CMSC 170

Introduction to Artificial Intelligence

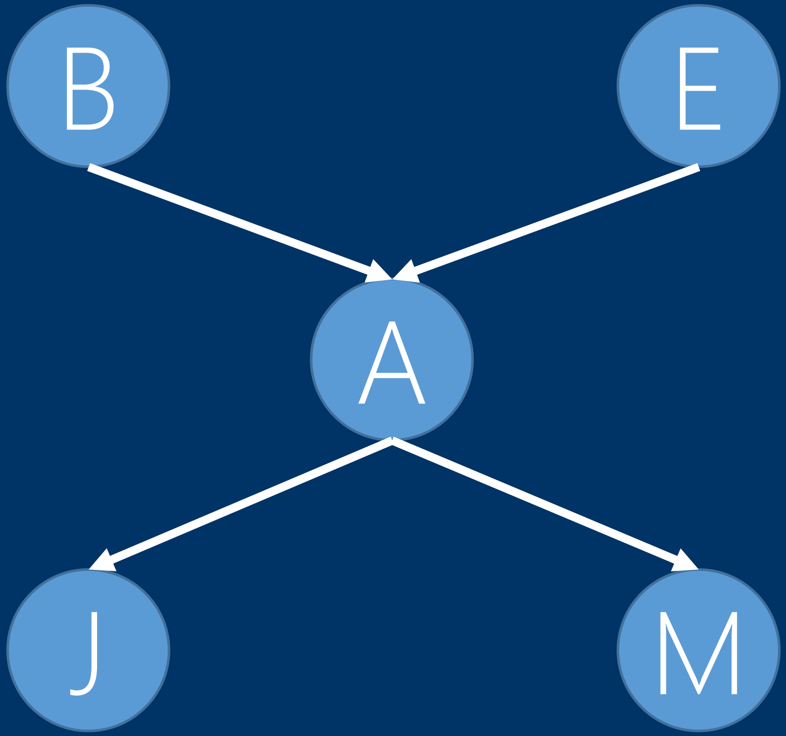
2^{nd} Semester AY 2013-2014

CNM Peralta

Probabilistic Inference

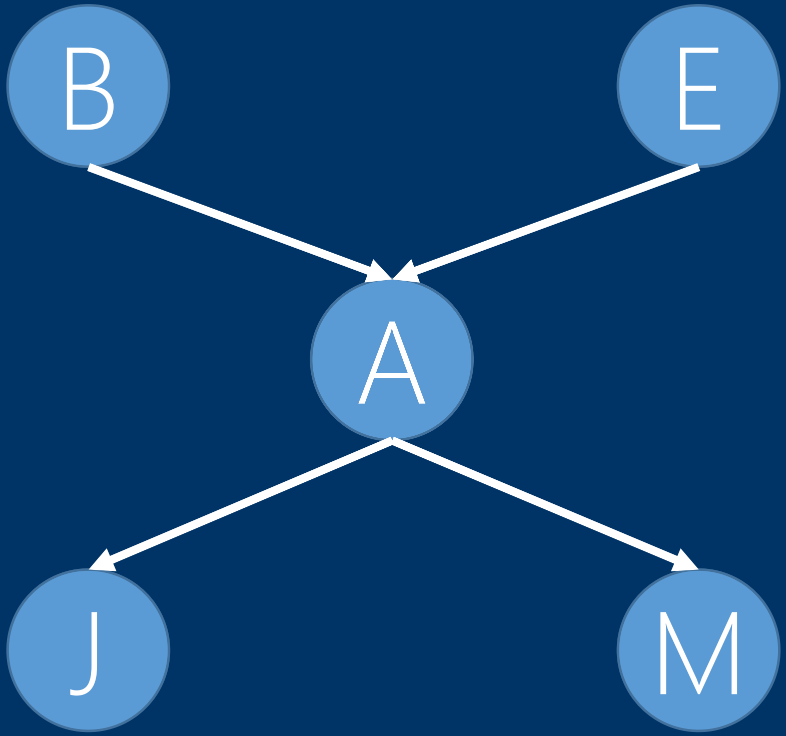
We will now use our knowledge of Bayes Networks for **inference**.

Say we have a Bayes Network...

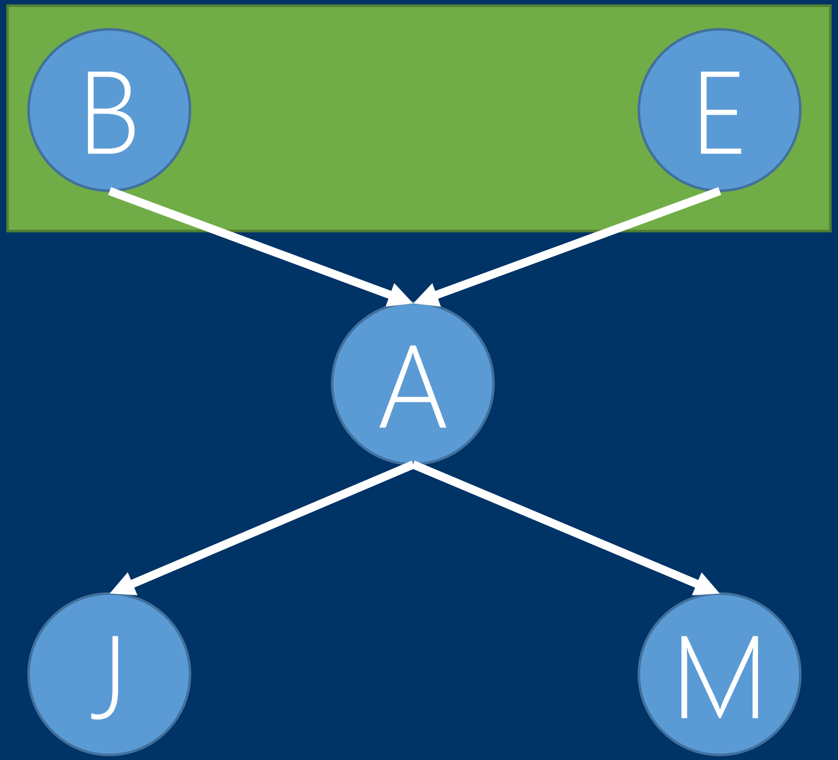


If there is a (B)urglary or an (E)arthquake, the house (A)larm will go off. If the house alarm goes off, then either (J)ohn or (M)ary will call.

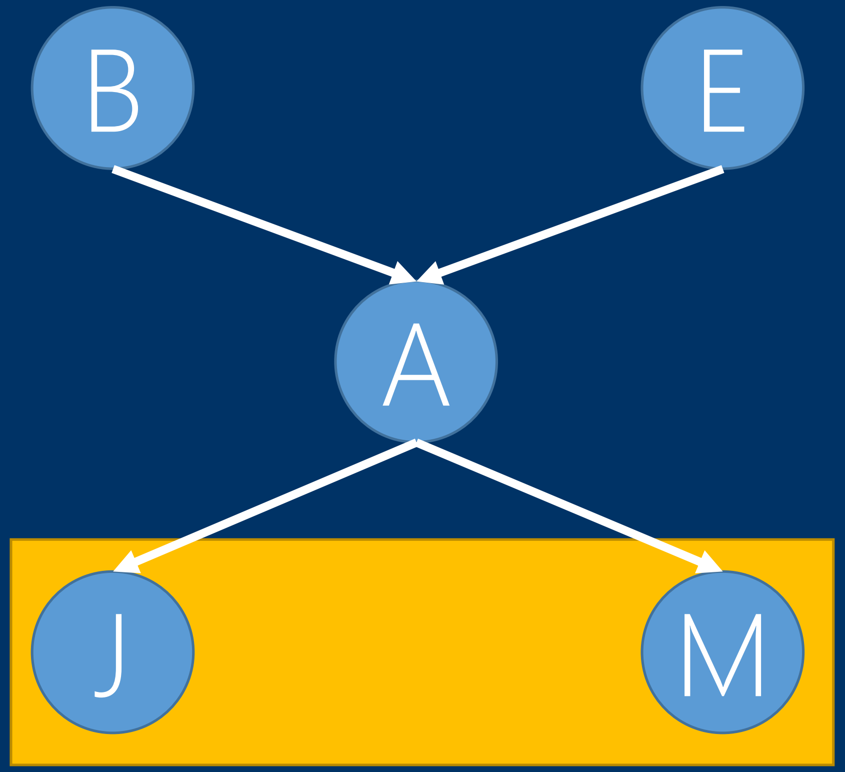
We can translate this as an input-output problem.



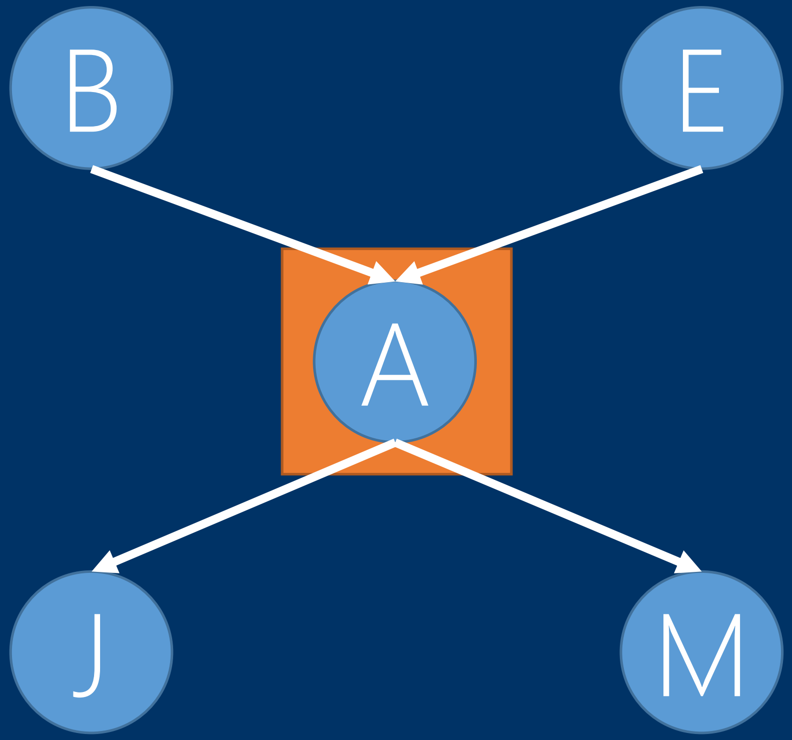
Given inputs (B and E), what are the outputs (J and M)?



The inputs are called the ***evidence***.



