Multivariate Distributions

In the previous lecture, we dealt only with two variables, even though one of them was not binary.

Let's take a look at a multi-variate distribution taking a (not large) step to three variables.

For this example, let's just assume we have 3 binary variables X, Y, and Z, with a joint probability table that looks like:

X	**Y**	**Z**	**Pr(X, Y, Z)**
0	0	0	0.030
0	0	1	0.120
0	1	0	0.105
0	1	1	0.245
1	0	0	0.105
1	0	1	0.045
1	1	0	0.280
1	1	1	0.070

Let's talk about a couple of observations we can make on multivariate distributions.

Suppose we didn't want the entire joint distribution, but wanted to look at the co-occurences of only some subset of variables.

We can **sum out (marginalize)** a variable by "collapsing" the distribution on the remaining variables we're not summing out. Formally, for variables of interest α and variables we're summing out β with $\alpha \cap \beta = \varnothing$

$$P(\alpha) = \sum_{\beta_i \in \beta} P(\alpha, \beta_i)$$

ℹ The resulting distribution over variables α that have not been summed out is called the **joint marginal** distribution over $P(\alpha)$.

So, when we marginalize, we get a new table, or distribution, whose rows are the sums of probability mass over the variables that were summed out.

❓ Marginalize Z in the above distribution, leaving a joint marginal on X and Y. To do so, simply look for every row where X and Y *agree* and then sum over the possible values for Z.

(the individual colors don't mean anything but observe the corresponding rows)

These corresponding rows are where X and Y agree (i.e., in both rows 1 and 4 Y has value y and X has value x)

So, to sum out Z, we simply collapse across the two rows! It's as though X were removed entirely from the equation, just leaving us with a distribution over X and Y.

This amounts to the following joint marginal on X and Y with Z summed out, written:

$$\sum_Z Pr(X, Y, Z)$$

X	Y	$\sum_Z Pr(X, Y, Z) = Pr(X, Y)$
0	0	0.15
0	1	0.35
1	0	0.15
1	1	0.35

(Aside: that table looks really rasta)

Any who, observe how we now have a joint distribution on Y and Z.

Independence

Let's turn our attention to independence relationships, remembering that our definition of independence was that:

If X is independent from Y, written

$$X \perp\!\!\!\perp Y$$

...then:

$$Pr(X|Y) = Pr(X), \forall x, y$$

So using our joint marginal over X and Y above, let's test them for independence.

Repeating our joint marginal from before:

X	Y	$\sum_Z Pr(X, Y, Z) = Pr(X, Y)$
0	0	0.15
0	1	0.35
1	0	0.15
1	1	0.35

Using the joint marginal above, determine whether or not $X \perp\!\!\!\perp Y$. Click for solution.

Neat! So this means that knowing something about X tells me nothing about the state of Y and vice versa...

E.g., knowing that it's Tuesday tells me nothing about my chances of winning the lottery.

If we were so inclined, we could repeat the process of summing out X from the original joint distribution to get:

Y	Z	$\sum_X Pr(X, Y, Z)$
0	0	0.135
0	1	0.165

Y	Z	
1	0	0.385
1	1	0.315

Performing the same test for independence, we would find that:

$$\begin{aligned} Pr(Y=0|Z=0) &= \frac{Pr(Y=0,Z=0)}{Pr(Z=0)} \\ &= \frac{0.135}{0.52} \\ &\approx 0.260 \\ &\neq Pr(Y=0) \\ \therefore Y &\not\perp\!\!\!\perp Z \end{aligned}$$

Conditional Independence

A topic we haven't talked about yet is a peculiar phenomenon known as conditional independence.

As it turns out, it's possible for two variables X and Y to be independent ONLY after we've observed (conditioned upon) some other variable(s) Z.

To compare:

Relationship	Description	Written
Independence	If we have knowledge that X occurred, and that tells us nothing about whether or not Y occured, then X is independent of Y and vice versa.	$X \perp\!\!\!\perp Y$
Conditional Independence	X and Y are conditionally independent if and only if, given information about Z, having knowledge about X tells us nothing about whether or not Y occured; i.e., X and Y may not be independent until *after* conditioning on a third variable / set of variables Z.	$X \perp\!\!\!\perp Y\|Z$

So you might be curious... what's an intuitive interpretation of conditional independence?

Here are some good good scenarios that explain it, using dice, because every statistician loves dice for some reason:

Concept	Description	Written
Independent	You roll two dice: A and B. Knowing the outcome of A tells you nothing about the outcome of B.	$A \perp\!\!\!\perp B$
Independent, but Conditionally Dependent	You roll two dice: A and B. If I tell you that the sum (S) of the two dice's totals is even, then knowing the value of either A or B actually *does* tell me something about the other, even though A and B are independent on their own.	$A \not\perp\!\!\!\perp B \vert S$
Independent, and Conditionally Independent	You roll two dice: A and B. If I tell you that the result (R) of A is not 3 and the result of B is not 2, I learn new information about each, but nothing that connects the two outcomes. So, the dice rolls are independent AND conditionally independent based on the new information.	$A \perp\!\!\!\perp B$ $A \perp\!\!\!\perp B \vert R$
Dependent, but Conditionally Independent	This one's a bit trickier and we'll need to develop a new toolkit to think about it, so let's hold off on it for now...	$A \not\perp\!\!\!\perp B$ $A \perp\!\!\!\perp B \vert C$

ℹ Two other, equivalent, ways to think about conditional independence are that:

$$Pr(X \vert Y, Z) = Pr(X \vert Z) \Leftrightarrow X \perp\!\!\!\perp Y \vert Z$$
$$Pr(X, Y \vert Z) = Pr(X \vert Z) Pr(Y \vert Z) \Leftrightarrow X \perp\!\!\!\perp Y \vert Z$$

So, let's take a look at a probability distribution that might elicit conditional independence relationships.

Example
☑ Consider the following example relating three variables in a (fictitious) study on health effects of smoking.

- **Smoking:** whether or not an individual smokes
- **Cancer:** whether or not an individual has cancer
- **Asthma:** whether or not an individual has asthma

Let's look at a made-up distribution:

Asthma	Smoking	Cancer	Pr(Asthma, Smoking, Cancer)
0	0	0	0.576
0	0	1	0.144
0	1	0	0.008
0	1	1	0.072
1	0	0	0.064
1	0	1	0.016
1	1	0	0.012
1	1	1	0.108

We can see that the instances of smoking are usually indicative of both cancer and asthma.

Now, we make the following observations:

1. We hypothesize that smoking causes cancer and asthma.
2. If this is the case, then knowing that someone smokes immediately tells us that they are likely to develop cancer and asthma.
3. Because of (1) and (2) above, we observe that IF we know someone smokes, then knowing that they have asthma tells us nothing more about them having cancer.

❓ What conditional independence relationship are the above observations making?

