Applying Bayes' Theorem

Last week we just saw the tip of the iceberg as far as what Bayes' Theorem can accomplish for us.

We've learned the probabilistic mechanisms that we'll use in the following lecture, so let's see how it all fits together.

Shall we start with a motivating example? Don't feel bad if you don't get this at first, about 85% of medical doctors asked the same question get it wrong.

Example

t Read the following problem description, and for each probability mentioned, formalize it into a Pr(x | y) statement (for example) and then solve for the correct quantity. Assume two binary variables: Test and Disease, as described by the problem.

> A very rare condition, Schistosoforneymiosis, is found in about 1/1000 of those tested for it; sufferers experience a consistently wet left foot and sentient freckles.
>
> The test is an elaborate procedure involving multiple probes, and returns an end result that is either positive (Test) or negative (¬Test). The problem is that the tests are not perfect, but 95% of people who have the disease will test positive, and 2% of people who do *not* have the disease will test positive.
>
> **If a patient tests positive, what is the probability that they have the disease?**

Let's dissect this problem and determine the proper quantities for each component to lead to our solution.

à Translate the sentence into a Pr statement: "A very rare condition, Schistosoforneymiosis, is found in about 1/1000 of those tested for it."

The above represents the **prior** probability that we discussed from last time, i.e., the chance that someone has the condition before accounting for any evidence.

à Translate the sentence into a Pr statement: "95% of people who have the disease will test positive."

This value represents the case of a **true positive** since the test is positive and so is the disease.

à Translate the sentence into a Pr statement: "2% of people who do *not* have the disease will test positive."

This value represents the case of a **false positive** since the test is positive but the disease is not.

à Translate the sentence into a Pr statement: "If a patient tests positive, what is the probability that they have the disease?"

Alright, off to a good start! Let's write out everything that we have so far:

```
Pr(Disease) = 0.001
Pr(Test | Disease) = 0.95
Pr(Test | ¬Disease) = 0.02

Pr(Disease | Test) = ???
```

So, if it wasn't obvious before, we need to use Bayes' Theorem in order to solve for our target quantity! Let's write out what we'll need for that:

```
Pr(Disease | Test) = (Pr(Test | Disease) * Pr(Disease)) / Pr(Test)
```

Do we have everything that we need to solve for Pr(Disease | Test)?

Well, yes and no; we have everything we need to calculate what we need, explicitly. In particular, we have everything on the RHS but Pr(Test).

Do we have a means of calculating the Pr(Test)?

ὰ Click here if you need a hint on how to calculate Pr(Test)...

"But Andrew, we don't have things in terms of Pr(α, β), only **conditions**!"

You're correct! You even said the strategy we need to use! Let's take a look:

```
; I claim that the choice for β is a partition
; in which each β_i is mutually exclusive, do you agree?
β = {{Disease}, {¬Disease}}

; So, by the Law of Total Probability:
Pr(α) = Σ_i Pr(α, β_i)

Pr(Test) = Σ_i Pr(Test, β_i)
         = Pr(Test, Disease) + Pr(Test, ¬Disease)

; Now, we can condition in order to get:
Pr(Test) = Pr(Test | Disease) * Pr(Disease) +
           Pr(Test | ¬Disease) * Pr(¬Disease)

; We do not *explicitly* have one of these quantities above,
; BUT, obvserve:
Pr(¬Disease) = 1 - Pr(Disease)
             = 1 - 0.001
             = 0.999

; Therefore:
Pr(Test) = 0.95 * 0.001 + 0.02 * 0.999
         = 0.02093

; And now, plugging back into our original:
Pr(Disease | Test) = (Pr(Test | Disease) * Pr(Disease)) / Pr(Test)
                   = (0.95 * 0.001) / 0.02093
                   = 0.045
```

Wow! That's a really small chance given that our incredibly accurate test was still positive!

Understanding Bayes' Theorem gives us a lot of power to detect non-obvious consequences like the above.

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	0.2
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.05
1	0	1	0.05
1	1	0	0.2
1	1	1	0.2

Well that's... err... pretty perfect... such neat probabilities! (we'd never get something so clean in the real world, of course, but play along for now...)

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

ά We can **sum out (marginalize)** a variable by "collapsing" the distribution on the remaining variables we're not summing out. This is formally defined as:
$P(Y) = \Sigma_{z \in Z} P(Y, z)$

Less formally, for example, if we wanted to sum out X in the above distribution, leaving a joint **marginal** on Y and Z, we simply look for every row where Y and Z *agree* and then sum over the possible values for X

Let's look at that in action; take a look at the following two pairs of rows, color-coded for your convenience:

X	**Y**	**Z**	**Pr(X, Y, Z)**
0	0	0	0.2
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.05
1	0	1	0.05
1	1	0	0.2
1	1	1	0.2

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

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

So, to sum out X, 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 Y and Z.

ά The distribution that results from summing out a variable from a joint is called a **joint marginal** distribution, to illustrate that a variable was summed out of the original joint.

This amounts to the following joint marginal on Y and Z with X summed out (typically written: `Σ_X Pr(Y, Z)`):

Y	**Z**	**Σ_X Pr(Y, Z)**

0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

(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 Y is independent from Z, written:
; Y ⊥ Z
; ...then:

Pr(Y | Z) = Pr(Y)

; ...∀ y, z
```

So let's assess that above (if it isn't obvious)...

Using Bayes' Conditioning, we see that:

```
Pr(Y | Z) = Pr(Y, Z) / Pr(Z)

; Using our new joint on Y and Z, it is
; easy to see that:
Pr(Y = 1, Z = 1) = 0.25
Pr(Y = 1, Z = 0) = 0.25
Pr(Z = 1) = Pr(Z = 0) = 0.5

; ...so:
Pr(Y = 1 | Z = 1) = 0.25 / 0.5
                  = 0.5
                  = Pr(Y = 1)

; ...and we can also show that:
Pr(Y = 1 | Z = 0) = Pr(Y = 1 | Z = 1)
                  = Pr(Y = 1)

∴ Y ⊥ Z
```

Neat! So this means that knowing something about Z tells me nothing about the state of Y...

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 Z from the original joint distribution to get:

X	Y	Σ_Z Pr(X, Y)
0	0	0.4
0	1	0.1
1	0	0.1
1	1	0.4

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

```
Pr(Y | X) = Pr(Y, X) / Pr(X)

Pr(Y = 1 | X = 1) = 0.8
                  ≠ Pr(Y = 1)

