STA 247 — Assignment #1, Due in class on November 3 at 3:10pm

Late assignments will be accepted only with a valid medical or other excuse.
This assignment is to be done by each student individually.
There are four pages to this assignment.

Part I

1. Suppose that A, B, and C are events for which $P(B)$, $P(B \cap C)$, and $P(B \cap C^c)$ are not zero. Using only the basic axioms of probability (on page 8), the definition of conditional probability (on page 27), and properties of sets, prove that

$$P(A \mid B) \;=\; P(A \mid B \cap C)P(C \mid B) \;+\; P(A \mid B \cap C^c)P(C^c \mid B)$$

2. A programmer writes a program to construct poems automatically, as follows. First, the program randomly selects ten distinct words from a dictionary. Then, starting with a poem with no words, the program repeatedly picks a word randomly from among these ten words, with equal probabilities, and adds it to the end of the poem. Words can be selected and added to the poem more than once, but when the word selected is the same as the word selected immediately before, the program stops, without adding this word to the end of the poem a second time.

   What is the probability that a poem produced by this program will be exactly three words long? If the poem is exactly three words long, how likely it is that the first and last words will be the same?

3. Suppose that computers A and C can communicate only via computer B. So, for a network packet to get from computer A to computer C, it must first be sent from computer A to computer B, and then from computer B to computer C. Packets sent from computer A to computer B usually take 5 milliseconds, but with probability 0.1, the packet will take 15 milliseconds. Packets sent from computer B to computer C usually take 10 milliseconds, but with probability 0.2, the packet will take 20 milliseconds. The time taken from A to B is independent of the time taken from B to C. The time required for other operations is negligible.

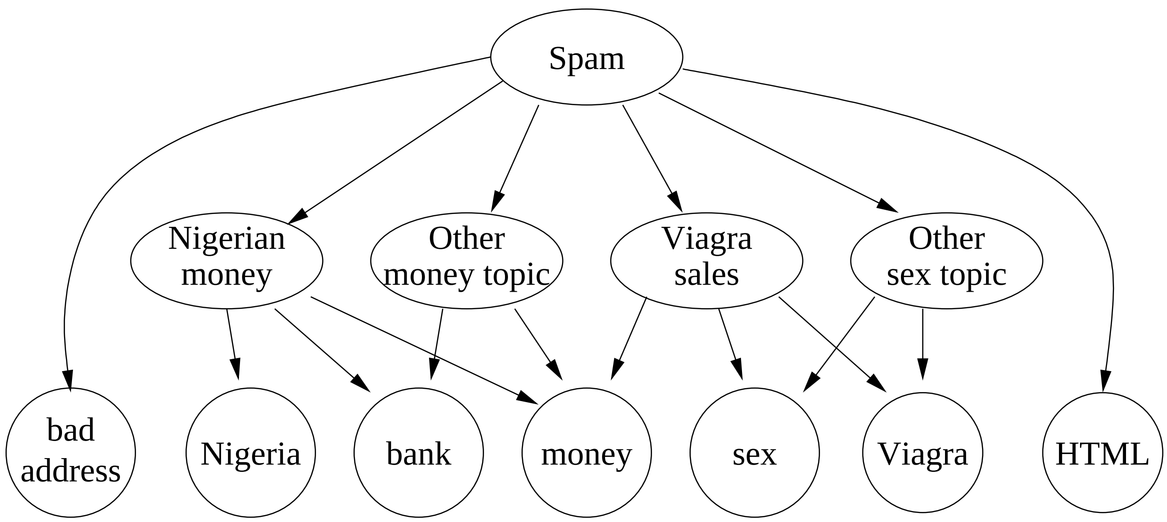
   Let X be the total time for a packet to be sent from computer A to computer C. Find the probability mass function for X, and compute $E(X)$ and $STD(X)$.

Part II

Unwanted e-mail — commonly called "spam" — is now a major problem. Many people use "spam filters" to automatically discard e-mail that looks like spam. Some of these filters use a probabilistic model to determine how likely a piece of e-mail is to be spam, and then discard it if the probability of it being spam is sufficiently high (eg, greater than 0.99, or whatever level the user thinks is appropriate).

Here, we'll consider a simple probabilistic model for e-mail, that incorporates some of the common characteristics of spam. (A real spam filter would need a more complex model.) Your task will be to write an R program to simulate generation of e-mail according to this model, and to use this simulation to determine how likely it is that a piece of e-mail with certain known characteristics is spam.