❓ Rationalize: why is this a conditional independence relationship and not an absolute independence relationship?

❓ Express this conditional independence in probability notation involving all three variables; how will we go about illustrating this independence relationship from the distribution?

Alright, now that we have our query in mind, let's observe the following:

Goal:

$$Pr(Asthma|Cancer, Smoking) = Pr(Asthma|Smoking)$$

We have neither of those tables, but we can compute them! Let's use Bayes' Conditioning and marginalization!

$$Pr(Asthma|Cancer, Smoking) = \frac{Pr(Asthma, Cancer, Smoking)}{Pr(Cancer, Smoking)}$$

$$Pr(Asthma|Smoking) = \frac{Pr(Asthma, Smoking)}{Pr(Smoking)}$$

Marginalizing from the joint distribution, we can get our two joint-marginals on $Pr(Cancer, Smoking)$ and $Pr(Asthma, Smoking)$:

Smoking	Cancer	Pr(Smoking, Cancer)
0	0	0.64
0	1	0.16
1	0	0.02
1	1	0.18

Smoking	Asthma	Pr(Smoking, Asthma)
0	0	0.72
0	1	0.08

Smoking	Asthma	Pr(Smoking, Asthma)
1	0	0.08
1	1	0.12

As a final ingredient for our proof of conditional independence, I'll save you the meager effort and tell you that:

$$Pr(Smoking = 1) = 0.2$$

Now, we just need to compute the two conditional distributions; we'll start with the Pr(Asthma | Cancer, Smoking) from the joint:

$$Pr(Asthma|Cancer, Smoking) = \frac{Pr(Asthma, Cancer, Smoking)}{Pr(Cancer, Smoking)}$$

Asthma	Smoking	Cancer	Pr(Asthma, Smoking, Cancer)	Pr(Asthma | Smoking, Cancer)
0	0	0	0.576	0.9
0	0	1	0.144	0.9
0	1	0	0.008	0.4
0	1	1	0.072	0.4
1	0	0	0.064	0.1
1	0	1	0.016	0.1
1	1	0	0.012	0.6
1	1	1	0.108	0.6

Interesting, looks like we have some flavor of uniformity going on here... hmmm... Let's compute Pr(Asthma | Smoking):

$$Pr(Asthma|Smoking) = \frac{Pr(Asthma, Smoking)}{Pr(Smoking)}$$

Smoking	Asthma	Pr(Asthma | Smoking)
0	0	0.9
0	1	0.1
1	0	0.4
1	1	0.6

Aha! We see that the two are in fact equivalent, regardless of what we know about whether or not a person has cancer! For brevity's sake, consider the positive variables below to be = 1, and the negated ones to be = 0:

$$\begin{aligned} Pr(\neg Asthma|\neg Cancer, \neg Smoking) &= Pr(\neg Asthma|Cancer, \neg Smoking) \\ &= Pr(\neg Asthma|\neg Smoking) \\ &= 0.9 \end{aligned}$$

$$\begin{aligned} Pr(\neg Asthma|\neg Cancer, Smoking) &= Pr(\neg Asthma|Cancer, Smoking) \\ &= Pr(\neg Asthma|Smoking) \\ &= 0.4 \end{aligned}$$

$$\begin{aligned} Pr(Asthma|\neg Cancer, \neg Smoking) &= Pr(Asthma|Cancer, \neg Smoking) \\ &= Pr(Asthma|\neg Smoking) \\ &= 0.1 \end{aligned}$$

$$\begin{aligned} Pr(Asthma|\neg Cancer, Smoking) &= Pr(Asthma|Cancer, Smoking) \\ &= Pr(Asthma|Smoking) \\ &= 0.6 \end{aligned}$$

$$\therefore Pr(Asthma|Cancer, Smoking) = Pr(Asthma|Smoking)$$

$$\therefore Asthma \perp\!\!\!\perp Cancer|Smoking$$

Whew! That was a lot of work! But we'll see in a moment that this is all worth it...

Bayesian Networks

Finally, we get to Bayesian Networks! I know you paid for your whole seat this discussion but you're only going to need the edge!

To commemorate the occasion, I've created an image deserving of the day's grandeur that I've wanted to make for some time now.

That, of course, is a picture of Father Thomas Bayes, the one responsible for the eponymous theorem, photoshopped onto David Hasselhoff's body from the hit drama series Baywatch from 1989.

I haven't had any means of working this joke into conversation, and the previous part of this lecture's been dull, so here we are...

You awake again? OK...

In the last section we derived that $Asthma \perp\!\!\!\perp Cancer | Smoking$ from our smoking example.

Let's look at an interesting consequence:

❓ Give a chain-rule factorization of the joint probability table: `Pr(Asthma, Cancer, Smoking)`

❓ Based on our conditional independence relation found from the previous section ($Asthma \perp\!\!\!\perp Cancer | Smoking$), can we reduce this to anything simpler?

Hmm, interesting... so we took a big joint probability table (Pr(Asthma, Cancer, Smoking)) and broke it down into 3 smaller tables!

There's something else interesting about our factorization...

Remembering that we presumed that both Asthma and Cancer were indicators of Smoking:

❓ Observation 1: Thinking about cause-effect relationships, what's interesting about the tables: Pr(Asthma | Smoking) and Pr(Cancer | Smoking)?

Alright, we're almost ready to hit the punchline... one final example.

Example
☑ Suppose we have 5 binary variables: A, B, C, D, and E. Each of these variables are pairwise independent.

❓ How many worlds are in the table / distribution:

$$Pr(A, B, C, D, E)$$

❓ Use the chain rule to factor this table. What is the resulting factorization?

❓ How many worlds are there in each of these factored tables? How many rows total?

ℹ Observation 2: The more independence relationships we're able to make on our distribution, the more compact we make the factored joint.

There are 32 rows in the full joint, but only 10 rows in the individual, factored tables!