∴ Y is dependent on X
```

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 ⊥ 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 ⊥ 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 ⊥ B
Independent, but	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	A /⊥ B | S (that's meant to be a little line through the ⊥

Conditionally Dependent	actually *does* tell me something about the other, even though A and B are independent on their own.	symbol but I can't find that in unicode so... use your imagination)
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 ⊥ B A ⊥ B | R

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

$\mathbf{Pr}(X \mid Y, Z) = \mathbf{Pr}(X \mid Z)$ if $X \perp Y \mid Z$

$\mathbf{Pr}(X, Y \mid Z) = \mathbf{Pr}(X \mid Z) * \mathbf{Pr}(Y \mid Z)$ if $X \perp Y \mid Z$

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

In the book, the example given is for observing relationships between three dental-exam variables:

- **Toothache:** subjective patient reported tooth pain; typically a sign of a cavity
- **Cavity:** whether the dentist found a cavity
- **Catch:** whether the dentist felt their metal teeth cleaner get caught on a tooth of interest; typically a sign of a cavity

...an example, which, frankly sounds a bit odd, but we'll run with it because it's the numbers we're interested in.

Let's look at that distribution:

Toothache	**Cavity**	**Catch**	**Pr(Toothache, Cavity, Catch)**
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 toothache are usually indicative of a cavity, which is also a cause of catches.

Now, the example makes the following observations (by example design, of course):

- We hypothesize that cavities cause toothaches and catches.
- If this is the case, then knowing that someone has a cavity immediately tells us that they are likely to have a toothache and catch.
- Furthermore, if we knew someone had a cavity, does knowing that they also have a toothache tell us more than we already know? No!

> à What conditional independence relationship are the above observations making?

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

> à Express this conditional independence in probability notation; 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(Toothache | Catch, Cavity) = Pr(Toothache | Cavity)

; We have neither of those tables, but we can compute them!
; Let's use Bayes' Conditioning and marginalization!
  Pr(Toothache | Catch, Cavity)
= Pr(Toothache, Catch, Cavity) / Pr(Catch, Cavity)

  Pr(Toothache | Cavity)
= Pr(Toothache, Cavity) / Pr(Cavity)
```

Marginalizing from the joint distribution, we can get our two joint-marginals on Pr(Catch, Cavity) and Pr(Toothache, Cavity):

Cavity	Catch	Pr(Cavity, Catch)
0	0	0.64
0	1	0.16
1	0	0.02
1	1	0.18

Toothache	Cavity	Pr(Toothache, Cavity)
0	0	0.72
0	1	0.08
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(Cavity) = 0.2

Now, we just need to compute the two conditional distributions; we'll start with the Pr(Toothache | Catch, Cavity) from the joint:

Toothache	Cavity	Catch	Pr(Toothache, Cavity, Catch)	Pr(Toothache | Cavity, Catch)
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(Toothache | Cavity):

Toothache	Cavity	Pr(Toothache | Cavity)
0	0	0.9
0	1	0.4
1	0	0.1
1	1	0.6

Aha! We see that the two are in fact equivalent, regardless of what we know about a Catch! Taking a closer look:

```
  Pr(¬Toothache | ¬Catch, ¬Cavity)
= Pr(¬Toothache | Catch, ¬Cavity)
= Pr(¬Toothache | ¬Cavity)
= 0.9

  Pr(¬Toothache | ¬Catch, Cavity)
= Pr(¬Toothache | Catch, Cavity)
= Pr(¬Toothache | Cavity)
= 0.4

  Pr(Toothache | ¬Catch, ¬Cavity)
= Pr(Toothache | Catch, ¬Cavity)
= Pr(Toothache | ¬Cavity)
= 0.1

  Pr(Toothache | ¬Catch, Cavity)
= Pr(Toothache | Catch, Cavity)
= Pr(Toothache | Cavity)
= 0.6

∴ Pr(Toothache | Catch, Cavity) = Pr(Toothache | Cavity)
∴ Toothache ⊥ Catch | Cavity
```

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 **`Toothache`** $\perp$ **`Catch`** | **`Cavity`** from our dentist example.

Let's look at an interesting consequence:

> à Give the chain-rule factorization of the joint probability table: **`Pr(Toothache, Catch, Cavity)`**

OK, and based on our conditional independence relation found from the previous section, can we reduce this to anything simpler?

```
; Yes! Since:
Toothache ⊥ Catch | Cavity

; Then:
  Pr(Toothache | Catch, Cavity)
= Pr(Toothache | Cavity)

; And so our factorization:
  Pr(Toothache, Catch, Cavity)
= Pr(Toothache | Catch, Cavity) * Pr(Catch | Cavity) * Pr(Cavity)
= Pr(Toothache | Cavity) * Pr(Catch | Cavity) * Pr(Cavity)
```

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

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

Remembering that we presumed that both toothaches and catches were indicators of cavities:

à Thinking about cause-effect relationships, what's interesting about the tables: Pr(Toothache | Cavity) and Pr(Catch | Cavity)

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

Example

t Using the following distribution properties, compare the full joint probability distribution's size (over all 5 variables) to its chain-rule factorization:

```
; Say we have 5 binary variables:
A, B, C, D, E

; They elicit a joint probability distrbution:
Pr(A, B, C, D, E)

; [?] How many rows are in this table?

; Now, say we had independence relationships
; that allowed us to factor that joint distribution
; into the following:
  Pr(A, B, C, D, E)
= Pr(A) * Pr(B) * Pr(C) * Pr(D) * Pr(E)

; [?] How many rows are in each of these smaller
; tables?

; [?] Do we witness a difference between the number
; of rows in the full joint vs. the factorization?
```

Yes! Big savings, in fact!

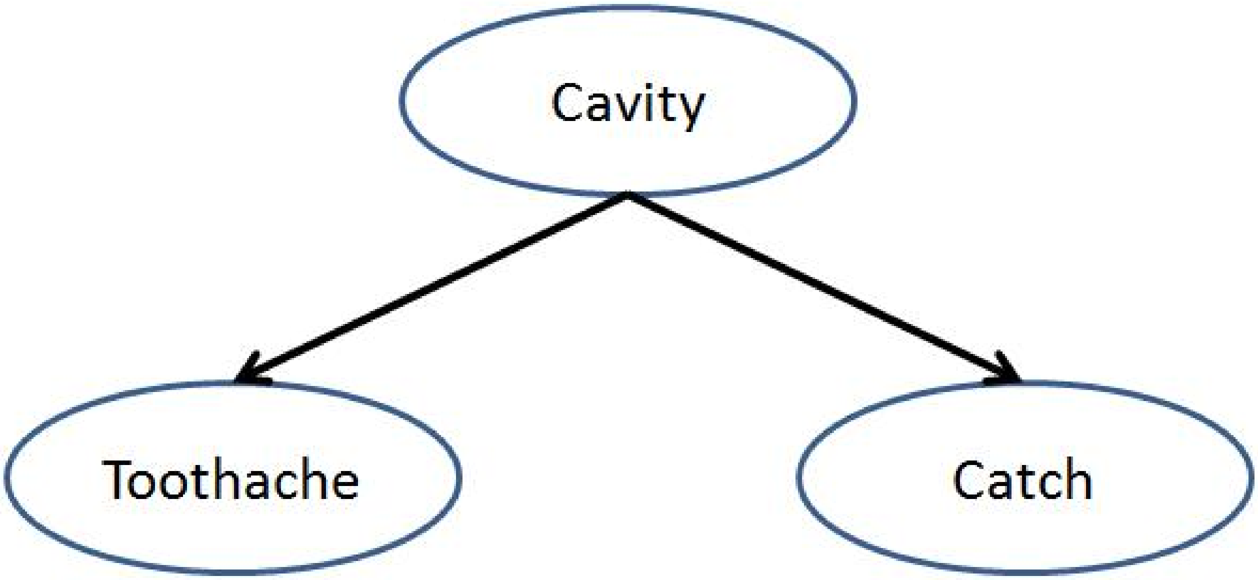
There are 32 rows in the full joint, but only 10 rows in the individual, factored tables!

