ECE/CS 498 DSU/DSG Spring 2020
In-Class Activity 2

Names: _____

NetID: _____

The purpose of the in-class activity is for you to:

(i) Review how to go from the word description of a problem to Bayesian Network (BN)
(ii) Factorize the joint distribution using conditional independence assumptions depicted by the BN
(iii) Determining the reduction in parameters in a BN over the joint distribution

Problem: Our TAs Vikram and Shengkun are working hard to analyze a new patient record dataset obtained from Carle Foundation Hospital. To help them in their data analysis, they install a software called Bayes Engine on their lab computer.

The Bayes Engine can either run fast or slow, depending on the computer configuration (high, medium, low), resource utilization by other processes (high, low) and presence of a computer virus. For the lab computer, assume the configuration, resource utilization and virus infection are independent from each other. Furthermore, the presence of a virus may cause 2 events: 1) the computer's anti-virus may raise a warning 2) spam messages are displayed on the home screen.

Today, Vikram and Shengkun observe that their Bayes Engine is running ***slow***. They suspect that their computer is infected with a virus since they are presented with spam messages. However, the anti-virus has not raised a warning. On further investigation, Vikram finds that their lab computer has a high configuration and Shengkun notes that there are many other applications running on the system, resulting in a high resource utilization. Can you help Vikram and Shengkun figure out if their computer is infected with a virus?

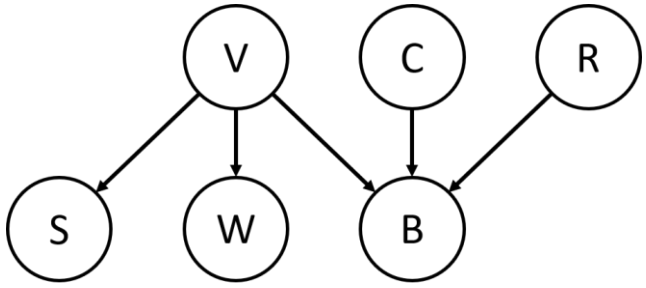
Part 1: Constructing the network

1. What are the variables in the above problem? What values can the variables take?
   Virus (V) ∈ {Yes, No}
   Spam (S) ∈ {Yes, No}
   Warning from Antivirus Software (W) ∈ {Yes, No}
   Configuration (C) ∈ {High, Medium, Low}
   Resource Utilization (R) ∈ {High, Low}
   Speed of Bayes Engine (B) ∈{Fast, Slow}

2. What are the nodes of the BN for the above problem?
   The variables defined in question 1 - V, S, W, C, R, B

3. What are the edges of the BN for the above problem? Draw the Bayesian Network.

   Virus (V) affects Spam (S)
   Virus (V) affects Warning of anti-virus software (W)
   Virus (V) affects the speed of Bayes Engine (B)
   Configuration (C) affects the speed of Bayes Engine (B)
   Resource Utilization (R) affects the speed of Bayes Engine (B)



Check your BN with a TA to confirm that you have a correct graph.

Graph constructed! You look up your notes to see the next steps. The lecture slides state that factorizing the joint distribution requires applying the chain rule and then applying local semantics (conditional independence assumptions) as specified by the structure of the graph.

Part 2: Applying local semantics

4. Apply chain rule to the joint distribution. [Hint: Recall that the parents should be conditioned on.]
   $P(V,S,W,B,C,R) = P(S|V,W,B,C,R)\ P(W|V,B,C,R)\ P(B|V,C,R)\ P(C|V,R)\ P(V|R)\ P(R)$

5. Simplify the terms of the joint distribution by using conditional independence assumptions. Specify the factorized joint distribution. [Hint: the probability of a node given its non-descendants is the probability of the node given its parents…]

   $P(S|V,W,B,C,R) = P(S|V)$
   $P(W|V,B,C,R) = P(W|V)$
   $P(B|V,C,R) = P(B|V,C,R)$
   $P(C|V,R) = P(C)$
   $P(V|R) = P(V)$

   $P(V,S,W,B,C,R) = P(S|V)P(W|V)P(B|V,C,R)P(C)P(V)P(R)$

Part 3: Determining Reduction in Required Parameters

6. How many parameters would a full joint distribution over the all variables require?
   Total number of parameters of joint distribution $P(V, S, W, C, R, B)$
   = (number of distinct combinations of (V, S, W, C, R, B)) - 1
   = (product of number of values each parameter can take) – 1
   = (2x2x3x2x2x2) -1= 96 – 1
   = 95 parameters

7. How many parameters are required to specify the factorized joint distribution?
   P(C): 2, P(R): 1, P(V): 1, P(B|V,C,R): 12, P(W|V) :2, P(S|V) :2
   Total = 2 + 1 + 1 + 12 + 2 + 2 = 20

8. Good going! By using conditional independence, you have reduced the number of parameters needed for joint distribution. Now let's look at what Vikram and Shengkun want you to judge... Express the question about the computer being infected with a virus, given all other observed conditions, *as a comparison of probability expressions*.

   Is P(V=Yes | C=High, R=High, S=Yes, W=No, B=Slow) > P(V=No | C=High, R=High, S=Yes, W=No, B=Slow)?

9. Suppose we are going to apply the MAP decision rule to determine if there is a virus given the evidence. List all the parameters needed to make this decision. What is the minimum number of required parameters?
   The parameters required to make the decision are shown below in red.

$$P(V = Yes \mid B = Slow, C = High, R = High, S = Yes, W = No)$$
$$\propto P(S = Yes|V = Yes) * P(W = No|V = Yes) * P(B = Slow|V = Yes, C = High, R = High) * P(C = High) * P(R = High) * P(V = Yes)$$

$$P(V = No \mid B = Slow, C = High, R = High, S = Yes, W = No)$$
$$\propto P(S = Yes|V = No) * P(W = No|V = No) * P(B = Slow|V = No, C = High, R = High) * P(C = High) * P(R = High) * P(V = No)$$

   Note that priors P(C=High) and P(R=High) have been used twice in the above expressions. Furthermore, we don't need the parameter P(V=No) since P(V=No) = 1 - P(V=Yes).

   List of **unique** parameters needed: $P(S = Yes|V = Yes)$, $P(S = Yes|V = No)$, $P(W = No|V = Yes), P(W = No|V = No), P(B = Slow|V = Yes, C = High, R = High)$, $P(B = Slow|V = No, C = High, R = High), P(C = High), P(R = High), P(V = Yes)$

   Minimum number of required parameters = 9.

In order to make the required MAP comparison, technically $P(C = High)$ and $P(R = High)$ don't need to be known since they are identical for both posteriors. Thus, we also accept that the **minimum number of required parameters is 7**

$[P(S = Yes|V = Yes),\ P(S = Yes|V = No),$
$P(W = No|V = Yes), P(W = No|V = No), P(B = Slow|V = Yes, C = High, R = High),$
$P(B = Slow|V = No, C = High, R = High), P(V = Yes)]$