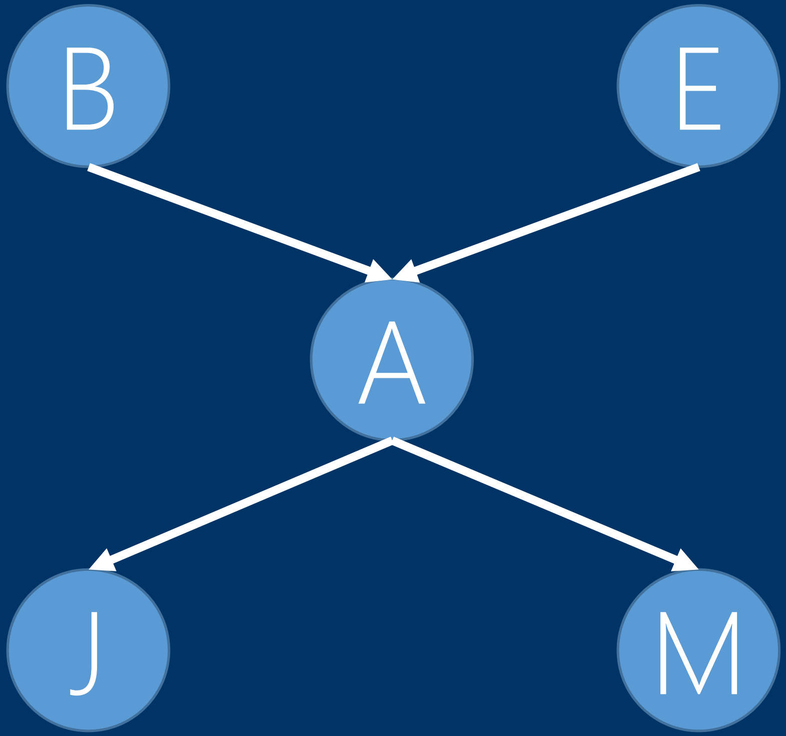
The outputs are called the ***query***.



Anything that isn't evidence or query is a ***hidden*** variable.

The output is a complete joint probability distribution over the query variables, called the

posterior distribution given the evidence.



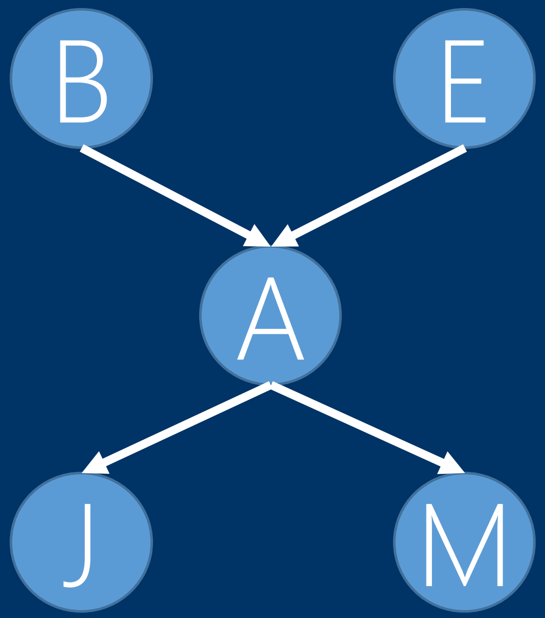
Output:

$P(Q_1,Q_2,\ldots|E_1=e_1,E_2=e_2\ldots)$

Probability distribution of one or more query variables…

…given the values of the evidence variables.

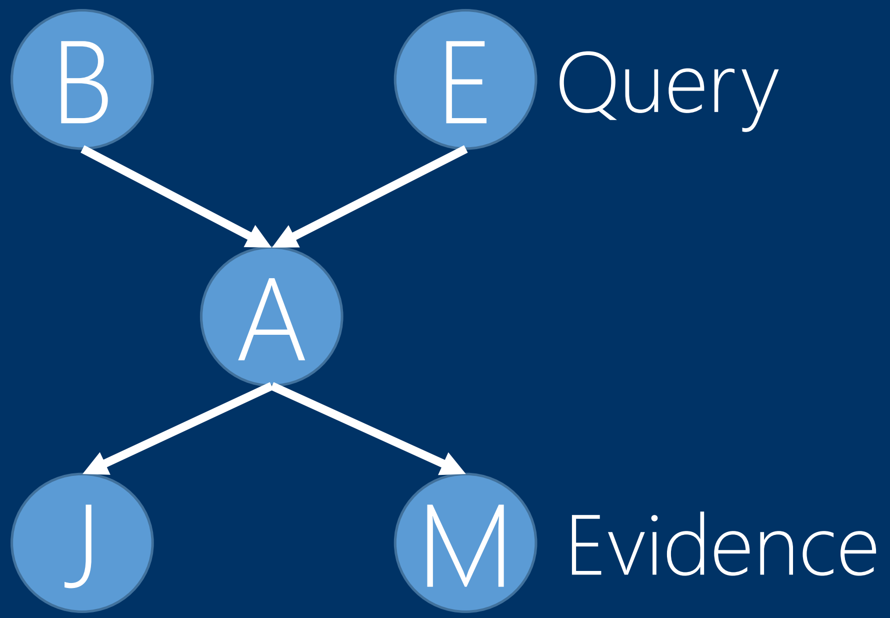
We can also ask what the **most likely explanation** is.



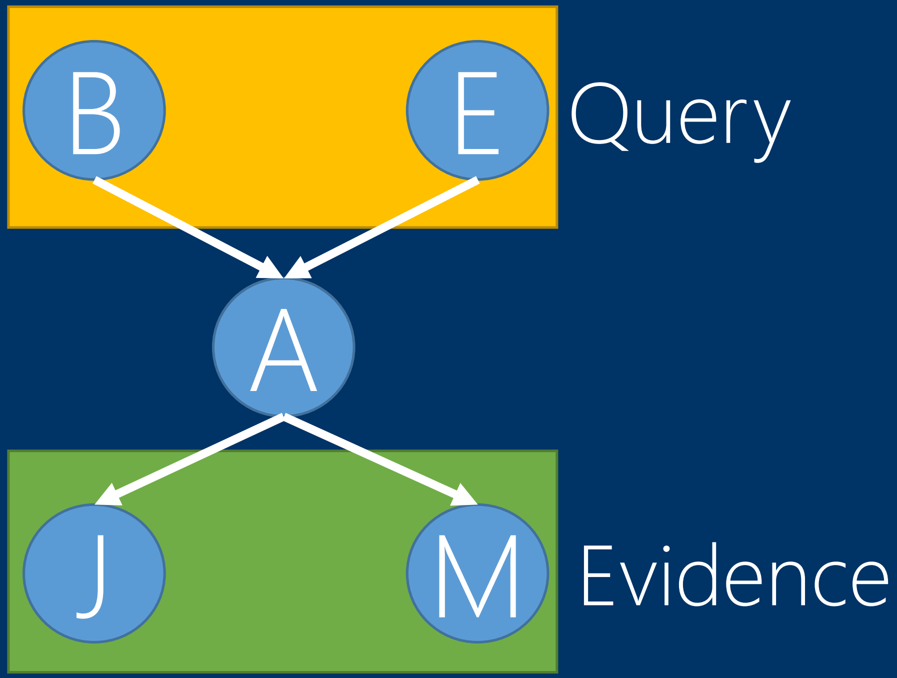
Output:

$$\max(P(Q_1=q_1,Q_2=q_2,\ldots \mid E_1=e_1,E_2=e_2\ldots))$$

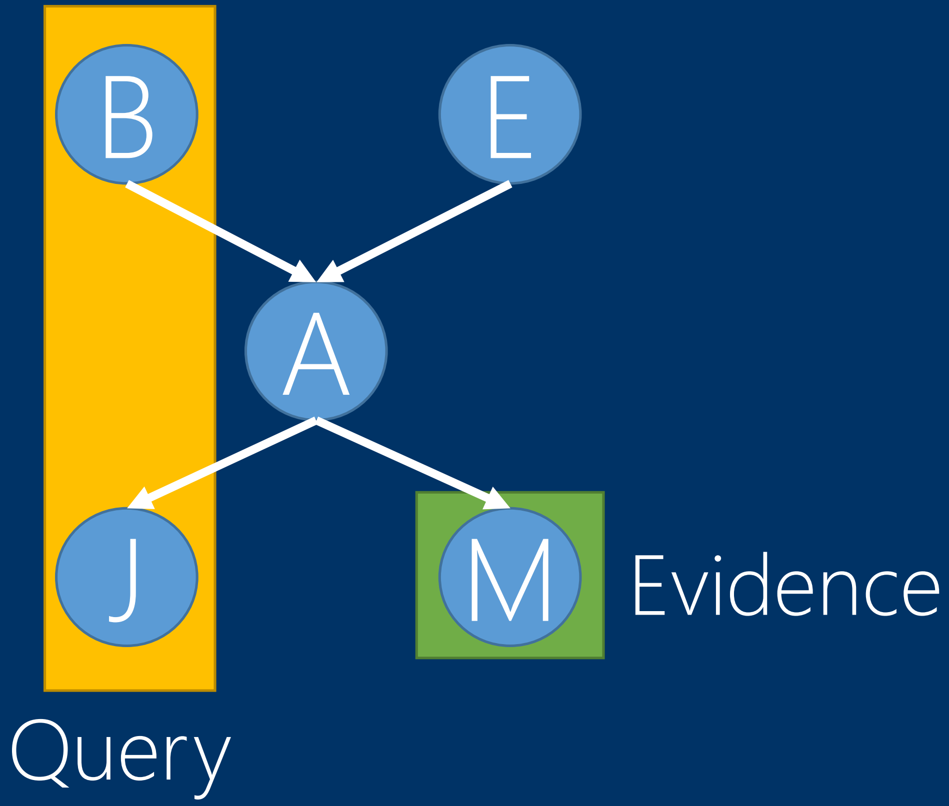
What combination of values of query variables (q_1, q_2) is most likely to occur given the evidence values (e_1, e_2).



Take note that we can also reverse the flow of input-output.



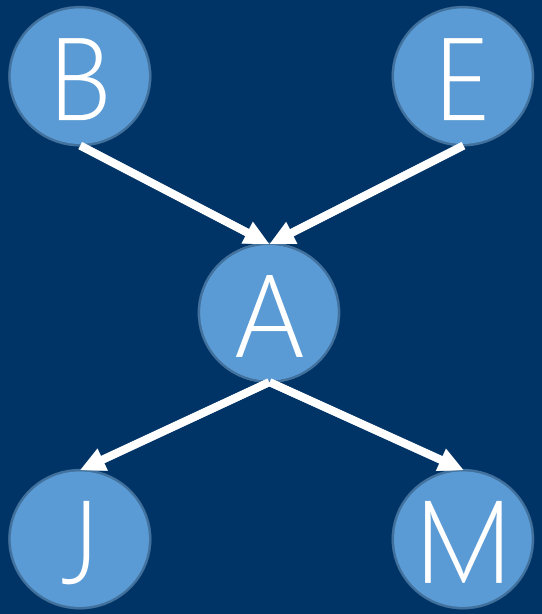
Take note that we can also reverse the flow of input-output…



...or have any combination of evidence and query variables, actually.

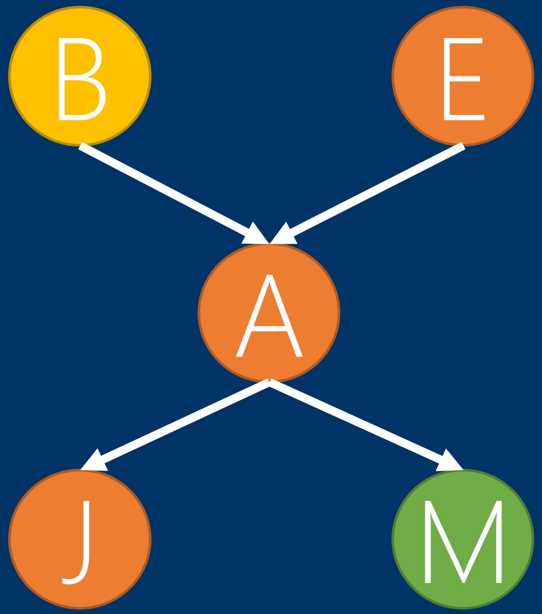
Example

If Mary calls and we want to know if there was a burglary, which of the nodes is considered evidence, query or, hidden?



