac{2}{3} * \frac{2}{3} * \frac{0}{3} * \frac{3}{8}\\
&= 0\\
&p_+ = 0/\underbrace{p(x_1=T, x_2=T, x_3=F)}_{constant} = 0
\end{aligned}$$

$$\begin{aligned}
&p(x_1=T, x_2=T, x_3=F|y=-)p(y=-)\\
&= p(x_1=T|y=-)p(x_2=T|y=-)p(x_3=F|y=-)p(y=-)\\
&= \frac{2}{5} * \frac{2}{5} * \frac{4}{5} * \frac{5}{8}\\
&= \frac{2}{25}\\
&= 0.08\\
&p_- = 0.08/\underbrace{p(x_1=T, x_2=T, x_3=F)}_{constant} > 0
\end{aligned}$$

Since $p_+ < p_-$, we should assign class "-" to this test example.

(b) ________ out of **(10 points)**: Repeat the question in part (a) using the following Bayesian Belief network as the classifier:

$$x_3 \to x_2 \to y \to x_1$$

Answer:

$$x_3 \to x_2 \to y \to x_1$$
$$\Downarrow$$

$$p(x_1, x_2, x_3, y) = p(x_1|y)\,p(x_2|x_3)\,p(x_3)\,p(y|x_2)$$

Using conditional probability, we get

$$\begin{aligned}
p_+ &= p(y=+|x_1=T,x_2=T,x_3=F) = \frac{p(x_1=T,x_2=T,x_3=F,y=+)}{p(x_1=T,x_2=T,x_3=F)} \\
&= \frac{p(x_1=T|y=+)\,p(x_2=T|x_3=F)\,p(x_3=F)\,p(y=+|x_2=T)}{p(x_1=T,x_2=T,x_3=F)} \\
p_- &= p(y=-|x_1=T,x_2=T,x_3=F) = \frac{p(x_1=T,x_2=T,x_3=F,y=-)}{p(x_1=T,x_2=T,x_3=F)} \\
&= \frac{p(x_1=T|y=-)\,p(x_2=T|x_3=F)\,p(x_3=F)\,p(y=-|x_2=T)}{p(x_1=T,x_2=T,x_3=F)}
\end{aligned}$$

Note that the denominator in both formulas above has the same constant value.

$$\begin{aligned}
&p(x_1=T,x_2=T,x_3=F,y=+) = \\
&= p(x_1=T|y=+)\,p(x_2=T|x_3=F)\,p(x_3=F)\,p(y=+|x_2=T) \\
&= (2/3)*(1/2)*(1/2)*(1/2) = \frac{1}{12} = 0.083\ldots \\
&p_+ = \frac{1}{12}/\underbrace{p(x_1=T,x_2=T,x_3=F)}_{constant} \\
&p(x_1=T,x_2=T,x_3=F,y=-) = (2/5)*(1/2)*(1/2)*(1/2) = \frac{1}{20} = 0.05 \\
&p_- = \frac{1}{20}/\underbrace{p(x_1=T,x_2=T,x_3=F)}_{constant}
\end{aligned}$$

Since $p_+ > p_-$, we should assign class "+" to this test example.

6. ______ out of (**10 points**): Show the key steps of the Junction Tree algorithm for the following DAG (Note: No need to show the details of the last step, i.e., the Belief Propagation inference step):

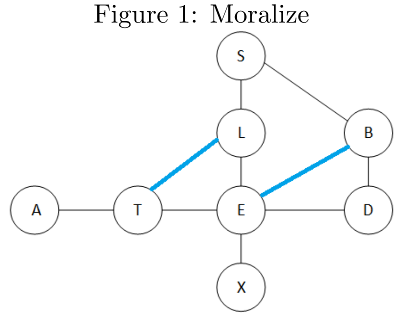
$$A \to T \to E \to D; S \to L \to E \to X; S \to B \to D$$

Answers:

1.) Start with a DAG.