The model is based on a "causal network" (also called a "belief network") that shows how each random variable depends on its "parent" random variables. Here is a picture of the model for an e-mail message:



The top node represents the **Spam** random variable, which has four possible values:

Spam = 0	Not a spam message (ie, not unwanted e-mail)
Spam = 1	Spam about Nigerian money transfers
Spam = 2	Spam offering to sell Viagra cheaply
Spam = 3	Some other spam message

The four nodes below this have possible values of 0 and 1, representing whether or not the message is about each of four possible topics (1 means it's about this topic, 0 that it isn't about this topic). The **Nigeria money** and **Viagra sales** random variables are 1 if the message is spam of the corresponding type, and have only a small probability of being 1 otherwise. The **Other money topic** and **Other sex topic** random variables have a fairly high probability of being 1 if the message is some other kind of spam (ie, Spam=3), but they are also sometimes 1 for messages that aren't spam.

The bottom row of random variables represent things we can find out about the message. They all have values of 0 or 1. The **bad address** random variable is 1 if the reply address for the e-mail appears to be invalid. This is more likely for spam messages than for messages that aren't spam. Similarly, the **HTML** random variable is 1 if the message is in HTML format, rather than plain text. It also is more likely for spam messages. The remaining five random variables indicate whether each of the words "Nigeria", "bank", "money", "sex", "Viagra" appear in the message, with the value 1 indicatating that the word does appear.

The model needs to specify the joint probability mass function for all possible combinations of values for these twelve random variables. There are 8192 such possible combinations, so we'd rather not specify the probability of each combination of values separately. Instead, we use the fact that we can always write the joint probability for a set of random variables — eg, W, X, Y, and Z as the following product, once we've chosen some order for them:

$$P(W = w,\, X = x,\, Y = y,\, Z = z)$$

$$= \; P(W = w)\, P(X = x \,|\, W = w)\, P(Y = y \,|\, W = w,\, X = x)\, P(Z = z \,|\, W = w,\, X = x,\, Y = y)$$

If we can furthermore simplify some of the factors on the right, we may be able to specify the model using many fewer numbers.

When the model is specified using a causal network, we order the variables so that all the arrows go forward (top to bottom in this case), and then use conditional probabilities that are conditional only on the "parents" of a variable. A variable X is a parent of variable Y if there is an arrow from X to Y. For instance, in the network above, we can do the following simplification:

$$P(\text{bank} = 1 \,|\, \text{Spam} = 3,\, \text{Nigerian-money} = 0,\, \text{Other-money-topic} = 1,\, \text{Viagra-sales} = 0,$$
$$\text{Other-sex-topic} = 1,\, \text{bad-address} = 0,\, \text{Nigeria} = 0)$$

$$= \; P(\text{bank} = 1 \,|\, \text{Nigerian-money} = 0,\, \text{Other-money-topic} = 1)$$

On the next page, all the probabilities needed to specify the joint distribution for the variables in this causal network are given. You are to use the network and these probabilities to write an R function called `generate.email`, which takes no arguments, and which returns a randomly generated set of values for all the random variables in the network. These values should be returned as an R "list" with twelve elements, named `Spam`, `Nigerian.money`, etc.

You should then write an R function called `spam.probability`, which takes as arguments the values of the seven random variables in the bottom row of the network, and returns an estimate of the conditional probability that the message is spam (ie, that **Spam** is not zero) given that the bottom seven random variables have the values specified. This is the procedure that the spam filter would use to decide whether to discard the e-mail, by comparing this probability with the threshold set by the user (eg, 0.99).

You should implement the `spam.probability` function by randomly generating a number of e-mail messages using `generate.email`. You should count how many of these messages match the values of the seven known random variables, and of these matching messages, how many are spam. The ratio of these numbers gives an estimate of the conditional probability that the actual message is spam. You should continue generating messages until *either* you have generated 10000 messages in total, *or* you have generated 1000 messages that match the known values of the seven random variables.

You should hand in (on paper) the following:

a) A listing of your R program, with suitable (but not excessive) comments.

b) The output of five calls of your `generate.email` function.

c) The output of your `spam.probability` function when called with the following five sets of arguments (values in order from left to right):

   0 0 0 0 0 0 0, 1 0 0 0 0 0 0, 1 1 0 1 0 0 0, 0 0 0 0 1 0 1, 0 0 1 1 0 1 0

d) A brief discussion of how much the answers you got for (c) vary when you run your function again (without resetting the random number seed).