> ά **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 dentist's problem:



There's a lot to note about Bayesian networks:

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

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

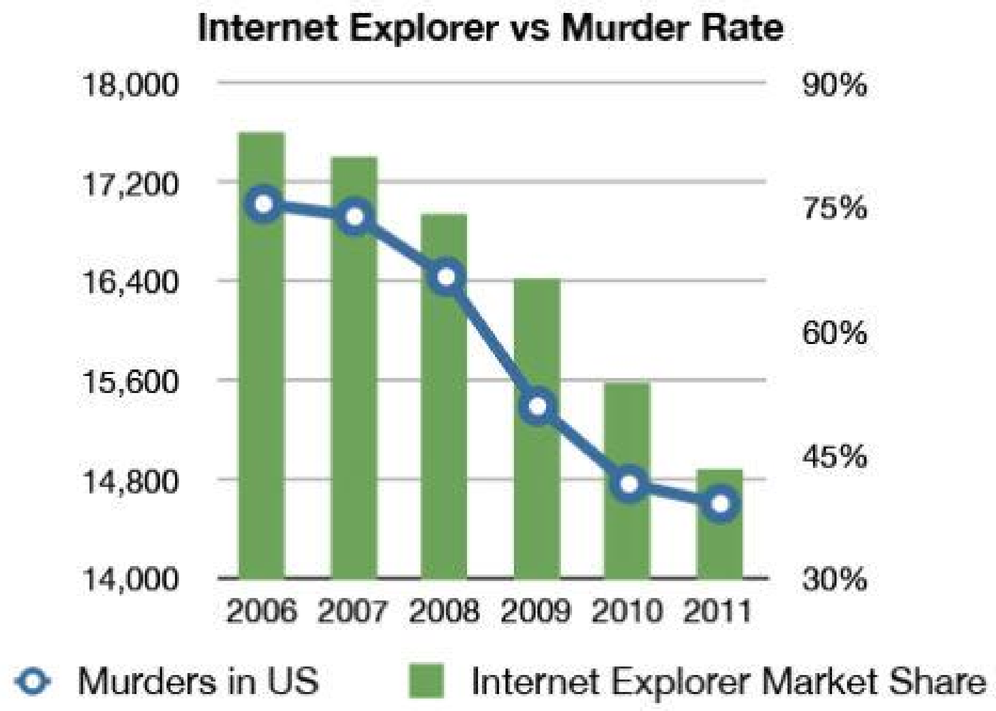
ά The network's **edges** represent **dependence** relationships, i.e., two connected nodes are dependent upon one another.

ά 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).

ά 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 dentist example: if we know that someone has a cavity (the cause), then knowing that they also have a toothache doesn't tell us more about their chance of having a catch.

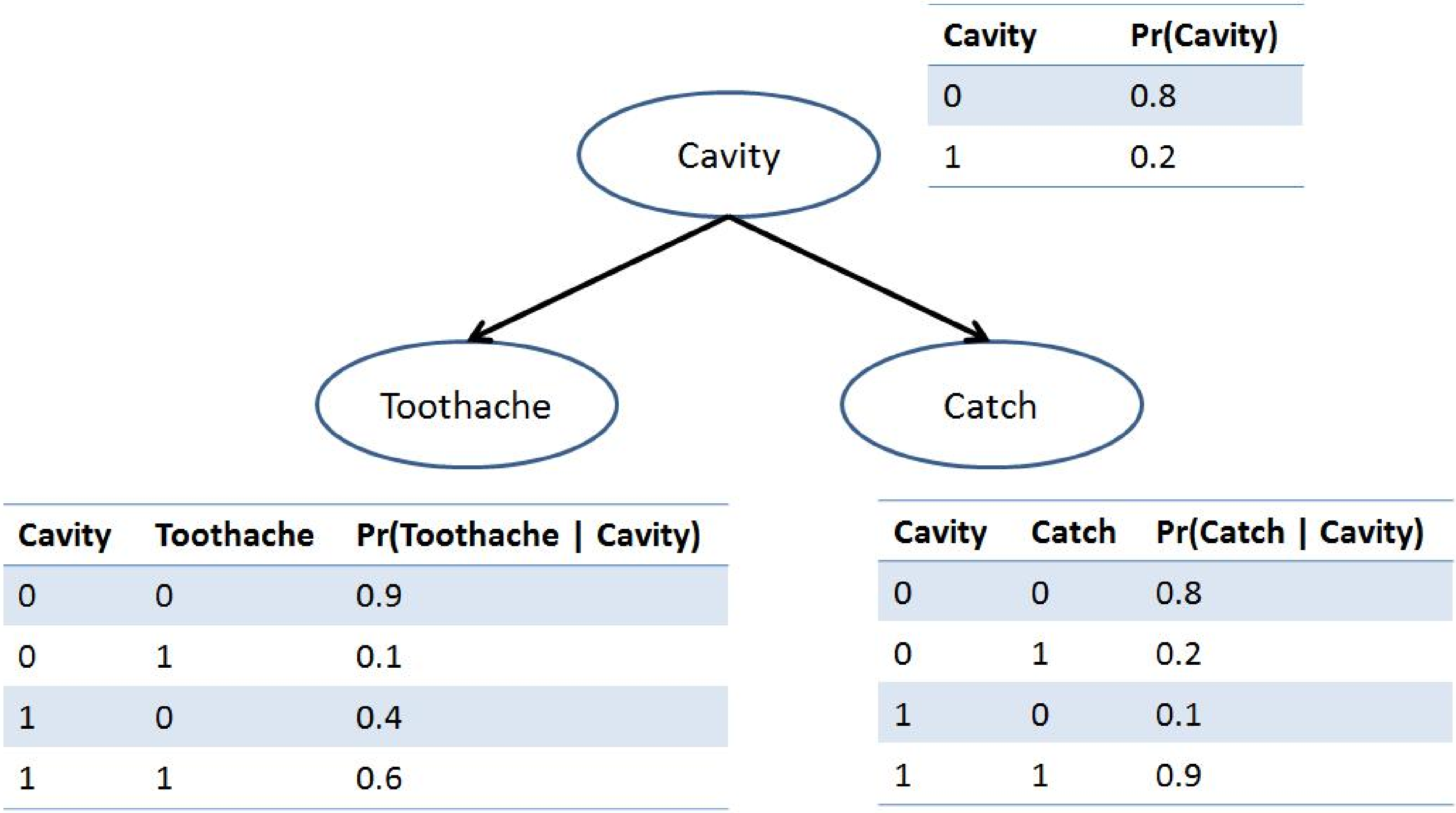
That is, as soon as we know the state of the causes, we don't get any new information about the effects!