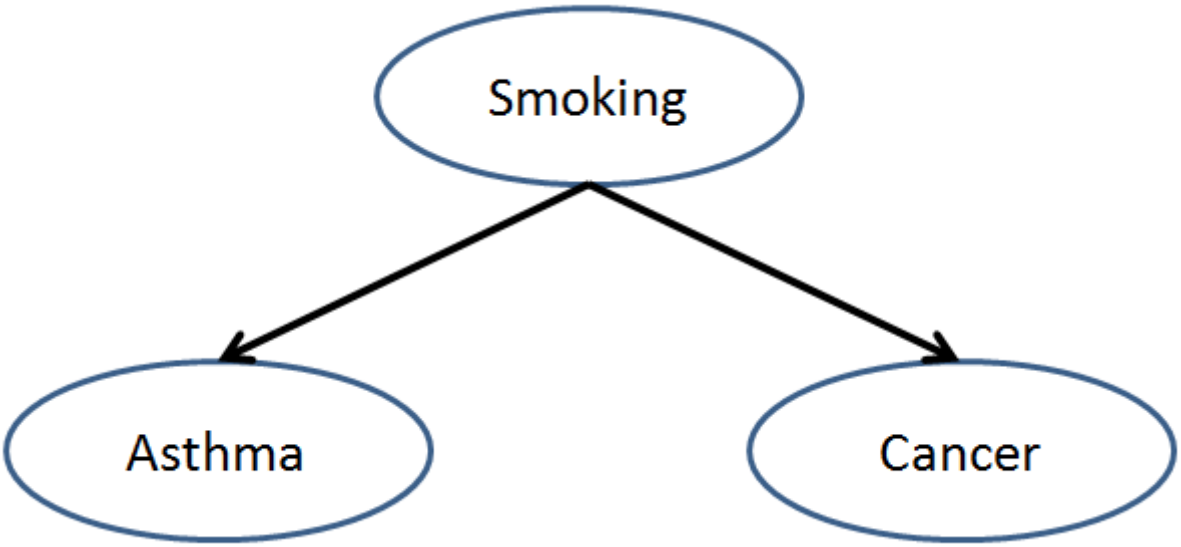
Observations 1 and 2 lead us to the beauty of Bayesian networks.

Bayesian Networks attempt to exploit independence relationships to reduce massive joint probability distributions to smaller tables, all while capturing the intuitive notions of cause and effect to structure the independences.

In the words of the great Judea Pearl, who was instrumental in their development, Bayesian networks are, "A parsimonious representation" of the joint distribution.

They're just really intuitive data structures!

Here's a simple Bayesian network representing our smoking problem:



Bayesian Network Properties

Bayesian networks belong to a class of graphs called **directed, acyclic graphs (DAGs)**, meaning that the edges between nodes are directed and they form no cycles (derp).

The network's **nodes** represent the variables in our distributions.

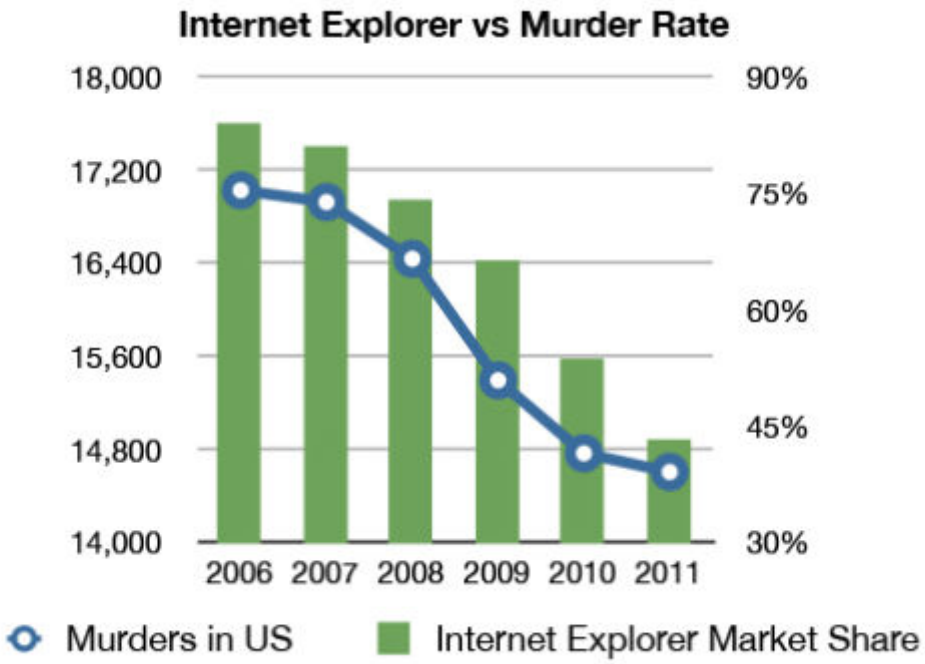
The network's **edges** represent **dependence** relationships, i.e., two connected nodes are dependent upon one another.

ℹ Edges illustrate only a dependence; effects can have multiple causes and causes can have multiple effects, any one of which illustrates an influence that may be negative OR positive (correlationally).

ℹ The **edge directions** are merely tools to represent the independence relationships, but are structured with potential causes pointing to potential effects.

Why do we say "potential causes" and "potential effects?"

Because our data is correlational!



(http://gizmodo.com/5977989/internet-explorer-vs-murder-rate-will-be-your-favorite-chart-today)
We would need stronger tools to claim confidence in a true causal relation, but for the purposes of Bayesian networks, it is convenient and intuitive to draw arrows from causes to effects.

The reason being the same as in our smoking example: if we know that someone has a Smoking (the cause), then knowing that they also have a Asthma doesn't tell us more about their chance of having a Cancer.

That is, as soon as we know the state of the causes, we don't get any new information about the effects!

Thus, we glean some notion of the **semantics / meaning** implicit within Bayesian networks:

ℹ The **Markovian factorization** of a Bayesian network says that we can use the chain rule to phrase the joint distribution as a product of "family" factors, composed of a node given its parents.

The joint distribution can be factored using independence relationships to:

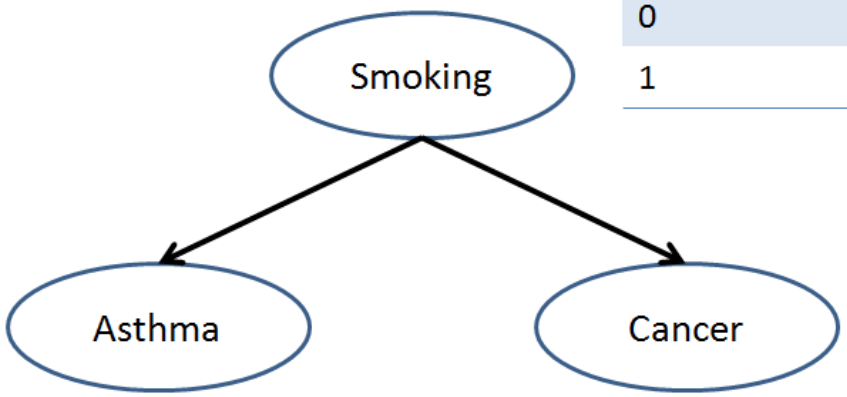
$$Pr(X_1, X_2, \ldots, X_n) = Pr(X_1|Parents(X_1)) * Pr(X_2|Parents(X_2))*\ldots$$

...which semantically represents the putative relationship between causes and effects:

$$\begin{aligned} & Pr(X_1|Parents(X_1)) * Pr(X_2|Parents(X_2))*\ldots \\ & = Pr(Effect_1|Causes(Effect_1)) * (Effect_2|Causes(Effect_2))*\ldots \\ & = \prod_{X_i \in Vars} Pr(X_i|Parents(X_i)) \end{aligned}$$

ℹ This factorization allows us to express the large joint distribution in terms of **conditional probability tables (CPTs)** of an effect given its parents.