Answer

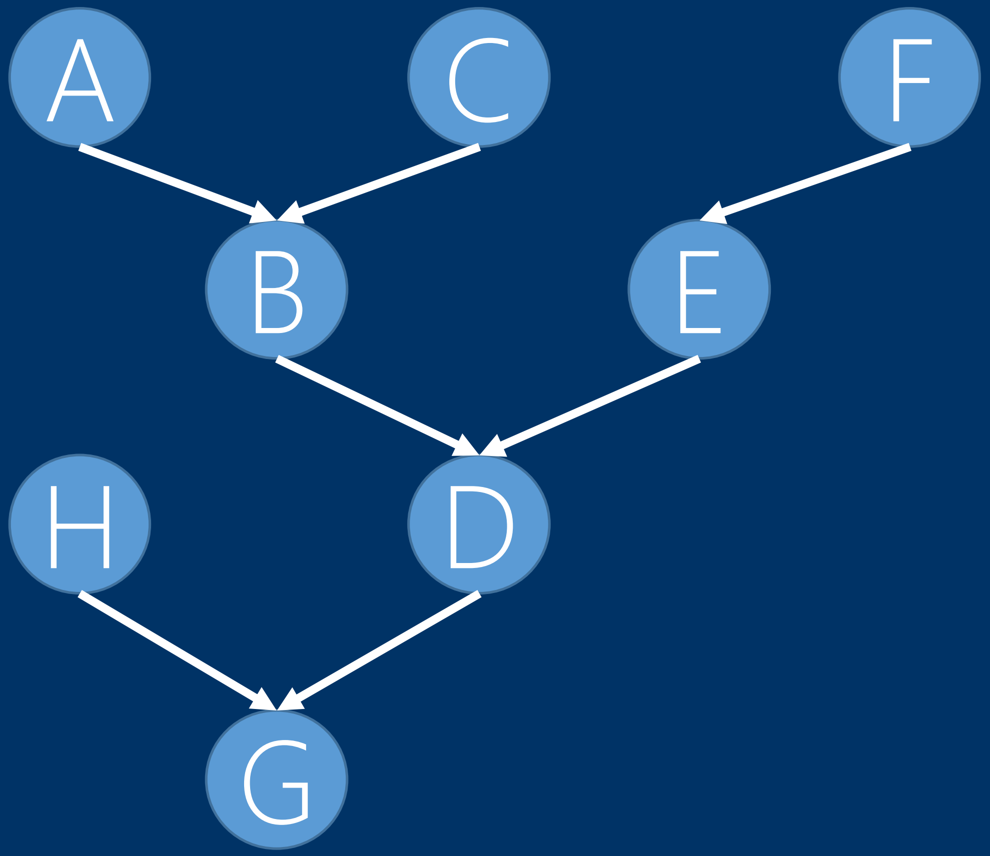
M is the evidence, B is the query, and the rest are hidden nodes.



Quiz (1/4)

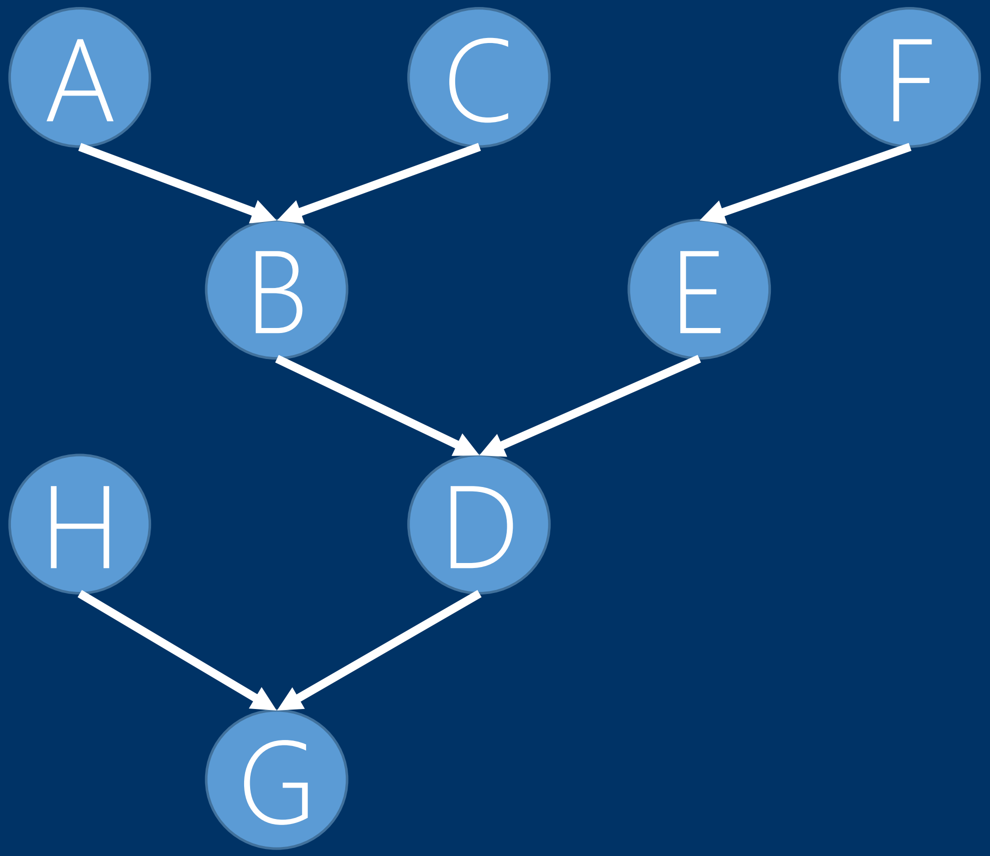
Given the Bayes net, answer the ff.:

1. $F \perp A$
2. $F \perp A | D$
3. $F \perp A | G$
4. $F \perp A | H$

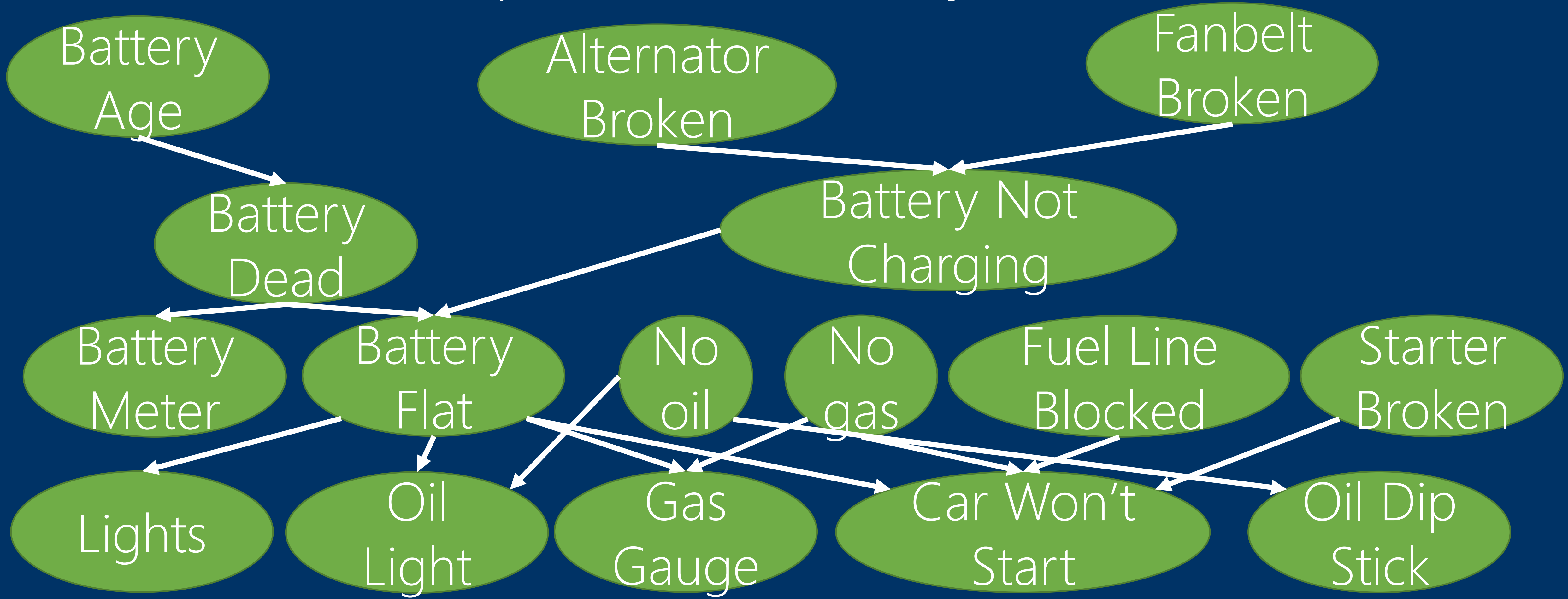


Quiz (1/4)

How many parameters are required to represent this Bayes' network?



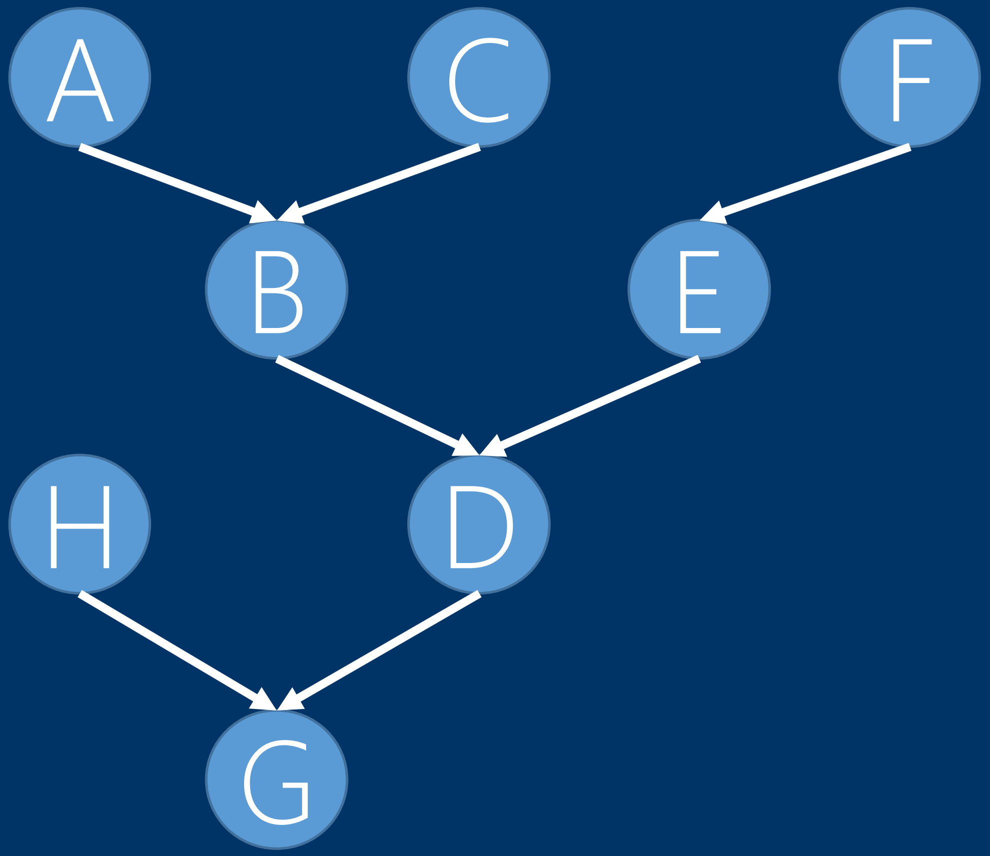
Quiz (1/4) How many parameters are needed to represent this Bayes network?