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

```
; The joint distribution:
  Pr(X1, X2, ..., Xn)

; ...can be factored using independence
; relationships to:
= Pr(X1 | Parents(X1)) * Pr(X2 | Parents(X2)) * ...

; ...which semantically represents the putative
; relationship between causes and effects:
= Pr(Effect1 | Causes(Effect1)) * (Effect2 | Causes(Effect2)) * ...

= Π_X∈Vars Pr(X | Parents(X))
```

á 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 dentist example, the CPTs would look like:



Cavity	Pr(Cavity)
0	0.8
1	0.2

Cavity	Toothache	Pr(Toothache | Cavity)
0	0	0.9
0	1	0.1
1	0	0.4
1	1	0.6

Cavity	Catch	Pr(Catch | Cavity)
0	0	0.8
0	1	0.2
1	0	0.1
1	1	0.9

à Using the above CPTs, compute: Pr(Cavity = 0, Toothache = 1, Catch = 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:

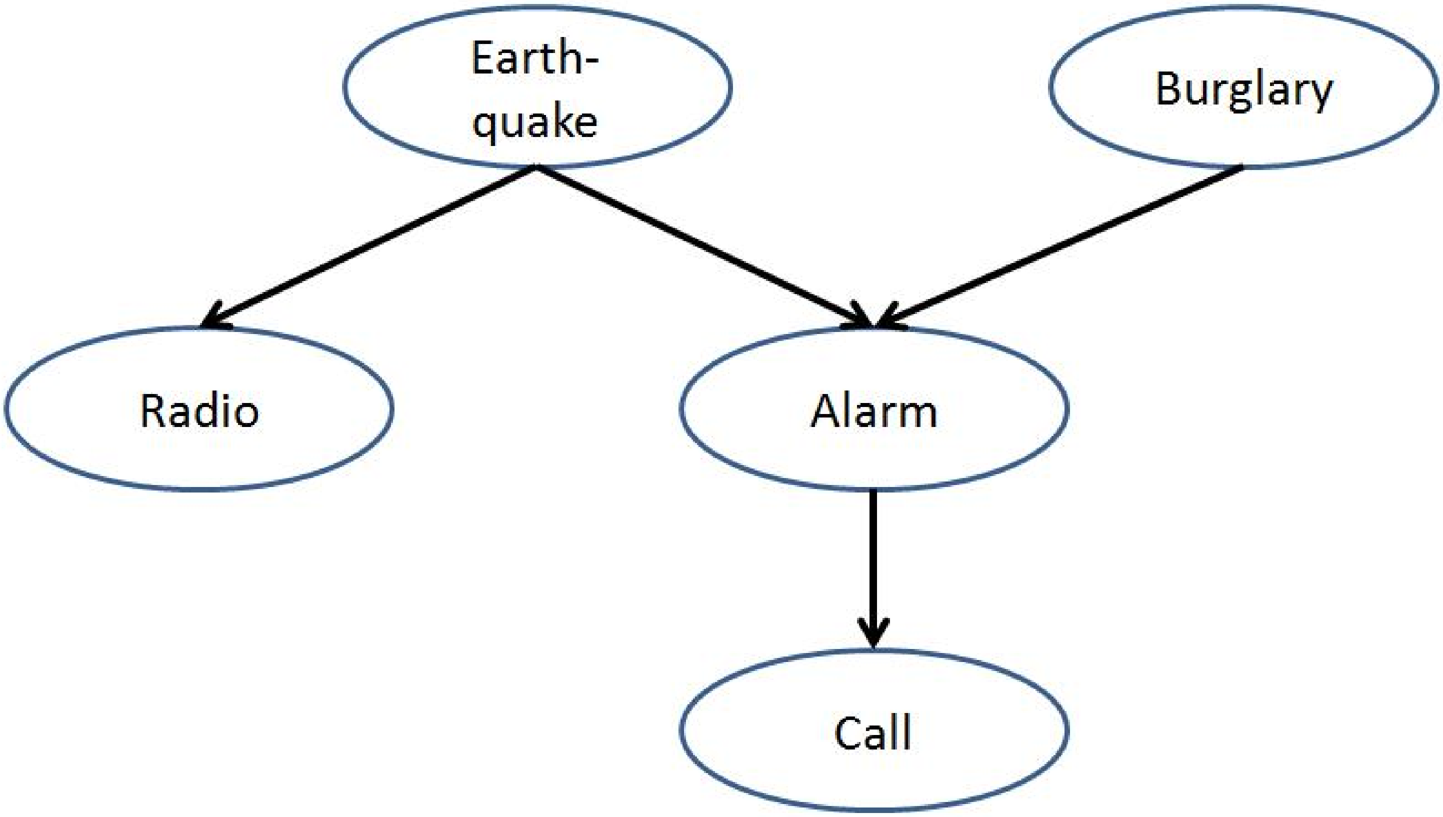
á The **Markovian assumption** states that every node is independent of its non-descendents given its parents.

(non-descendents of node X are defined as any node that you could not reach taking a directed path starting at X)

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

Example

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



à Click for answer.

We notice a couple things about our Markovian assumptions:

1. There are some redundancies (e.g., Burglary ⊥ Radio, but we also have that Radio ⊥ Burglary | Earthquake)
2. There are some independence relations that they miss...

Let's look at a formalization of independences implicit within our Bayesian Network next.

D-Separation

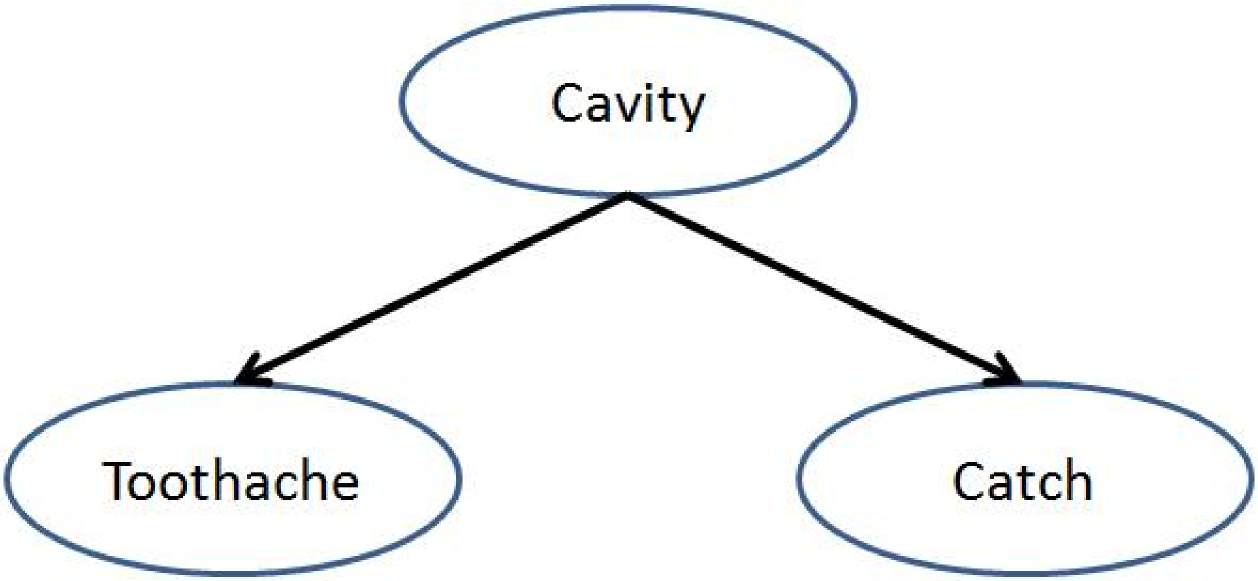
Let's start off by considering 3 different simple Bayesian networks and observe how we can generalize their characteristics.