For our smoking example, the CPTs would look like:



Smoking	Pr(Smoking)
0	0.8
1	0.2

Smoking	Asthma	Pr(Asthma | Smoking)
0	0	0.9
0	1	0.1
1	0	0.4
1	1	0.6

Smoking	Cancer	Pr(Cancer | Smoking)
0	0	0.8
0	1	0.2
1	0	0.1
1	1	0.9

❓ Using the above CPTs, compute: Pr(Smoking = 0, Asthma = 1, Cancer = 0)

Since our CPTs describe a world in which the effects can be screened off from other portions of the network by knowing about their causes, we say that Bayesian networks make the Markovian assumption:

The **Markovian assumption** states that every node is **independent** of its non-descendents **given** its parents, or formally, for node X, we write:

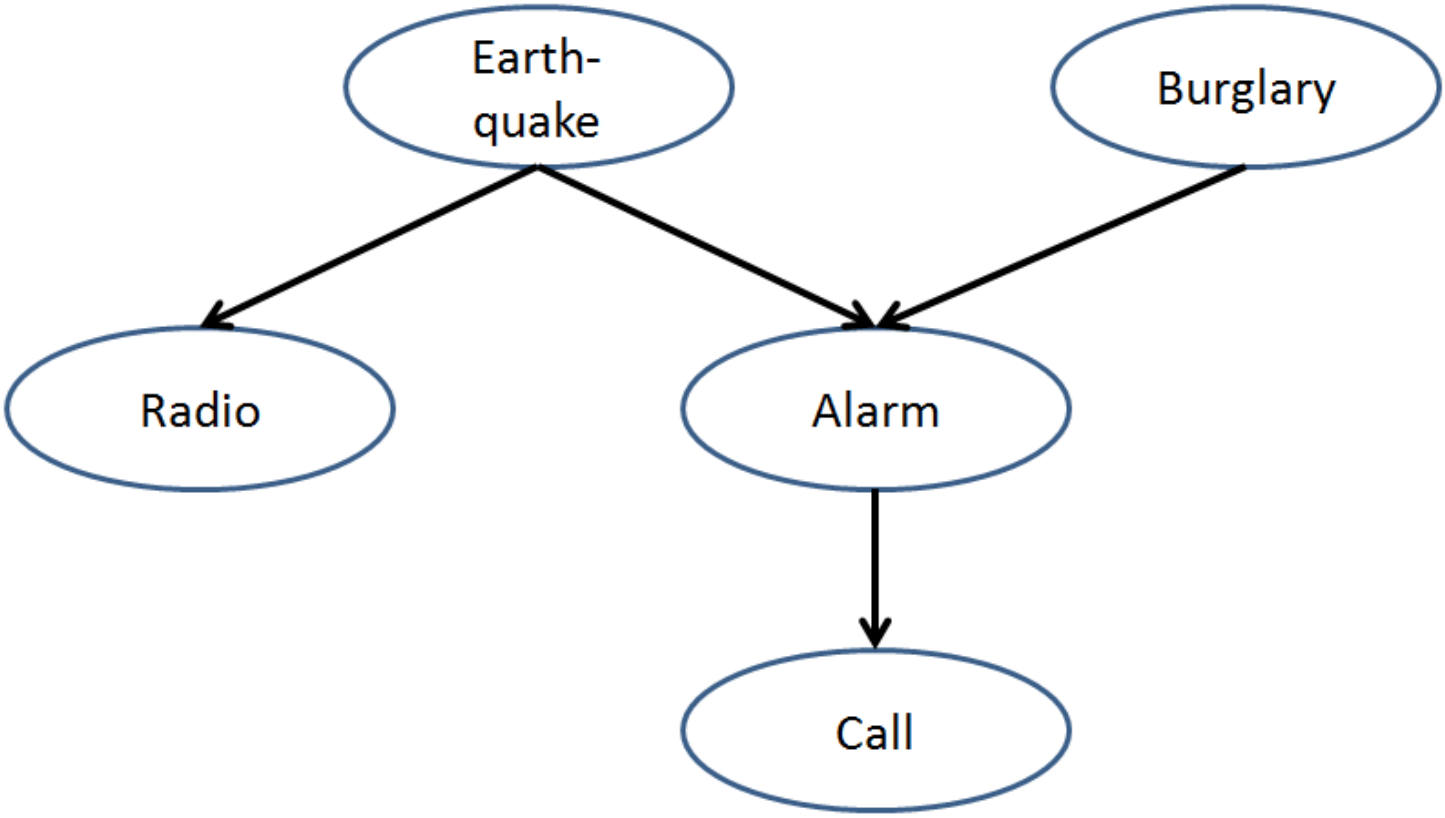
$$X \perp\!\!\!\perp NonDescendants(X) | Parents(X)$$

Non-descendants of node X are any node that can NOT be reached by taking a directed path starting at X.

You should repeat that a couple times... make it your mantra.

The Markovian assumptions are intuitive because they convey the "screening off" of extraneous effects once we know all of the causes for a given variable.

Example
What are the Markovian assumptions implicit in the following Bayesian network?



Click for answer.

Markovian assumptions are great and can give us some off-the-top independence relationships encoded by our networks...

...however, they don't give us a clean way of asking arbitrary claims of independence between nodes given other nodes.

To help us with such queries, we turn to the notion of d-separation.

D-Separation

Let's start off by considering 3 different simple Bayesian networks and observe how we can generalize their characteristics.

Consider each of the following 3-node Bayesian networks as a plumbing network where:

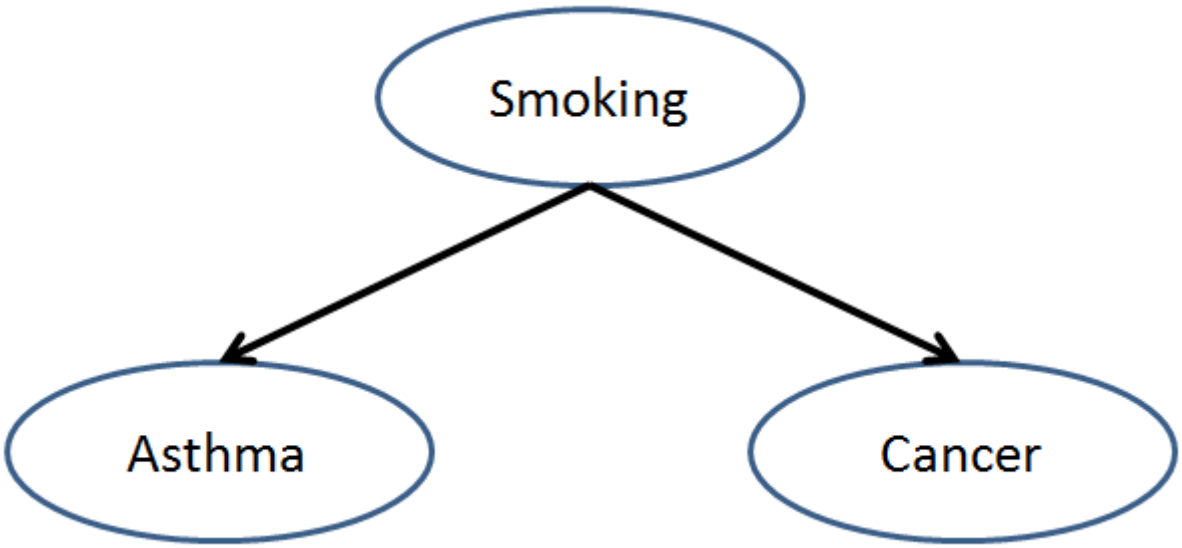
1. Nodes are "valves" that allow water (information) to flow through
2. Edges are "pipes" that connect the valves

3. Dependence is the process of determining whether water (information) could flow from some set of valves (variables) X to some other set of valves Y, accounting for whether or not any valves along any pipe-path are closed or not (our evidence, some set of variables Z)

We'll start with a familar problem:

ℹ A valve Z is **divergent / a fork** along some path if it is a common cause of two effects, X and Y. Moreover, whenever we're given Z, X is independent from Y.

$$X \leftarrow Z \rightarrow Y \Rightarrow X \perp\!\!\!\perp Y|Z$$



❓ Intuit: why does knowing whether or not someone Smokes make knowing whether or not they have Asthma tell us nothing more about their likelihood of having Cancer?

Divergent nodes are sometimes referred to as common causes. Here, we see that if we know whether or not someone has a Smoking, then information does NOT flow from Asthma to Cancer.

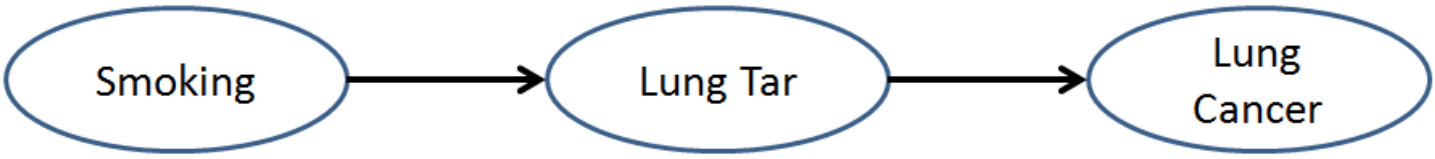
So, the rule for divergent valves along some path is that when we have a triplet X ← Z → Y, given Z blocks information flow from X to Y.

ℹ A valve Z is **sequential / a chain** along some path if some other variable X is its cause and Z has some effect Y.

$$X \rightarrow Z \rightarrow Y \Rightarrow X \perp\!\!\!\perp Y|Z$$

$$Y \leftarrow Z \leftarrow X \Rightarrow X \perp\!\!\!\perp Y|Z$$

Here's an example path of a chain:



❓ Intuit: why does knowing whether or not someone has Lung Tar make knowing whether or not they Smoke tell us nothing more about their likelihood of having Cancer?

Here, knowing that someone has tar in their lungs means that we no longer get information flowing from any knowledge that the person smoked to whether they'll have lung cancer.

"If Lung Tar (the medical term, of course) is the true cause of Lung Cancer, then it's irrelevant how the tar got there in the first place to whether or not you have Lung Cancer."

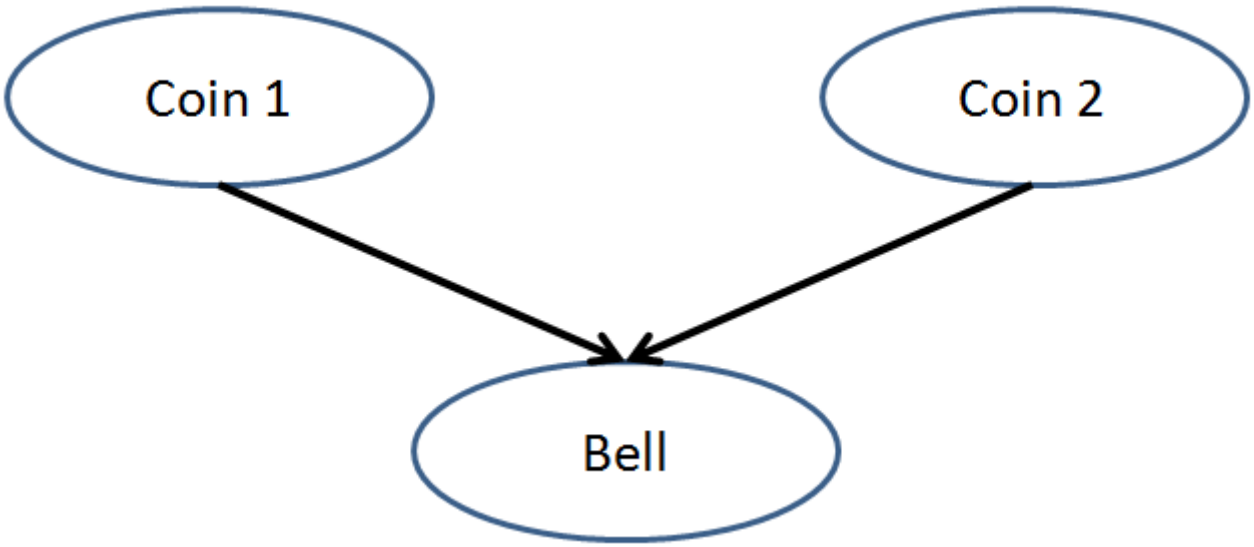
So, the rule for chain valves along some path is that for triple $X \rightarrow Z \rightarrow Y$, given Z blocks information flow from X to Y.

ℹ A valve Z is **convergent / a sink** along some path if it is the common effect of two causes, X and Y.

$$X \rightarrow Z \leftarrow Y \Rightarrow X \perp\!\!\!\perp Y|Z$$
$$\Rightarrow X \perp\!\!\!\perp Y|Descendants(Z)$$

Convergent nodes are a special beast, so let's look at the following scenario:

> Consider the scenario where a bell rings if and only if the outcome of two coin flips (from two separate coins) are identical (i.e., both heads or both tails).