Answers

Given the Bayes net, answer the ff.:

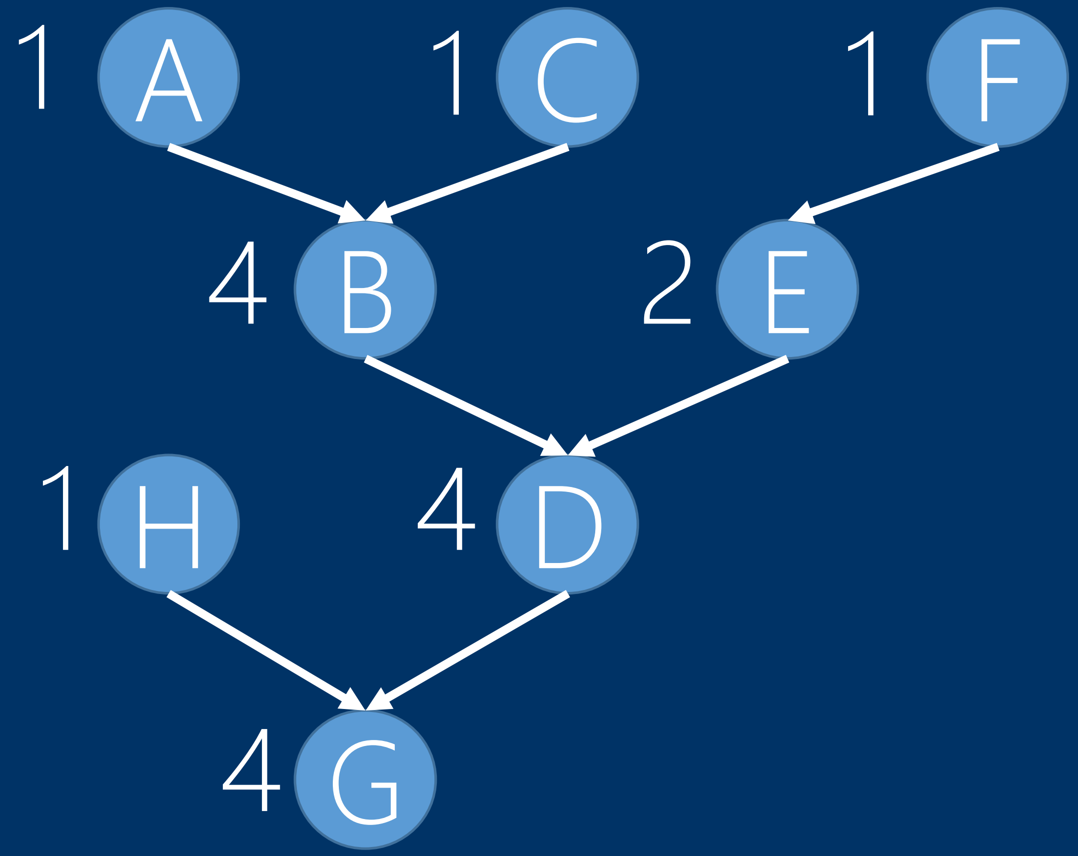
1. $F \perp A$ - Yes
2. $F \perp A | D$ - No
3. $F \perp A | G$ - No
4. $F \perp A | H$ - Yes



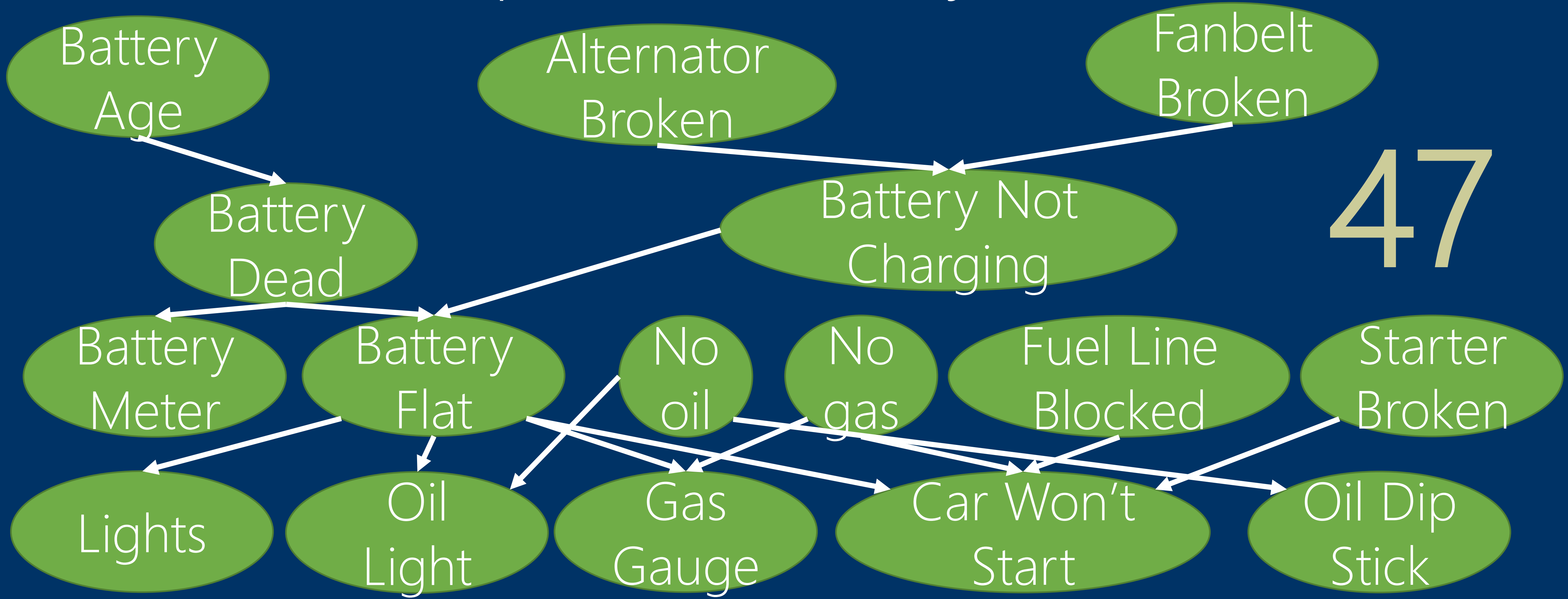
Answers

How many parameters are required to represent this Bayes' network?

18



Answers How many parameters are needed to represent this Bayes network?



47

Enumeration

is a method of probabilistic inference.

Enumeration

goes through all possibilities, adds them up and comes up with an answer.

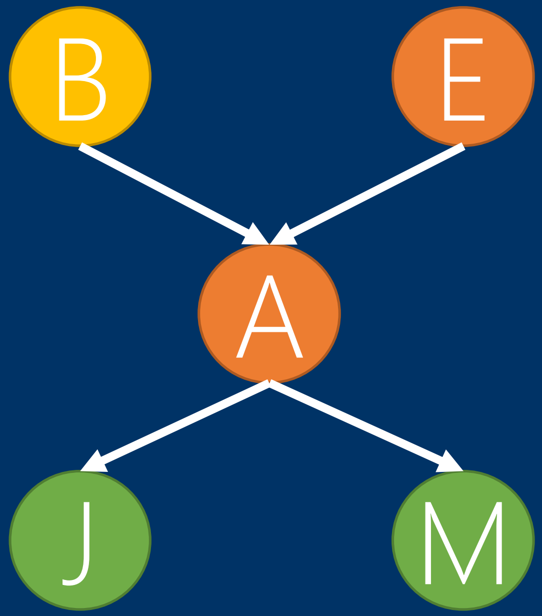
Before we continue...

$$P(Q|E) = \frac{P(Q,E)}{P(E)}$$

Conditional Probability

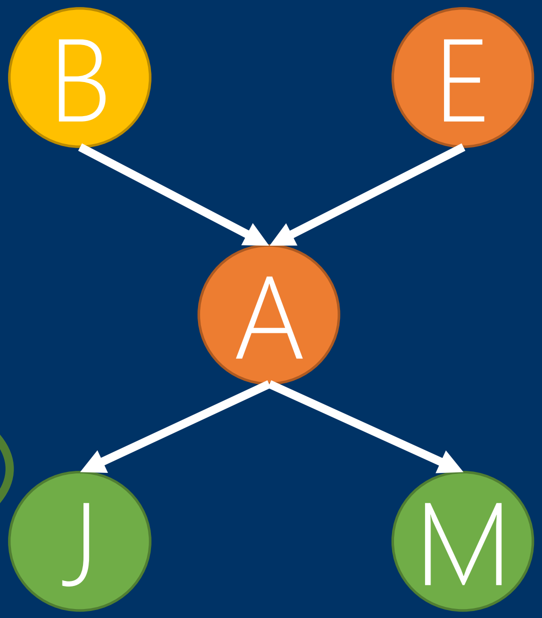
Example

$$P(+B|+J + M) = \frac{P(+B, +J, +M)}{P(+J, +M)}$$



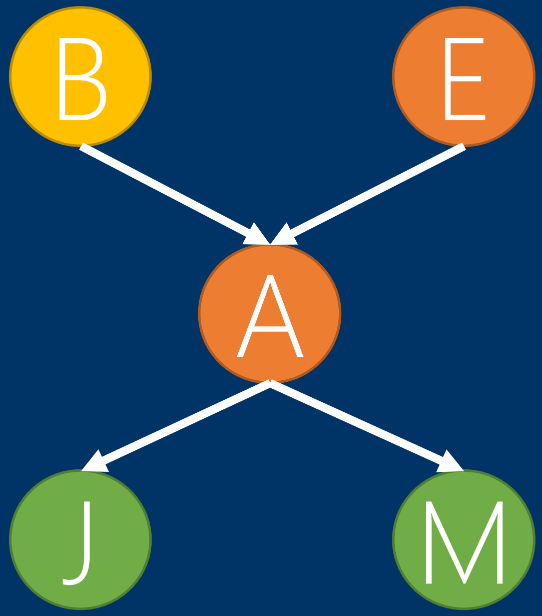
Example

$$P(+B,+J,+M)$$
$$=\sum_{E}\sum_{A}P(+B,+J,+M,E,A)$$



We need to expand this using the flow of arrows from the Bayes network.