Here are the probabilities you will need:

$P(Spam = 0) \;=\; 0.6, \;\; P(Spam = 1) \;=\; 0.1, \;\; P(Spam = 2) \;=\; 0.1, \;\; P(spam = 3) \;=\; 0.2$

$P(bad\text{-}address = 1 \,|\, Spam = 0) \;=\; 0.1, \;\; P(bad\text{-}address = 1 \,|\, Spam \neq 0) \;=\; 0.8$

$P(HTML = 1 \,|\, Spam = 0) \;=\; 0.2, \;\; P(HTML = 1 \,|\, Spam \neq 0) \;=\; 0.4$

$P(Nigerian\text{-}money = 1 \,|\, Spam = 0) \;=\; 0.001, \;\; P(Nigerian\text{-}money = 1 \,|\, Spam = 1) \;=\; 1$
$P(Nigerian\text{-}money = 1 \,|\, Spam = 2) \;=\; 0, \;\; P(Nigerian\text{-}money = 1 \,|\, Spam = 3) \;=\; 0$

$P(Other\text{-}money\text{-}topic = 1 \,|\, Spam = 0) \;=\; 0.1, \;\; P(Other\text{-}money\text{-}topic = 1 \,|\, Spam = 1) \;=\; 0$
$P(Other\text{-}money\text{-}topic = 1 \,|\, Spam = 2) \;=\; 0, \;\; P(Other\text{-}money\text{-}topic = 1 \,|\, Spam = 3) \;=\; 0.3$

$P(Viagra\text{-}sales = 1 \,|\, Spam = 0) \;=\; 0.001, \;\; P(Viagra\text{-}sales = 1 \,|\, Spam = 1) \;=\; 0$
$P(Viagra\text{-}sales = 1 \,|\, Spam = 2) \;=\; 1, \;\; P(Viagra\text{-}sales = 1 \,|\, Spam = 3) \;=\; 0$

$P(Other\text{-}sex\text{-}topic = 1 \,|\, Spam = 0) \;=\; 0.1, \;\; P(Other\text{-}sex\text{-}topic = 1 \,|\, Spam = 1) \;=\; 0$
$P(Other\text{-}sex\text{-}topic = 1 \,|\, Spam = 2) \;=\; 0, \;\; P(Other\text{-}sex\text{-}topic = 1 \,|\, Spam = 3) \;=\; 0.4$

$P(Nigeria = 1 \,|\, Nigerian\text{-}money = 0) \;=\; 0.05, \;\; P(Nigeria = 1 \,|\, Nigerian\text{-}money = 1) \;=\; 1$

$P(bank = 1 \,|\, Nigerian\text{-}money = 0, Other\text{-}money\text{-}topic = 0) \;=\; 0.1$
$P(bank = 1 \,|\, Nigerian\text{-}money = 1, Other\text{-}money\text{-}topic = 0) \;=\; 0.5$
$P(bank = 1 \,|\, Nigerian\text{-}money = 0, Other\text{-}money\text{-}topic = 1) \;=\; 0.5$
$P(bank = 1 \,|\, Nigerian\text{-}money = 1, Other\text{-}money\text{-}topic = 1) \;=\; 0.5$

$P(money = 1 \,|\, Nigerian\text{-}money = 0,\; Other\text{-}money\text{-}topic = 0, Viagra\text{-}sales = 0) \;=\; 0.2$
$P(money = 1 \,|\, Nigerian\text{-}money = 0,\; Other\text{-}money\text{-}topic = 0, Viagra\text{-}sales = 1) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 0,\; Other\text{-}money\text{-}topic = 1, Viagra\text{-}sales = 0) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 0,\; Other\text{-}money\text{-}topic = 1, Viagra\text{-}sales = 1) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 1,\; Other\text{-}money\text{-}topic = 0, Viagra\text{-}sales = 0) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 1,\; Other\text{-}money\text{-}topic = 0, Viagra\text{-}sales = 1) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 1,\; Other\text{-}money\text{-}topic = 1, Viagra\text{-}sales = 0) \;=\; 0.7$
$P(money = 1 \,|\, Nigerian\text{-}money = 1,\; Other\text{-}money\text{-}topic = 1, Viagra\text{-}sales = 1) \;=\; 0.7$

$P(sex = 1 \,|\, Viagra\text{-}sales = 0, Other\text{-}sex\text{-}topic = 0) \;=\; 0.05$
$P(sex = 1 \,|\, Viagra\text{-}sales = 1, Other\text{-}sex\text{-}topic = 0) \;=\; 0.6$
$P(sex = 1 \,|\, Viagra\text{-}sales = 0, Other\text{-}sex\text{-}topic = 1) \;=\; 0.6$
$P(sex = 1 \,|\, Viagra\text{-}sales = 1, Other\text{-}sex\text{-}topic = 1) \;=\; 0.6$

$P(Viagra = 1 \,|\, Viagra\text{-}sales = 0, Other\text{-}sex\text{-}topic = 0) \;=\; 0.01$
$P(Viagra = 1 \,|\, Viagra\text{-}sales = 1, Other\text{-}sex\text{-}topic = 0) \;=\; 1$
$P(Viagra = 1 \,|\, Viagra\text{-}sales = 0, Other\text{-}sex\text{-}topic = 1) \;=\; 0.1$
$P(Viagra = 1 \,|\, Viagra\text{-}sales = 1, Other\text{-}sex\text{-}topic = 1) \;=\; 1$