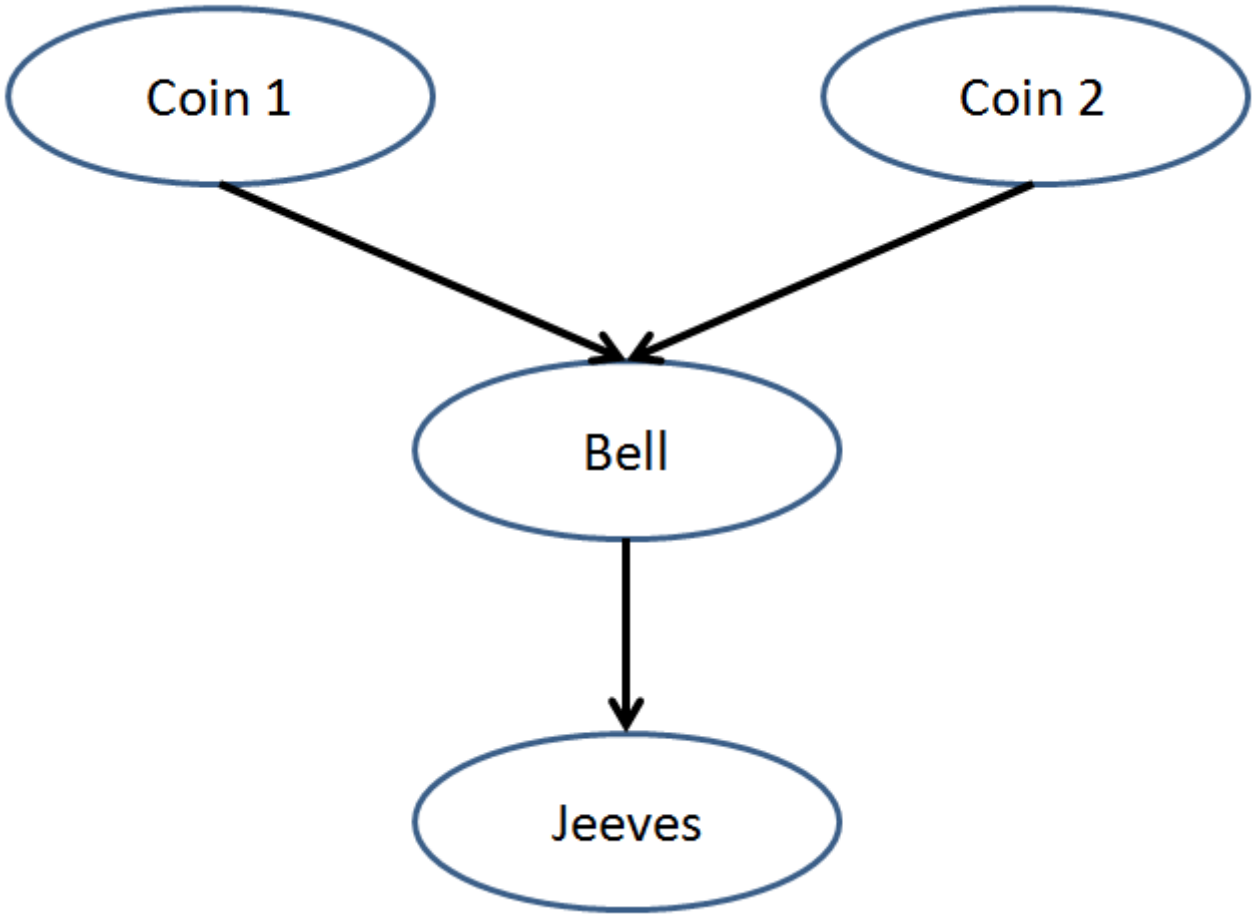


❓ If I know whether or not **the bell rings**, does information flow from knowing about coinflip 1 to coinflip 2?

❓ If I DO NOT know whether or not the bell rang, does information flow from knowing about coinflip 1 to coinflip 2?

Now consider the following scenario:

> A bell rings if and only if the outcome of two coin flips (from two separate coins) are identical (i.e., both heads or both tails). Additionally, if the bell rings, our elderly butler, Jeeves, has an 80% chance of bringing tea to our room.



❓ If I know whether or not **Jeeves brings tea**, does information flow from knowing about coinflip 1 to coinflip 2?

❓ If I DO NOT know whether or not the bell rang AND I DO NOT know whether Jeeves brought tea, does information flow from knowing about coinflip 1 to coinflip 2?

So, the rule for convergent valves is the following: for common effect Z in configuration: $X \rightarrow Z \leftarrow Y$, then Z is blocked if neither it NOR any of its descendants are given!

d-Separation

Now, we can think about these three cases of valves as being elements along a path in a Bayesian network.

d-separation (directional separation) allows us to determine all independence relations implicit from the graph just by looking at its structure.

We use the following notation to pose d-separation queries and assert independence through d-separation between variables:

$$dsep(X, Z, Y) \Leftrightarrow X \perp\!\!\!\perp Y|Z$$

The rules of d-separation are as follows:

```
; To determine if some set of nodes X is
; d-separated from some set of nodes Y by
; some (possibly empty) set of nodes Z:

trace ALL undirected paths from each X to each Y
    for each valve V on the current path
        if V is a fork and given (i.e., V∈Z)
            then this path is blocked
        if V is a chain and given
            then this path is blocked
        if V is a sink and neither it NOR its
          descendents are given
            then this path is blocked

if ALL paths blocked from each X to each Y given Z
    then X is d-separated from Y given Z
else there was an open path
    then X is NOT d-separated from Y given Z
```