We'll examine 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:

> ά A valve Z is **divergent / a fork** along some path if it is a common cause of two effects, X and Y.
> X ← Z → Y

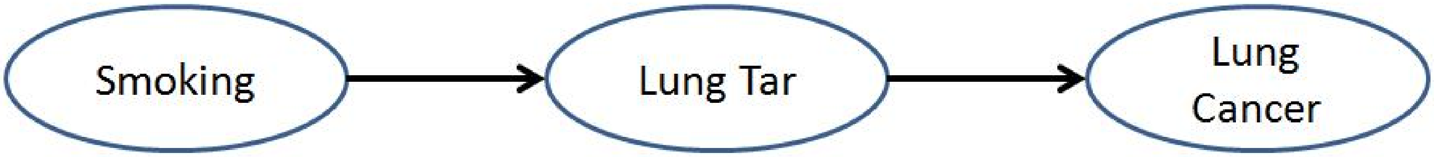


Divergent nodes are sometimes referred to as common causes. Here, we see that if we know whether or not someone has a cavity, then information does NOT flow from toothache to catch.

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

ά A valve Z is **sequential / a chain** along some path if some other variable X is its cause and Z has some effect Y.
$X \rightarrow Z \rightarrow Y$ OR $Y \leftarrow Z \leftarrow X$

Here's an example path of a chain:



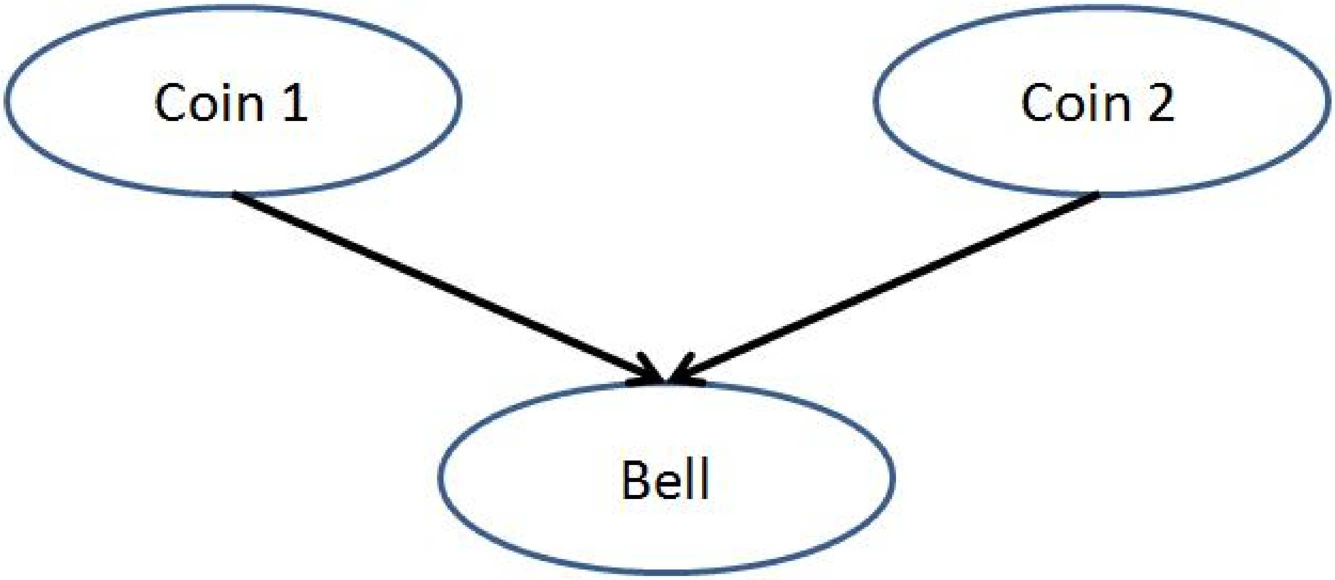
Here, knowing that someone has tar in their lungs means that we no longer get information flowing from any knowledge that the person smoked to whether they'll have lung cancer.

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.

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

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).

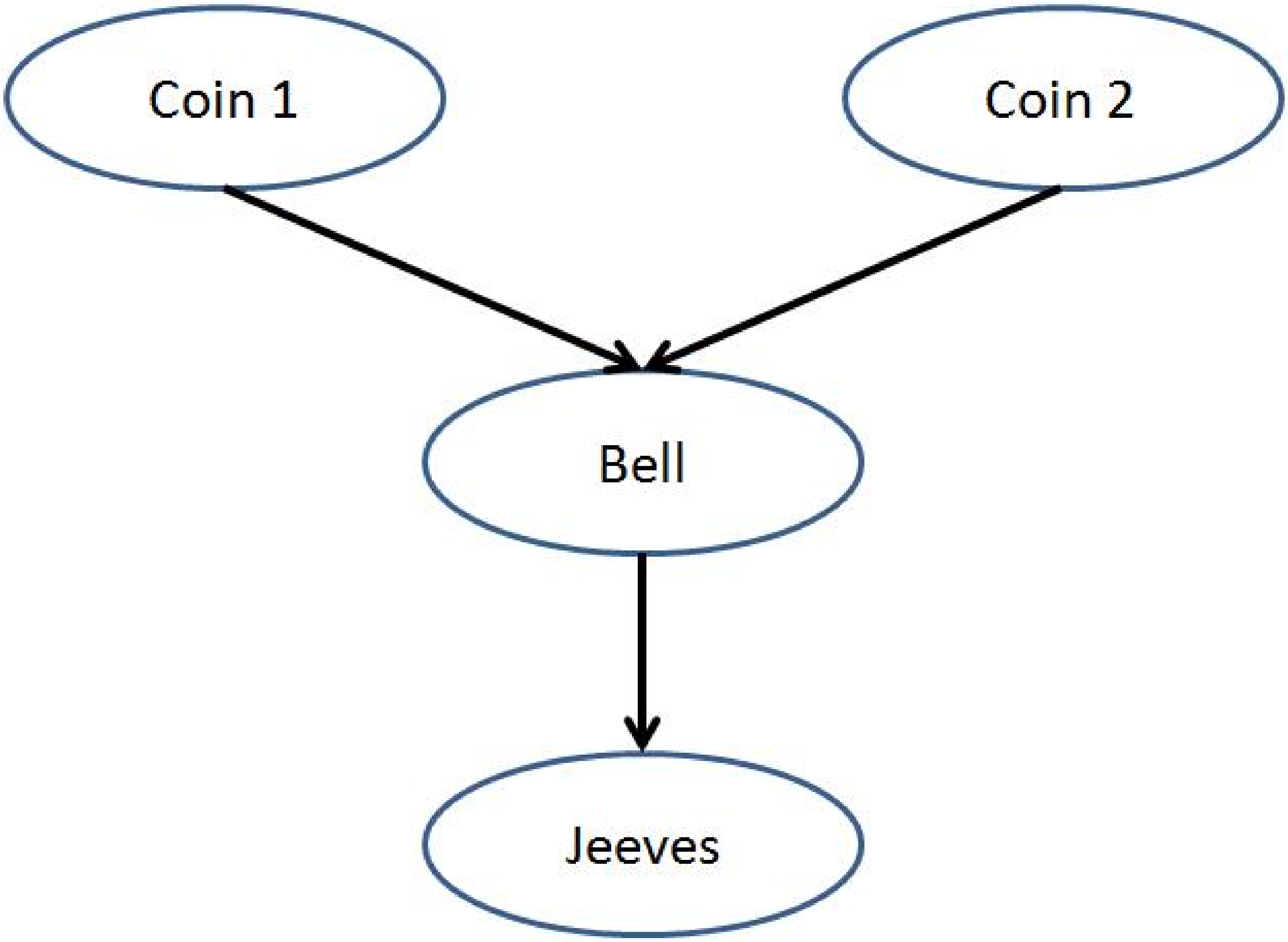


à If I know whether or not the bell rings, does information flow from coinflip 1 to coinflip 2?