Example



$$P(+B,+J,+M)$$

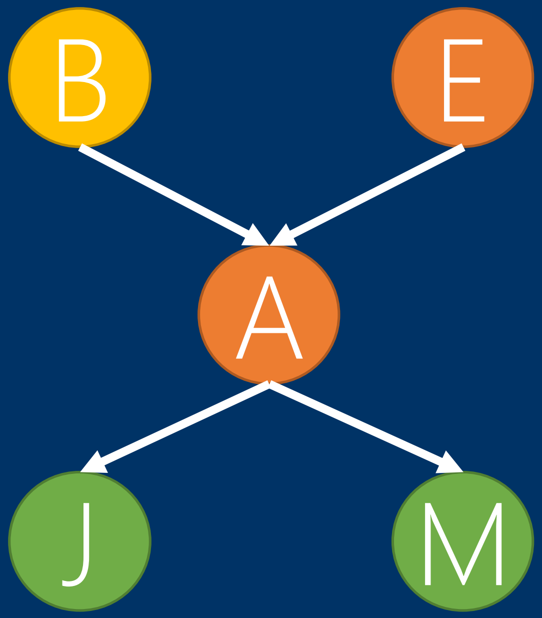
$$= \sum_{E}\sum_{A} P(+B,+J,+M,E,A)$$

$$P(+B)P(E)P(A|+B,E)P(+J|A)P(+M|A)$$

Let's call this f(e,a).

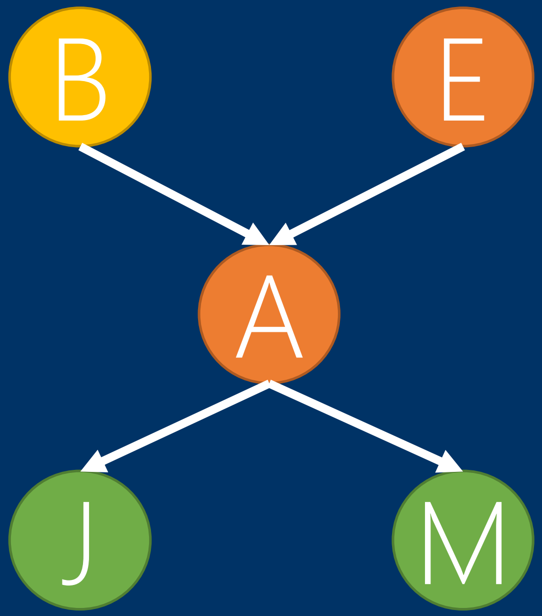
Example

$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(E,A) =$$
$$P(+B)P(E)P(A|+B,E)P(+J|A)P(+M|A)$$

B	P(B)
+B	0.001
¬B	0.999



E	P(E)
+E	0.002
¬E	0.998

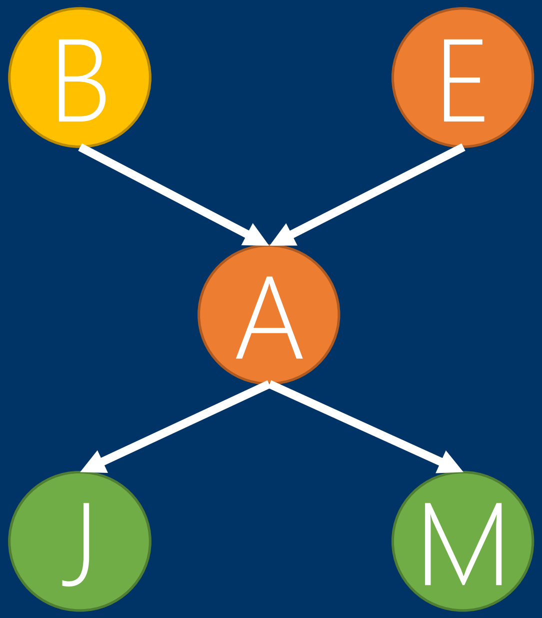
B	E	A	P(A|B,E)
+B	+E	+A	0.95
+B	+E	¬A	0.05
+B	¬E	+A	0.94
+B	¬E	¬A	0.06
¬B	+E	+A	0.29
¬B	+E	¬A	0.71
¬B	¬E	+A	0.001
¬B	¬E	¬A	0.999

A	J	P(J|A)
+A	+J	0.9
+A	¬J	0.1
¬A	+J	0.05
¬A	¬J	0.95

A	M	P(M|A)
+A	+M	0.7
+A	¬M	0.3
¬A	+M	0.01
¬A	¬M	0.99

Example

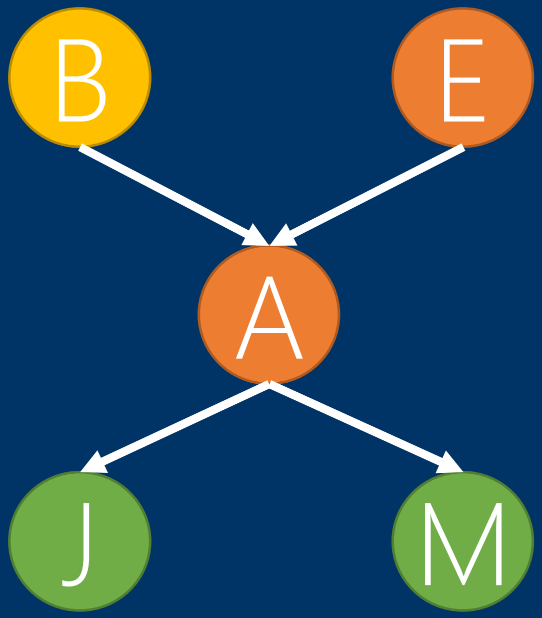
$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(+E,+A) =$$
$$0.001 \cdot 0.002 \cdot 0.95 \cdot 0.9 \cdot 0.7$$

Example

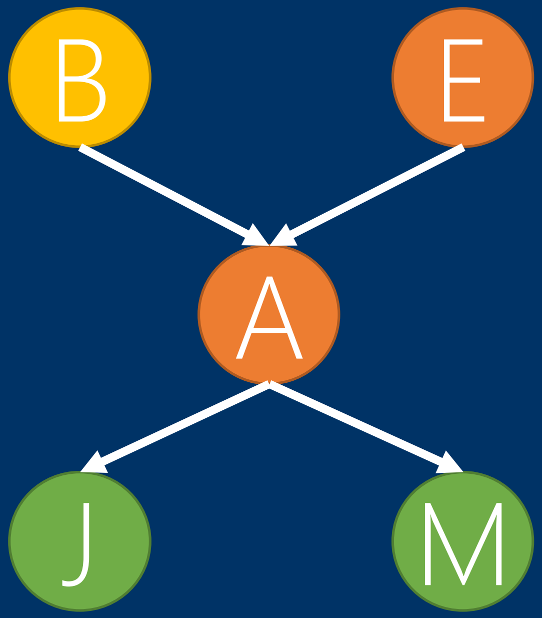
$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(+E,+A) =$$
$$0.001 \cdot 0.002 \cdot 0.95 \cdot 0.9 \cdot 0.7 = 0.000001197$$

Example

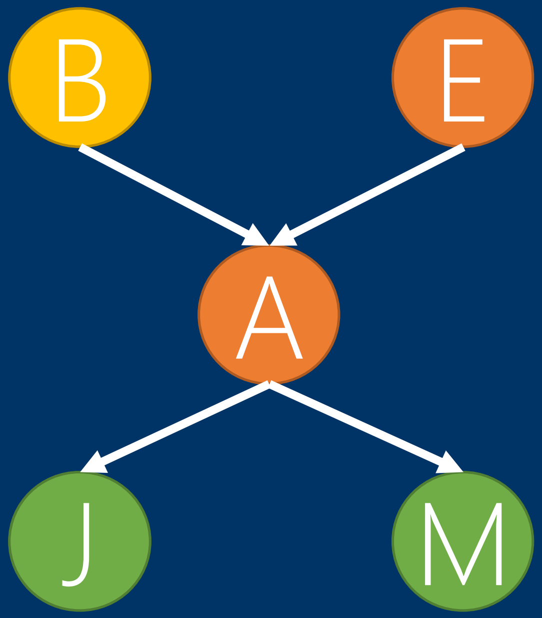
$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(+E,\neg A) =$$
$$0.001 \cdot 0.002 \cdot 0.05 \cdot 0.05 \cdot 0.01$$

Example

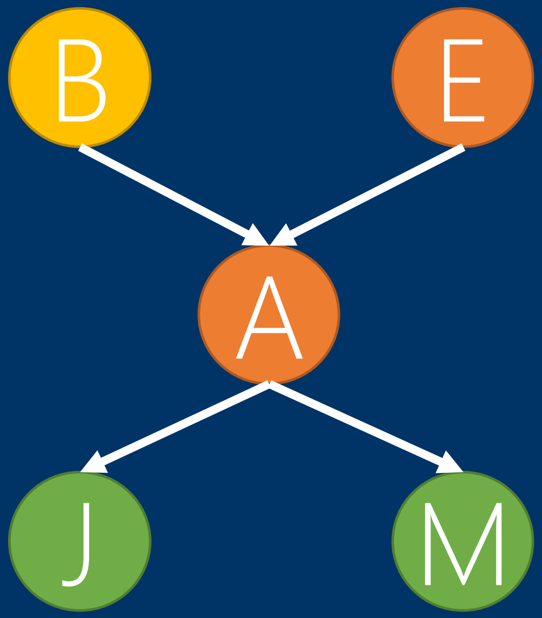
$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(+E,\neg A) =$$
$$0.001 \cdot 0.002 \cdot 0.05 \cdot 0.05 \cdot 0.01$$
$$= 0.000000000005$$

Example

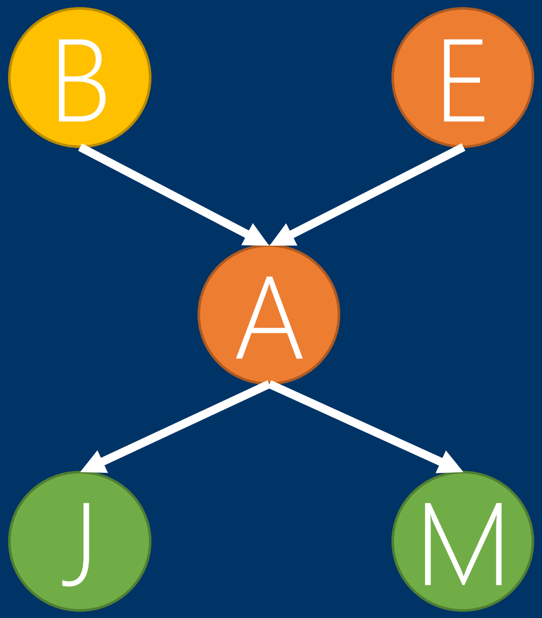
$$P(+B, +J, +M, E, A)$$
$$= f(+E, +A) + f(+E, \neg A)$$
$$+ f(\neg E, +A) + f(\neg E, \neg A)$$



$$f(\neg E, +A) =$$
$$0.001 \cdot 0.998 \cdot 0.94 \cdot 0.9 \cdot 0.7$$

Example

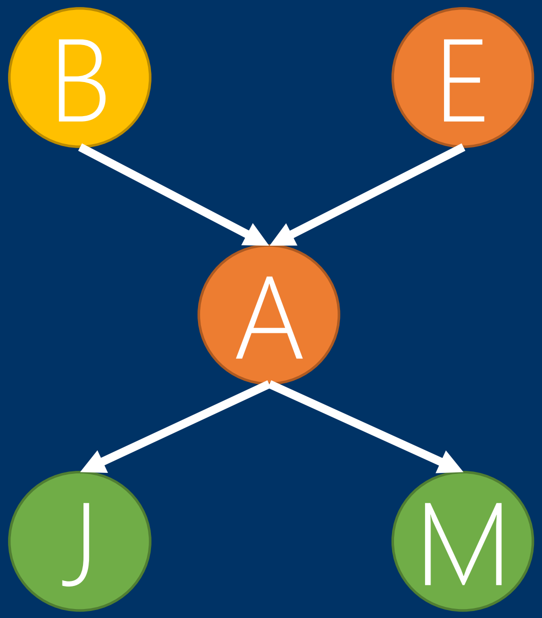
$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(\neg E,+A) =$$
$$0.001 \cdot 0.998 \cdot 0.94 \cdot 0.9 \cdot 0.7$$
$$= 0.0005910156$$