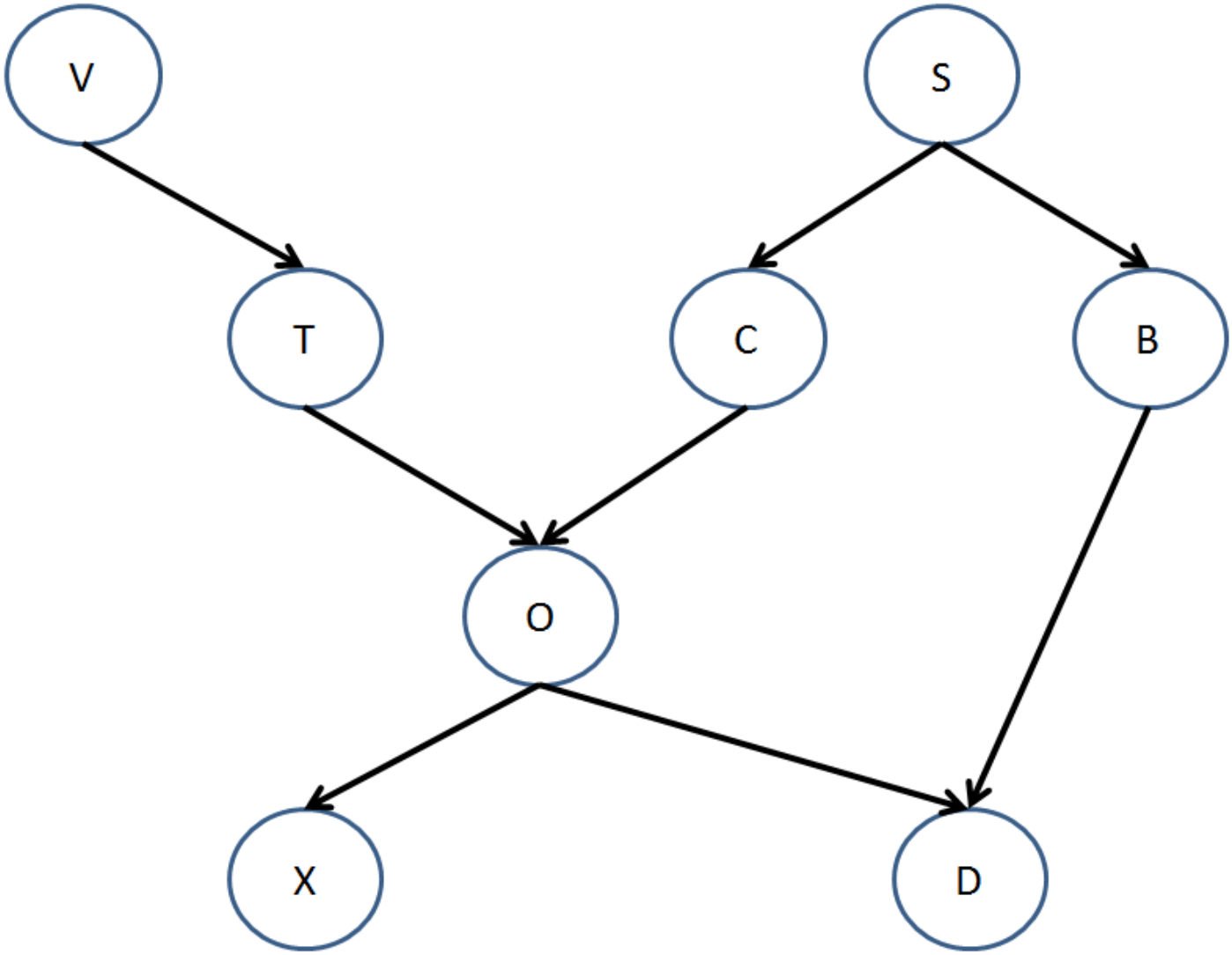
Here's that algorithm at-a-glance:

Valve Type	Example (Z is the valve)	Rule for being blocked
Fork	X ← Z → Y	Path is blocked if Z is **given**
Sequence	X → Z → Y	Path is blocked if Z is **given**
Collider	X → Z ← Y	Path is blocked if **NEITHER Z NOR ANY OF ITS DESCENDANTS** are given

It might look complicated, but the fact that it's a strictly graphical criterion for independence makes it easy to trace.

Let's do a bunch of examples:

Example
☑ Use the following Bayesian network to determine if each variable set X is separated from Y given Z. If it is not, show the open path.



Query	Answer

❓ $dsep(\{V\}, \{\}, \{T\})$	
❓ $dsep(\{V\}, \{T\}, \{O\})$	
❓ $dsep(\{T\}, \{O\}, \{S\})$	
❓ $dsep(\{T\}, \{D\}, \{S\})$	
❓ $dsep(\{T\}, \{\}, \{S\})$	
❓ $dsep(\{T, X\}, \{B\}, \{S\})$	

And now, one final, conceptual question:

❓ Are the independence relationships given by the Markovian assumptions implied by d-separation? Is the reverse true?

Inference

So now that we have our Bayesian networks defined and we know what independence relationships they claim... what do we do with them?

Well, we can ask them questions of course!

For our networks, those questions will be of the form: "What's the probability of witnessing events Q given that I've seen evidence e?" (where Q is a set of variables and e is an instantiation of evidence)

Inference queries, for query sentence Q and evidence e, compute the quantity $Pr(Q|e)$ using the network CPTs.

Every variable in the network can be classified under one of three categories while performing inference:

- **Query Variables (Q):** variables in which we are interested in computing the probability mass.
- **Evidence Variables (e):** events that we have witnessed that will serve as our conditioned variables.
- **"Hidden" Variables (V):** variables in the network that are neither query nor evidence variables, but may still propagate information through the network. Unlike evidence variables, we don't know the state of hidden variables, so we have to sum over all of their possible values.

A couple things to note about the inference problem:

- This problem would be pretty easy if we had our joint distribution... but we don't with Bayesian networks! In fact, the whole point of Bayesian networks is to avoid having to reconstruct our giant joint table. We just have CPTs.
- Accounting for evidence in terms of these CPTs is therefore less trivial because of the "information flow" we talked about in the last section, where conditioning on some evidence might actually OPEN flow from one variable to another.
- We might have a lot of superfluous variables that don't contribute to our query and need a means of getting rid of them before beginning inference.

Enumeration Inference is a strategy for computing an arbitrary Bayesian network query via a sum of products of CPT values from the network.

Let's consider how we might compute an arbitrary query on a Bayesian network using Enumeration Inference.

Here's what we want:

$$Pr(Q|e)$$

Now, we don't have (nor want to recreate) the joint distribution, but we do have a set of our network's conditional probability tables.

So, using Bayes' conditioning, we can phrase any query as:

$$Pr(Q|e) = \alpha \sum_{v \in V} Pr(Q, e, v)$$

Above, α is simply a normalizing constant that is equivalent to 1 / Pr(e).

Finally, we know that the joint that we're summing over can be factorized into the Markovian family factors for all nodes X:

$$Pr(Q, e, V) = \Pi_{x \in X} Pr(x|Parents(x))$$

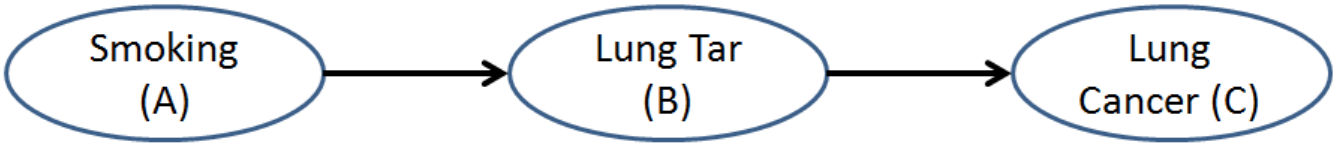
Example

Use the following Bayesian network and its CPTs to compute:

$$Pr(Cancer = 1|Smoking = 1)$$

A	Pr(A)
0	0.4
1	0.6

B	C	Pr(C | B)
0	0	0.5
0	1	0.5
1	0	0.7
1	1	0.3



A	B	Pr(B | A)
0	0	0.8
0	1	0.2
1	0	0.1
1	1	0.9

What are our query, evidence, and hidden variables above?