à If I DO NOT know whether or not the bell rang, does information flow from 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.



à If I know whether or not Jeeves brings tea, does information flow from coinflip 1 to coinflip 2?

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

So, the rule for convergent valves is the following: for common effect Z in configuration: X → Z ← 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.

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

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
```

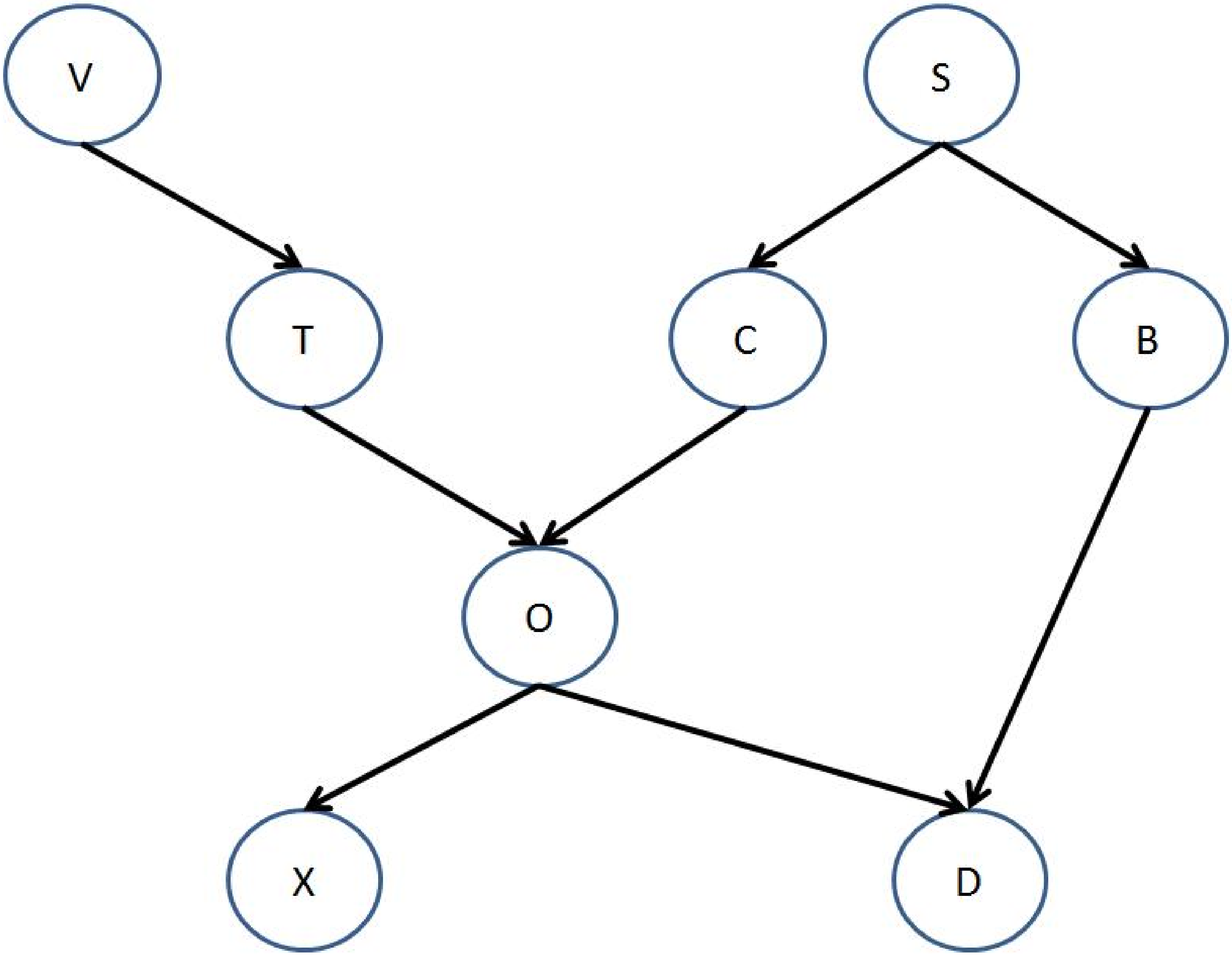
It might look complicated, but the fact that it's a strictly graphical criterion for independence makes it easy to trace.

á We use the following notation to establish d-separation relationships: `d-sep(X, Z, Y) ≡ X ⊥ Y | Z`

Let's do a bunch of examples:

Example

t 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.



à Is d-sep({V}, {}, {T})

à Is d-sep({V}, {T}, {O})

à Is d-sep({T}, {O}, {S})

à Is `d-sep( {T}, {D}, {S} )`

à Is `d-sep( {T}, {}, {S} )`

à Is `d-sep( {T, X}, {B}, {S} )`

And now, one final, conceptual question:

à 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)

This problem would be pretty easy if we had our joint distribution... but we don't with Bayesian networks! We just have CPTs.

Furthermore, we want to avoid reconstructing the joint because it might be massive!