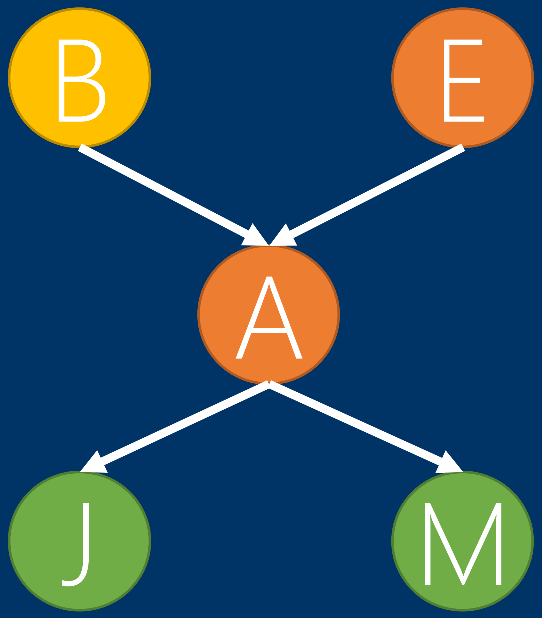
Example

$$P(+B,+J,+M,E,A)$$
$$= f(+E,+A) + f(+E,\neg A)$$
$$+ f(\neg E,+A) + f(\neg E,\neg A)$$



$$f(\neg E,\neg A) =$$
$$0.001 \cdot 0.998 \cdot 0.06 \cdot 0.05 \cdot 0.01$$

Example

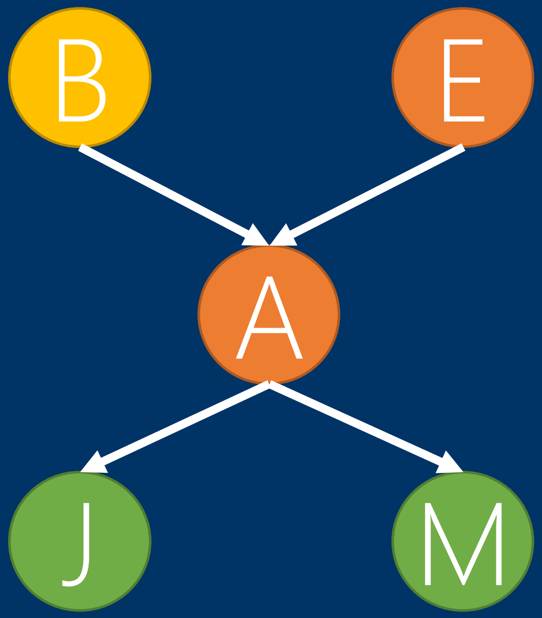


$$P(+B, +J, +M, E, A)$$
$$= f(+E, +A) + f(+E, \neg A)$$
$$+ f(\neg E, +A) + f(\neg E, \neg A)$$

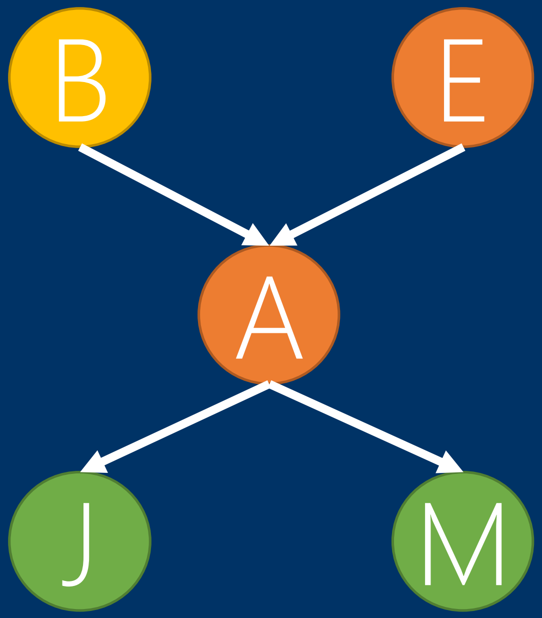
$$f(\neg E, \neg A) =$$
$$0.001 \cdot 0.998 \cdot 0.06 \cdot 0.05 \cdot 0.01$$
$$= 0.00000002994$$

Example

$$P(+B,+J,+M,E,A)$$
$$= 0.000001197$$
$$+ 0.00000000005$$
$$+ 0.0005910156$$
$$+ 0.00000002994$$

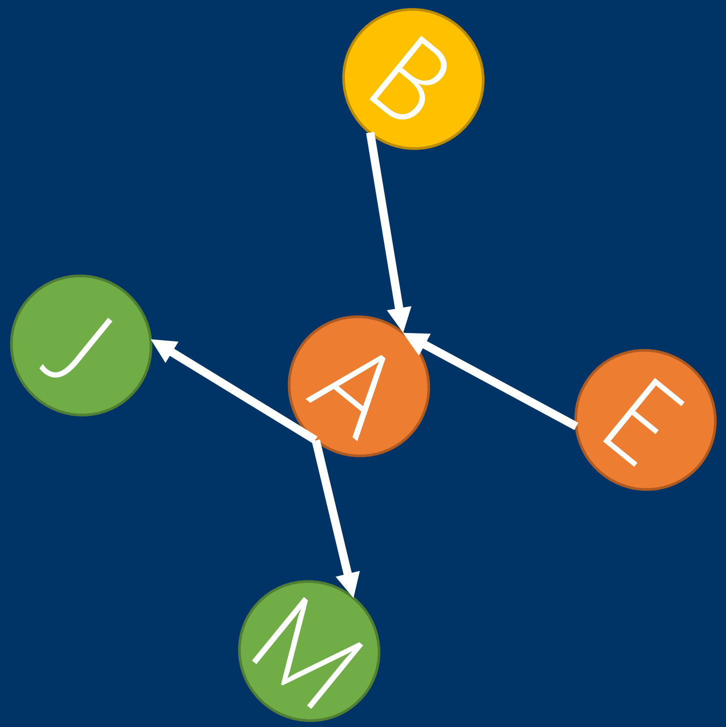


Example



$$P(+B,+J,+M,E,A)$$
$$= 0.000001197$$
$$+ 0.0000000005$$
$$+ 0.0005910156$$
$$+ 0.0000002994$$
$$= 0.00059224259$$

Example



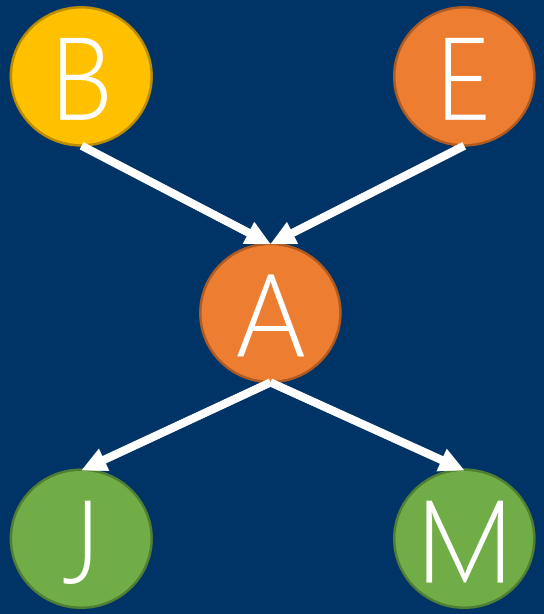
$$P(+B|+J + M) = \frac{0.00059224259}{P(+J,+M)}$$

Oops, that just the numerator. :>

Example

We still need to solve for the denominator:

$$P(+J, +M) = \sum_B \sum_E \sum_A P(+J, +M, B, E, A)$$



Fast-forward

$$P(+B|+J + M) = \frac{0.00059224259}{P(+J, +M)} = 0.284$$

Why is the probability of a burglary, even though both John and Mary called, low?

PROBLEM:

Enumeration takes too long when the number of nodes is large.

We can "pull out terms."

$$\sum_E \sum_A P(+B)P(E)P(A|+B,E)P(+J|A)P(+M|A)$$

$$= P(+B)\sum_E P(E)\sum_A P(A|+B,E)P(+J|A)P(+M|A)$$

BUT.

Pulling out terms will only reduce the number of times certain values need to be used.

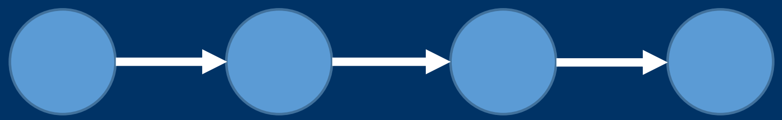
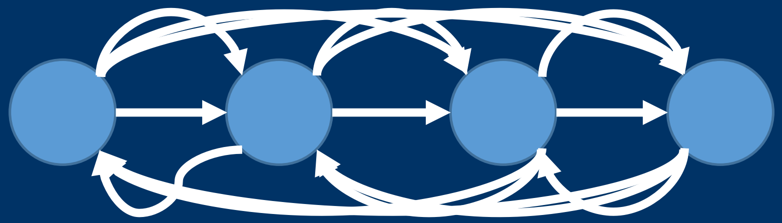
BUT.

The running time, and the number of probability values are still the same.

Another approach is called

"**Maximizing Independence.**"

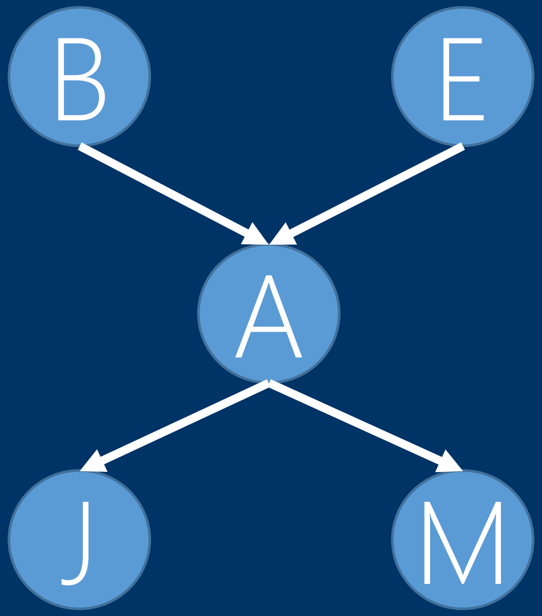
Bayes network structure dictates how efficiently inference can be performed on it.