Provide the query statement in terms of a sum of products over our network CPTs.

Now, notice that above expression (the answer to the question above) says to "Sum over all of the possible values of B and multiply the family factors."

This works, but we could improve it slightly, because the first factor Pr(A = 1) does not rely on the summation over B (see how B isn't mentioned in Pr(A = 1)?).

So, we can simply move it outside of the summation to simplify, giving us:

$$\begin{aligned} Pr(C=1|A=1) &= \alpha * Pr(A=1) \sum_{b \in B} Pr(b|A=1) * Pr(C=1|b) \\ &= \alpha * Pr(A=1)[Pr(B=0|A=1) * Pr(C=1|B=0) + Pr(B=1|A=1) * Pr(C=1|B=1)] \\ &= \alpha * 0.6 * [0.1 * 0.5 + 0.9 * 0.3] \\ &= \alpha * 0.192 \end{aligned}$$

So close to finishing our computation! Now we just need to find a value for

$$\alpha = 1/Pr(e) = 1/Pr(A=1)$$

Easy enough! We happen to know this from our CPT for A. This gives us:

$$\begin{aligned} Pr(C=1|A=1) &= \alpha * 0.192 \\ &= \frac{0.192}{Pr(A=1)} \\ &= \frac{0.192}{0.6} \\ &= 0.32 \end{aligned}$$

And there you have it! So in summary, the steps of enumeration inference are:

1. Identify your query (Q), evidence (e), and hidden variables (V) amongst all variables in the network (X).
2. Formulate your query expression in terms of the network CPTs, by the formula:

$$Pr(Q|e) = \alpha \sum_{v \in V} Pr(Q, e, v)$$

3. [Optional] (But time saving) Re-arrange summations and factors to reduce number of multiplications (like when we pulled out Pr(A) above)

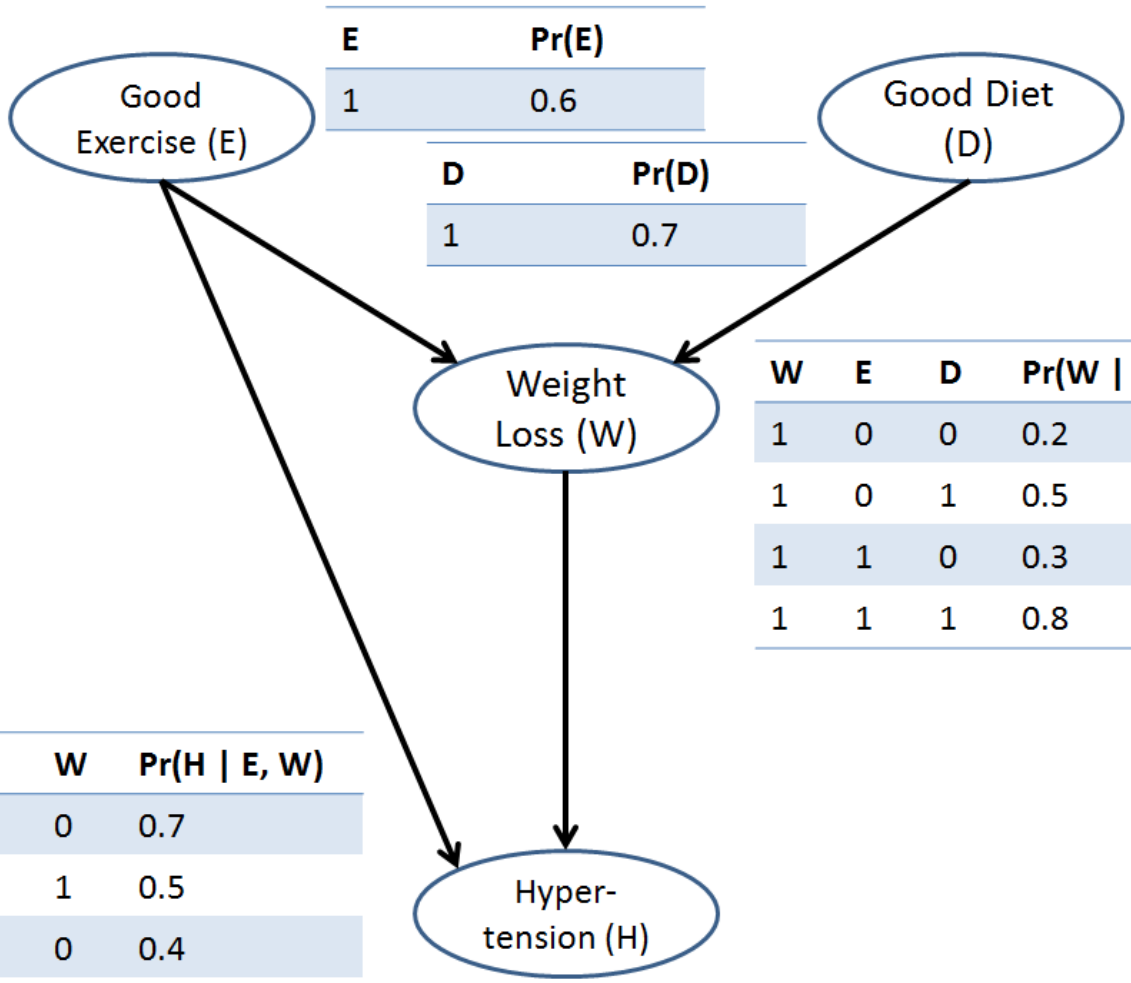
4. Solve for, and substitute into (2) above:

$$\alpha = 1/Pr(e)$$

Additional Practice

Example

Use the following Bayesian network to answer the queries that follow. Assume that each CPT denotes 0 as an absence of the variable and 1 as the presence of it (e.g. H = 1 means that you have hypertension). Use Enumeration Inference as described above.



E	Pr(E)
1	0.6

D	Pr(D)
1	0.7

W	E	D	Pr(W | E, D)
1	0	0	0.2
1	0	1	0.5
1	1	0	0.3
1	1	1	0.8

H	E	W	Pr(H | E, W)
1	0	0	0.7
1	0	1	0.5
1	1	0	0.4
1	1	1	0.1

What is the probability that an individual has hypertension given that they exercise? [Click for solution]

Homework 4

Your last grueling exercise set! Let's go over it now and look at some hints...

Problem	Tip
1. FR-TO-ENG	Deceptively intricate, but not impractical; consider breaking your solution down into several helper functions: • One that finds the pattern associated with the current top-level predicate. • One that performs the translation of the found pattern(s). • One that glues the replacement-patterns of a pattern with a decision tree together.
2. EVAL-D-TREE	Trivial. Big hint: remember what the return value is for Clisp's logical AND!

Hope those tips help a bit! I think you'll find this assignment a vacation compared to the previous homeworks.