á **Variable elimination** is an inference strategy designed to use only our CPTs to pose queries on our network.

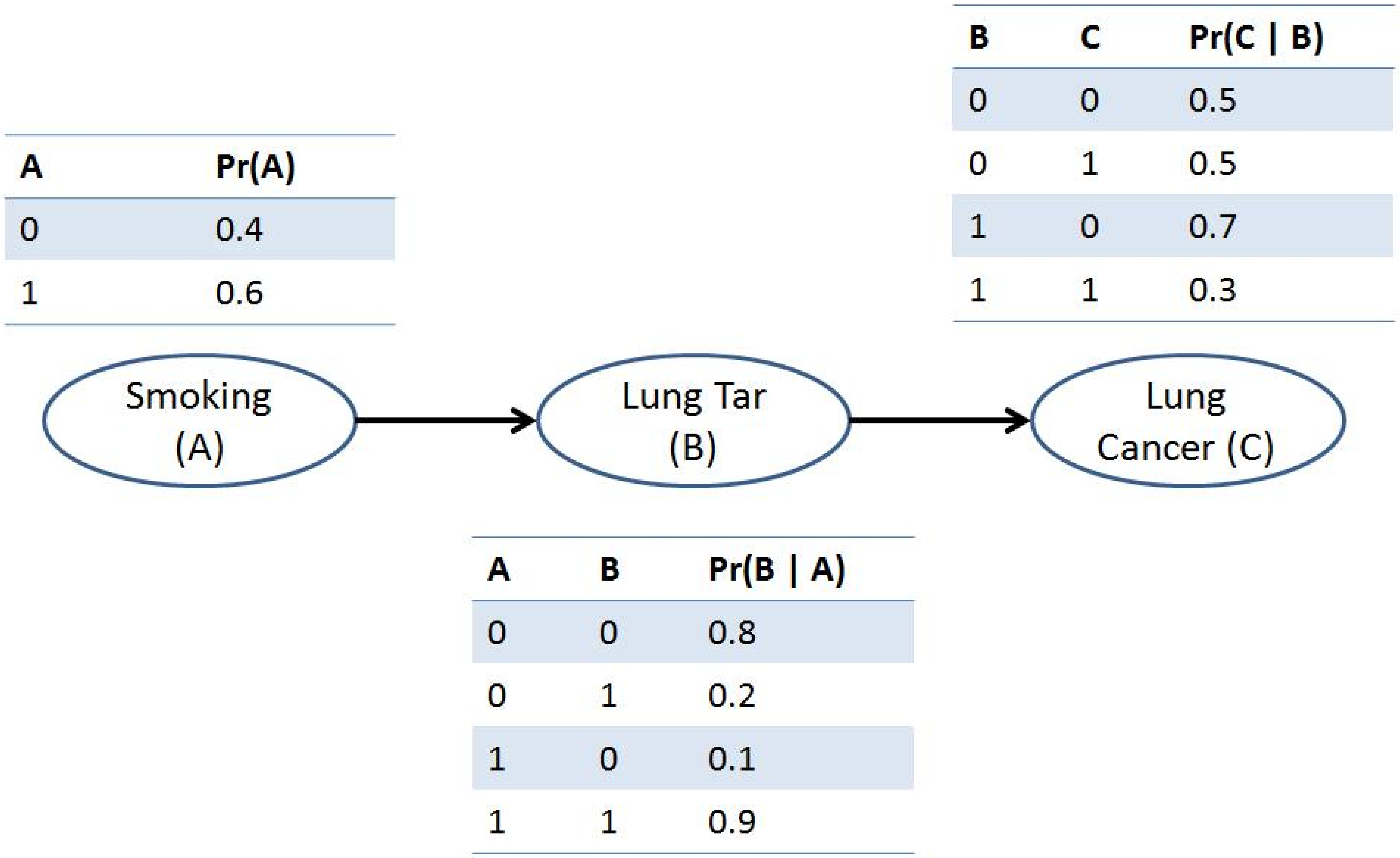
ň WARNING: The topic of Bayesian network inference is massive (there's a whole class on it, 262a, which I genuinely suggest taking!) and so this section is very hand-wavy.

The idea of variable elimination is that, with some query and evidence variables, we want to find Pr(Q | e).

For example, we might be interested in Pr(Cancer | Smoking = true)

Above, Q = {Cancer} and e = {Smoking = true}

Let's use this as our example! (example credit to Dr. Darwiche, and my notebook for surviving this long to be useful)



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

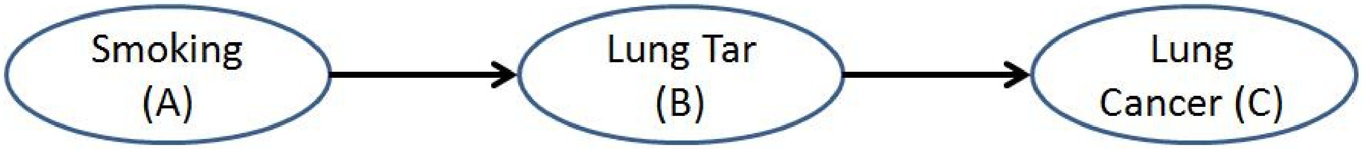
A	Pr(A)
0	0.4
1	0.6

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

Step One: Zero out CPT rows that are inconsistent with the evidence (these inconsistent rows are often just omitted entirely)

A	Pr(A)
0	0
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
0	1	0
1	0	0.1
1	1	0.9

Step Two: Choose an "elimination order"

Our goal is to get ONE CPT at the end that only mentions variables in Q; we do this by successively multiplying table values and summing out irrelevant variables.

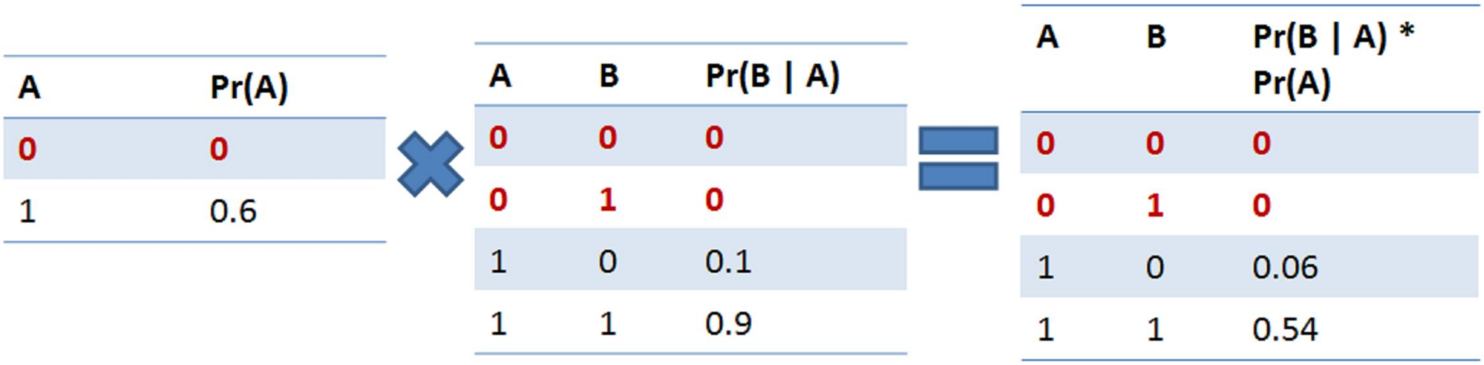
The order in which you do this matters greatly for computational complexity, but that discussion is omitted herein; for now, let's just choose an order:

```
; Π is used to designate an ordering
Π = {A, B, C}

; Meaning A first, then B, etc.
```

Since A is up first, this means we multiply all the CPTs mentioning A and get back a **factor**, which is simply some function mapping variables to some positive numbers.

Factors are NOT distributions, but simply the intermediary computational elements of variable elimination.

So, the two CPTs mentioning A in our example are: Pr(A) and Pr(B | A) so:



A	Pr(A)
0	0
1	0.6

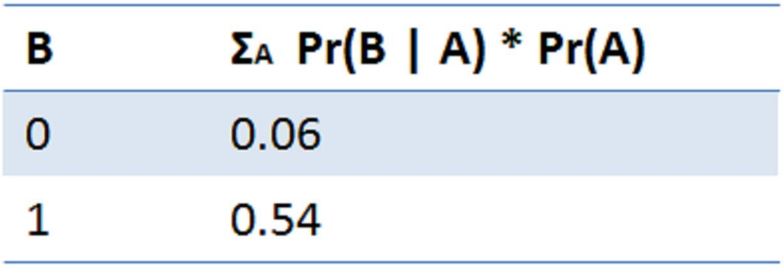
×

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

=

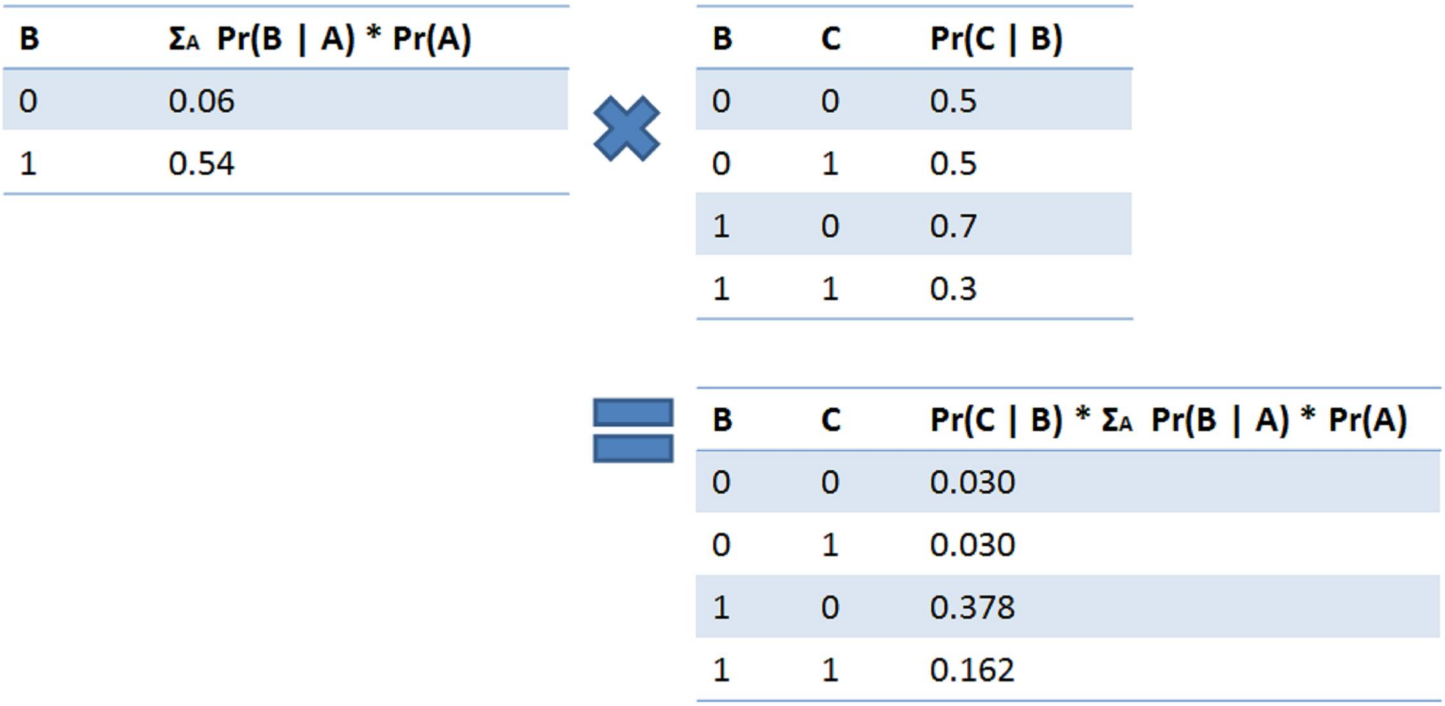
A	B	Pr(B | A) * Pr(A)
0	0	0
0	1	0
1	0	0.06
1	1	0.54

Since we've now dealt with all of the tables mentioning A, we can **eliminate** it by summing out, giving us:



B	Σ_A Pr(B | A) * Pr(A)
0	0.06
1	0.54

Our next variable to eliminate is B, which is mentioned in both our current factor (Σ_A Pr(B | A) * Pr(A)) and Pr(C | B), so do the same thing there!



B	Σ_A Pr(B | A) * Pr(A)
0	0.06
1	0.54

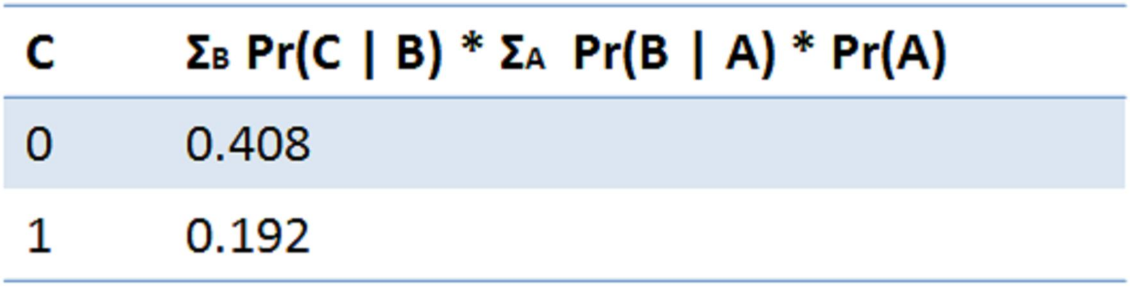
×

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

=

B	C	Pr(C | B) * Σ_A Pr(B | A) * Pr(A)
0	0	0.030
0	1	0.030
1	0	0.378
1	1	0.162

Almost there! But now B is irrelevant, so we can sum it out to get:



C	Σ_B Pr(C | B) * Σ_A Pr(B | A) * Pr(A)
0	0.408
1	0.192

So close! But let's look at what we have so far:

```
; The term currently in our table:
Σ_B Pr(C | B) * Σ_A Pr(B | A) * Pr(A)

; ...is really just the joint with B and A summed out!
; (after accounting for evidence A = true, of course)

= Σ_B Σ_A Pr(C | B) * Pr(B | A) * Pr(A)
= Σ_B Σ_A Pr(A, B, C)
```

So, since our final factor has only C remaining, and accounted for A = true, the values it represents are a marginal on C and A = true!

```
; So, from our final factor:
Pr(C = true, A = true) = 0.192
Pr(C = false, A = true) = 0.408

∴ Pr(A = true) = 0.600

; Now we have everything we need to compute
; our query:
; Pr(Q | e) = Pr(C | A = true)

; ...using Bayes conditioning:
Pr(C = true | A = true) = Pr(C = true, A = true) / Pr(A = true)
                        = 0.192 / 0.600
                        = 0.32

Pr(C = false | A = true) = Pr(C = false, A = true) / Pr(A = true)
                         = 0.408 / 0.600
                         = 0.68
```

And that's that!

Like I say, there's a whole class on Bayesian network inference, so this is but a small example and one strategy that can be used for such.

Thus concludes our brief tour of Bayesian networks... hope you enjoyed the ride!

Homework 4

Now that I've had a chance to look through the homework and do some solutions, I have a few hints that might be helpful in crafting your own!

Problem	Tip
1. SF-UNIFY	Remember, to watch out for ordering in frames and lists of frames! The book algorithm attempts to unify slots in frames, and frames in lists, *in the same order*, but our frames don't care about that!

	As such, you might find it useful to have multiple helper functions to handle the cases mentioned by the book, accounting for frames that might successfully unify but that simply aren't ordered the same.
2. SF-SUBST	As you might have already seen, the solution to this one looks a lot like a couple of other functions we've done in the past that have to replace fillers with other values...
3. SF-INFER	Fairly straightforward once you have SF-UNIFY working, since it does most of the heavy lifting...
4. FORWARD1	Be careful to avoid infinite loops with the facts that you learn! Hmm, how to screen the results of SF-INFER?
5. BACKWARD1	Now much easier since we're only doing the first step of backward chaining, this one too is reduced to a search problem on the first consequent that successfully unifies to the query.
6. C-GEN	Deceptively intricate; make sure you read through the examples and piece together a plan of attack. I suggest dividing the algorithm into several steps: (1) isoltaing the pattern to use, (2) picking apart each pattern component, and (3) recursing as necessary.
7. C-GENS	Trivial once Problem 6 is solved.

Hope those tips help a bit!