Structure	Running Time
 n linear nodes	$O(n)$
 Complete graph	$O(2^n)$

We can try to rearrange Bayes networks, but still keeping in mind the (in)dependence of the nodes on each other.

Example

Original

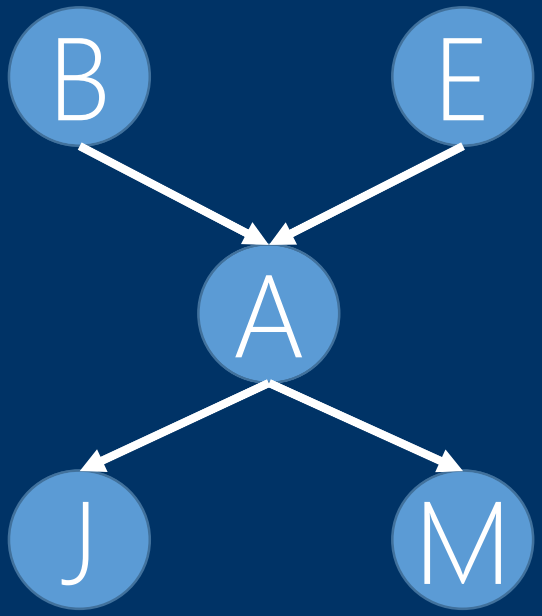


Restructured

Are J and M independent?

Example

Original



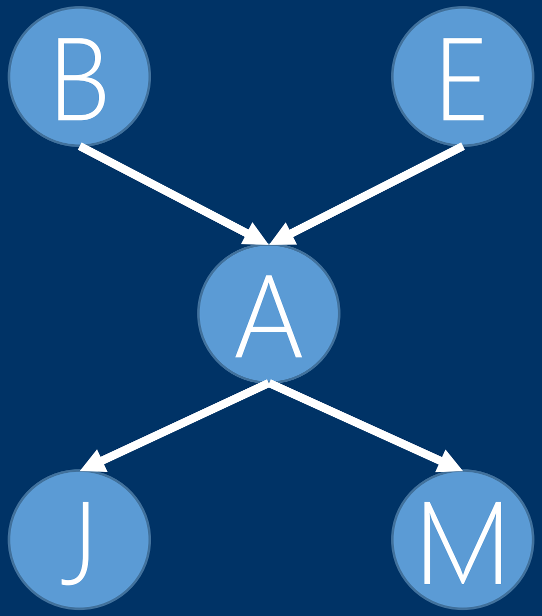
Restructured

Are J and M independent?

No, they aren't!

Example

Original

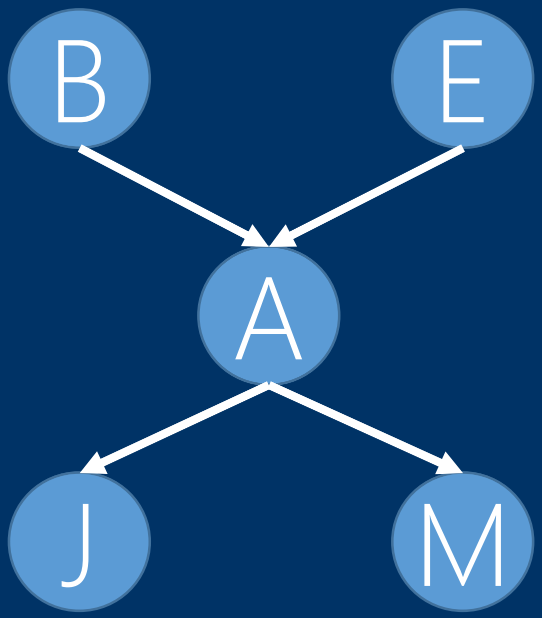


Restructured

Here, J & M are independent if A is known.

Example

Original

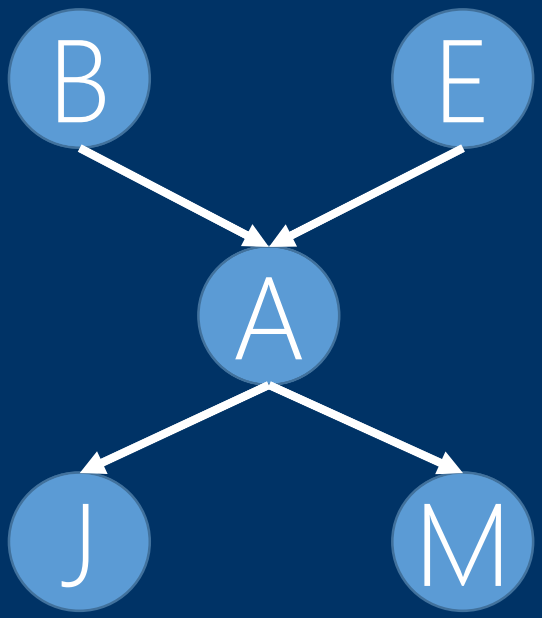


Restructured

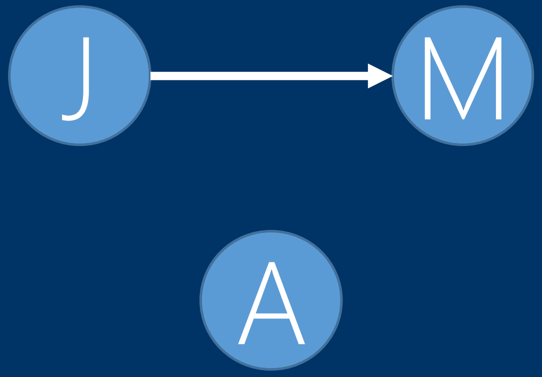
Since A is not known in this restructured network, J & M are dependent.

Example

Original



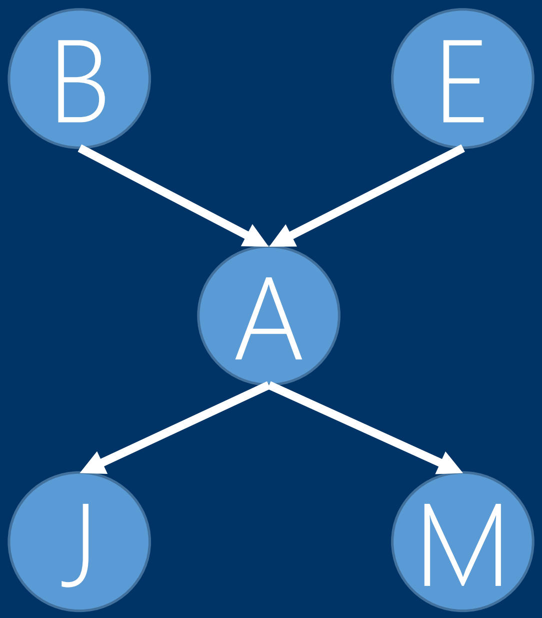
Restructured



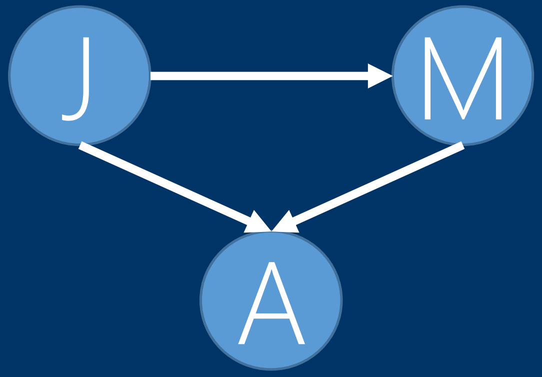
Is A dependent on J or/and M?

Example

Original



Restructured

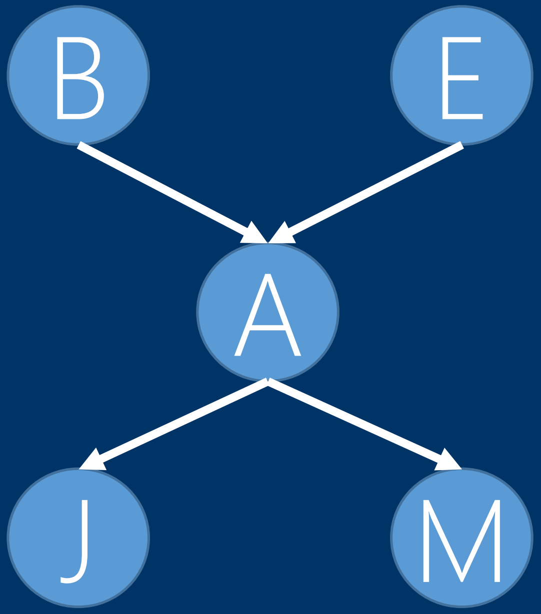


Is A dependent on J or/and M?

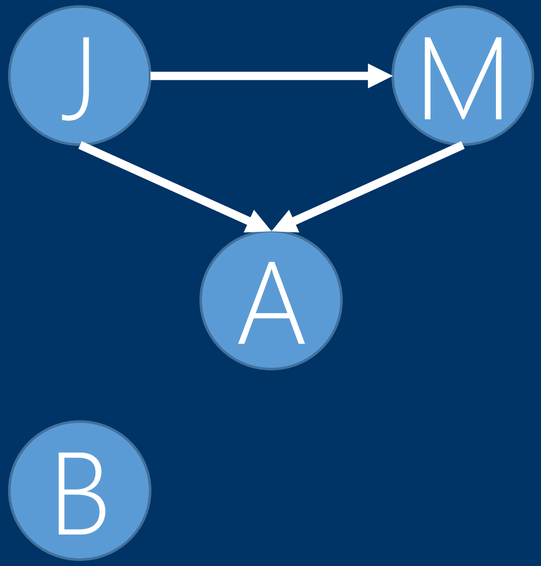
Yes to both.