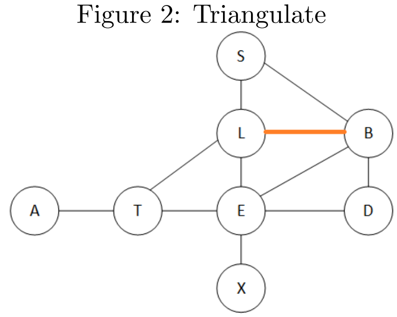
2.) Moralize: marry parents of each node (add edges $L - T$ and $B - E$) and remove edge directions.

Figure 1: Moralize



3.) Triangulate: if there is loops of length > 3, split it into triangles (add edge $L - B$ to triangulate).

Figure 2: Triangulate



4.) Find cliques: there are 6 cliques $AT, ELT, SLB, BLE, BDE, EX$.

5.) Form a junction tree of cliques (3 possibilities for a minimum spanning tree).

Figure 3: Junction tree (possibility 1)

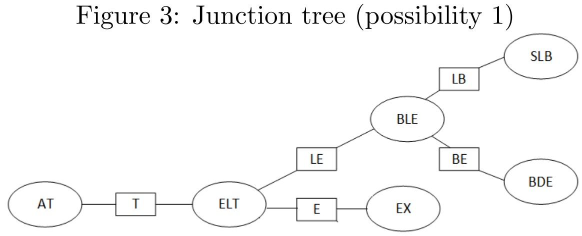


Figure 4: Junction tree (possibility 2)

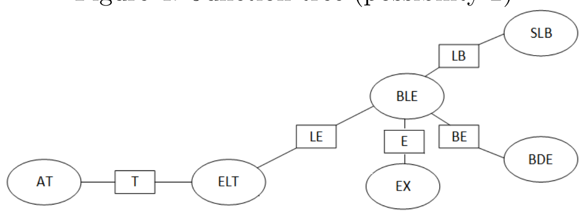
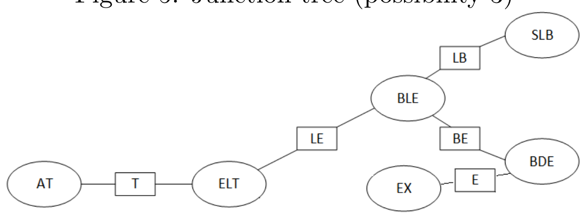


Figure 5: Junction tree (possibility 3)



6.) Do belief propagation in junction tree.

7. ______ out of (**15 points**): Using the Maximum Likelihood estimation, find the parameters for the following Bernoulli Trials problem (show all the derivations):
INPUT:

- M iid coin flips: $D = H, H, T, H, ...$
- Model: $p(H) = \theta$ and $p(T) = 1 - \theta$

OUTPUT: θ^*_{ML}

(a) ______ out of (**10 points**): What is the likelihood function $l(\theta; D) = \log p(D|\theta)$? (Hint: Introduce a random variable x that is equal to 1 if the trial is a Head and it is equal to zero if the trial is a Tail.)
Answer:

$$
\begin{aligned}
& l(\theta; D) = \log p(D|\theta) \\
& = \log\left(\prod_{m=1}^{M} \theta^{x^m} (1-\theta)^{1-x^m} \right) \\
& = \sum_{m=1}^{M} log(\theta^{x^m} (1-\theta)^{1-x^m}) \\
& = \log(\theta) \sum_m x^m + \log(1-\theta) \sum_m (1 - x^m) \\
& = \log(\theta) N_H + \log(1-\theta) N_T
\end{aligned}
$$

(b) ______ out of (**5 points**): What is θ^*_{ML}?
Answer:

$$
\begin{aligned}
& \frac{\partial l}{\partial \theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} \\
& \Rightarrow \theta^*_{ML} = \frac{N_H}{N_H + N_T}
\end{aligned}
$$