Example

Original

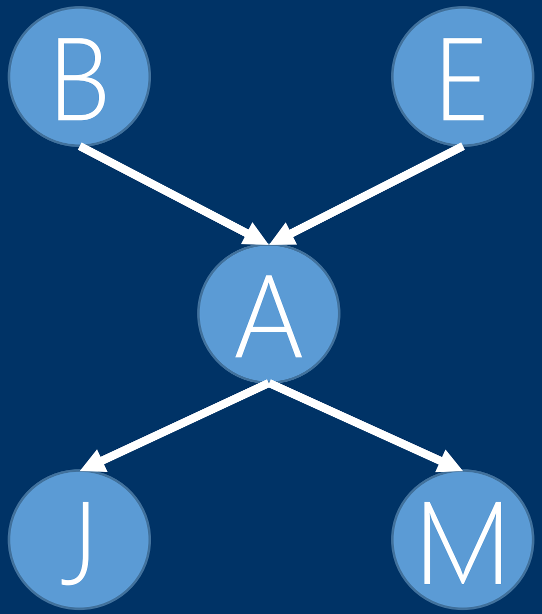


Restructured

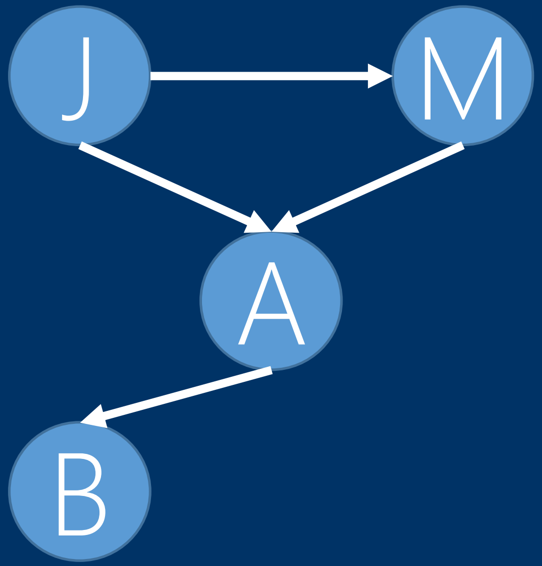


Example

Original

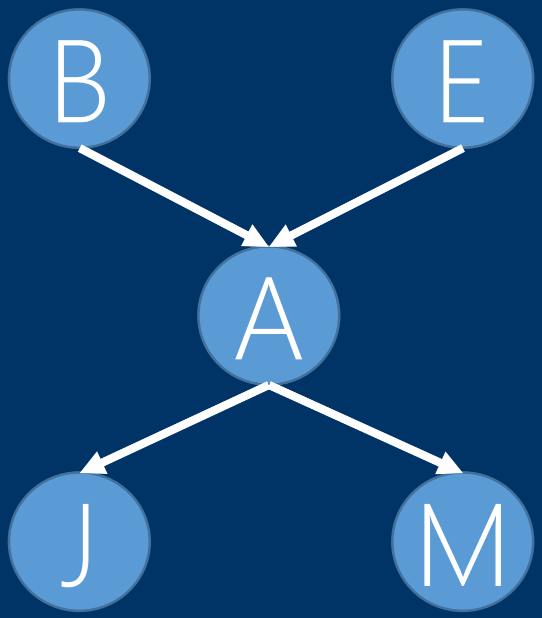


Restructured

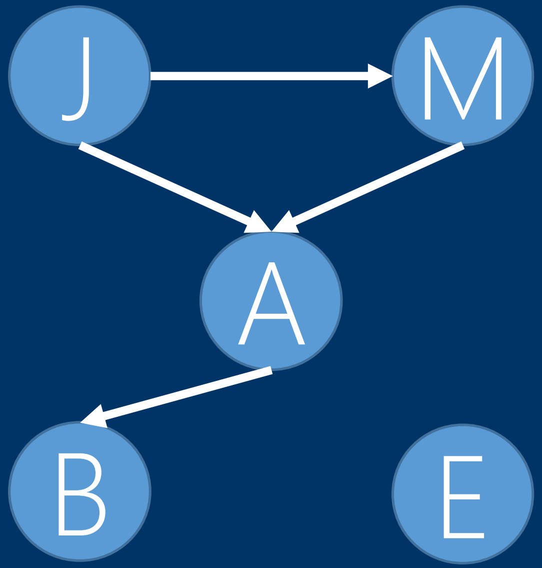


Example

Original

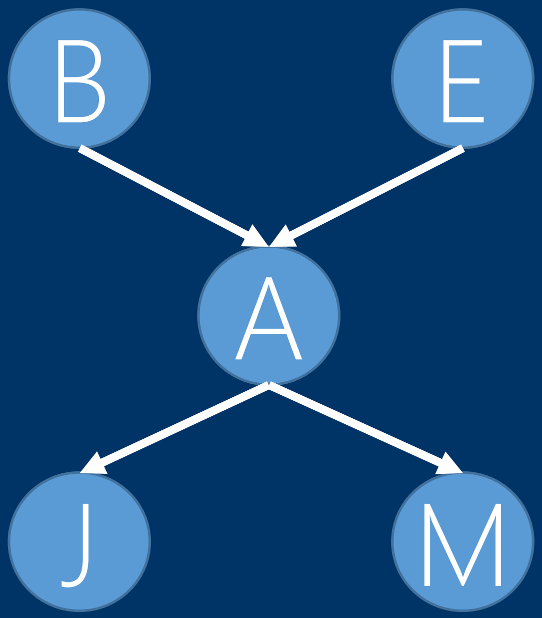


Restructured

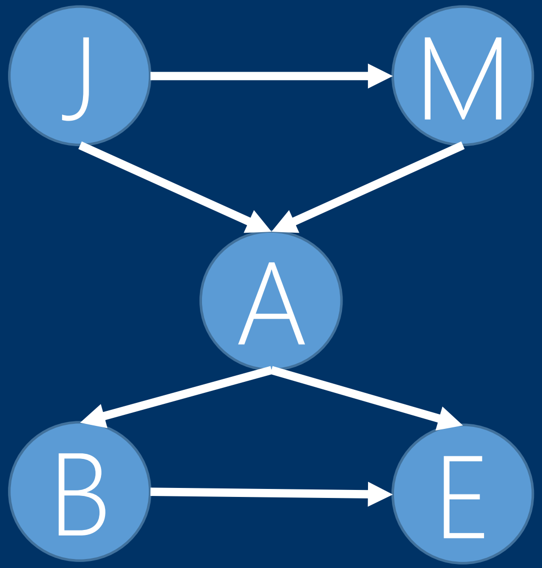


Example

Original



Restructured



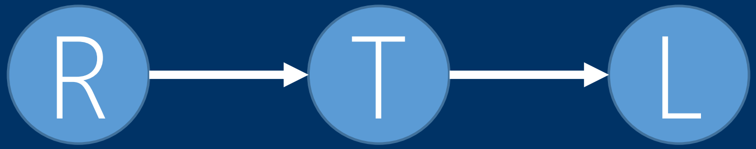
Bayes networks are most efficient when they are structured in the ***causal direction***.

An alternative approach for inference is ***variable elimination***.

Variable elimination

is faster than enumeration in most practical networks.

Example



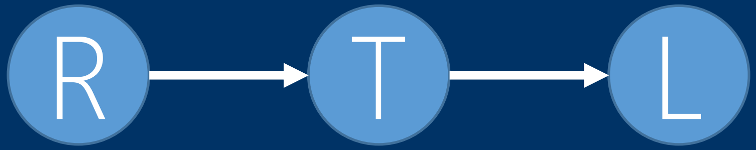
R	P(R)
+R	0.1
-R	0.9

R	T	P(T|R)
+R	+T	0.8
+R	-T	0.2
-R	+T	0.1
-R	-T	0.9

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

Example

First, we need to join factors.



R	P(R)
+R	0.1
-R	0.9

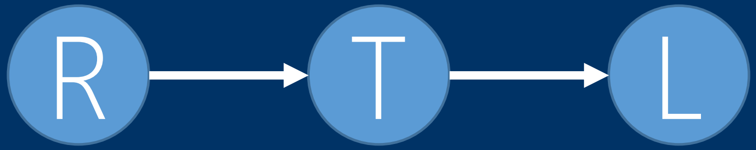
R	T	P(T|R)
+R	+T	0.8
+R	-T	0.2
-R	+T	0.1
-R	-T	0.9

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

Example

We will join R and T.

R	P(R)
+R	0.1
-R	0.9

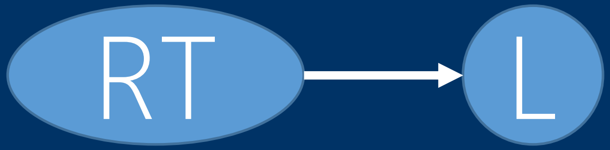


R	T	P(T|R)
+R	+T	0.8
+R	-T	0.2
-R	+T	0.1
-R	-T	0.9

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

Example

We will join R and T.



R	P(R)
+R	0.1
-R	0.9

X

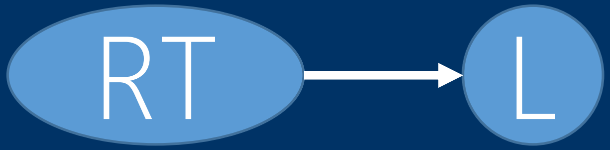
R	T	P(T|R)
+R	+T	0.8
+R	-T	0.2
-R	+T	0.1
-R	-T	0.9

=

R	T	P(R,T)
+R	+T	
+R	-T	
-R	+T	
-R	-T	

Example

We will join R and T.



R	P(R)
+R	0.1
-R	0.9

X

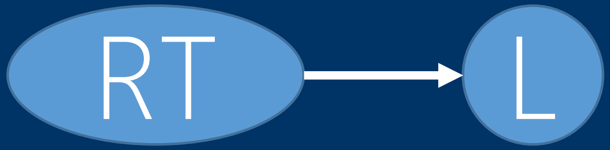
R	T	P(T|R)
+R	+T	0.8
+R	-T	0.2
-R	+T	0.1
-R	-T	0.9

=

R	T	P(R,T)
+R	+T	0.08
+R	-T	0.02
-R	+T	0.09
-R	-T	0.81

Example

Now, we will eliminate (or marginalize) R from RT



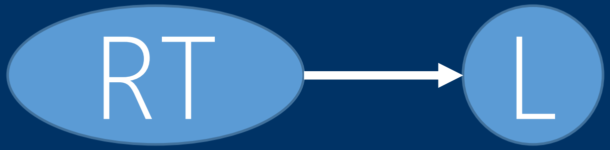
R	T	P(R,T)
+R	+T	0.08
+R	-T	0.02
-R	+T	0.09
-R	-T	0.81

- R =

T	P(T)
+T	
-T	

Example

Now, we will eliminate (or marginalize) R from RT

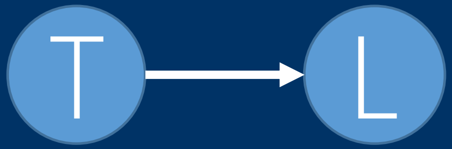


R	T	P(R,T)
+R	+T	0.08
+R	-T	0.02
-R	+T	0.09
-R	-T	0.81

- R =

T	P(T)
+T	0.17
-T	0.83

Example



T	P(T)
+T	0.17
-T	0.83

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

Example

Next, we join T and L

T,L

T	P(T)
+T	0.17
-T	0.83

X

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

=

T	L	P(L,T)
+T	+L	
+T	-L	
-T	+L	
-T	-L	

Example

Next, we join T and L

T,L

T	P(T)
+T	0.17
-T	0.83

X

T	L	P(L|T)
+T	+L	0.3
+T	-L	0.7
-T	+L	0.1
-T	-L	0.9

=

T	L	P(L,T)
+T	+L	0.051
+T	-L	0.119
-T	+L	0.083
-T	-L	0.747

Example

Lastly, we eliminate T.

T,L

T	L	P(L,T)
+T	+L	0.051
+T	-L	0.119
-T	+L	0.083
-T	-L	0.747

- T =

L	P(L)
+L	
-L	

Example

Lastly, we eliminate T.

T,L

T	L	P(L,T)
+T	+L	0.051
+T	-L	0.119
-T	+L	0.083
-T	-L	0.747

\- T =

L	P(L)
+L	0.134
-L	0.866

Example

Lastly, we eliminate T.

L	P(L)
+L	0.134
-L	0.866

The order of joining factors and elimination will dictate if variable elimination will be more efficient than enumeration.

Remember…

Probabilistic inference over Bayes networks in general is actually NP-hard.