COMP 424 - Artificial Intelligence
Lecture 15: Bayesian Networks 2

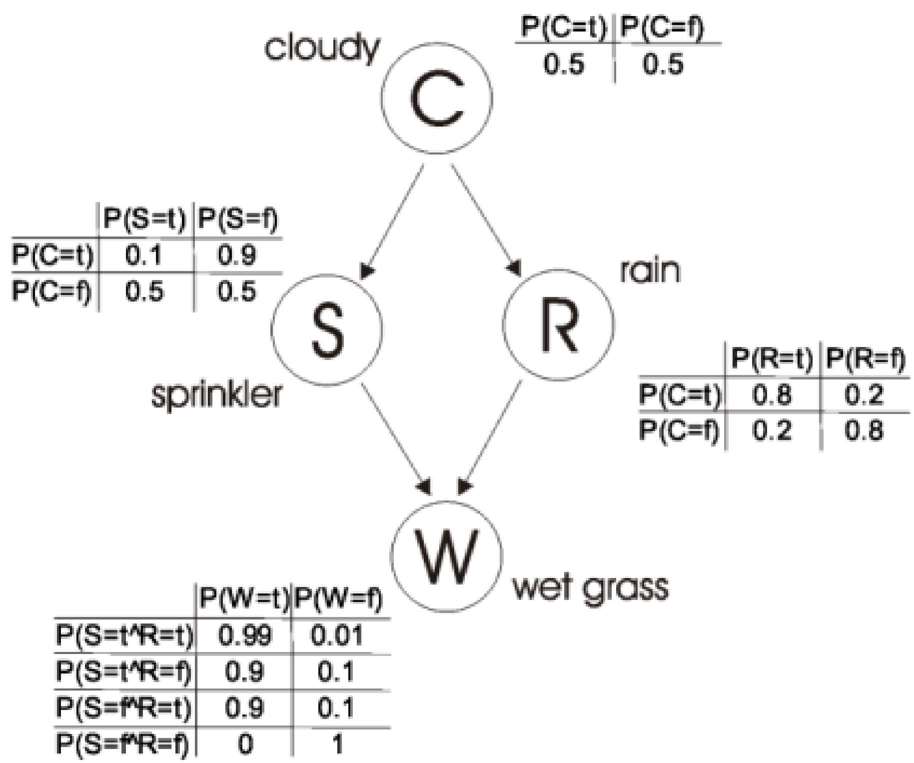
Instructor: Jackie CK Cheung (jcheung@cs.mcgill.ca)

Readings: R&N Ch 14

Quiz: Bayes net inference

1. What is the unconditional probability of WetGrass, P(W=1)?
    a) P(W=1|S,R)P(S,R)
    b) P(W=1|S,R)P(S|C)P(R|C)P(C)
    c) $\Sigma_{s,r}$ P(W=1 | S=s, R=r)
    d) $\Sigma_{s,r}$ P(W=1, S=s, R=r)
    e) $\Sigma_{s,r,c}$ P(W=1,S=s, R=r, C=c)

2. What is the conditional probability of Rain, given that the grass is wet, P(R=1|W=1)?
    a) P(R=1,W=1) / (P(W=1) + P(R=1))
    b) P(R=1|W=1,C=0)+P(R=1|W=1,C=1)
    c) $\Sigma_{s,c}$ P(W=1, R=1, S=s, C=c)
    d) $\Sigma_{s,c}$ P(W=1, R=1, S=s, C=c) / $\Sigma_{s,c,r}$ P(W=1, R=r, S=s, C=c)



P(C=t)	P(C=f)
0.5	0.5

	P(S=t)	P(S=f)
P(C=t)	0.1	0.9
P(C=f)	0.5	0.5

	P(R=t)	P(R=f)
P(C=t)	0.8	0.2
P(C=f)	0.2	0.8

	P(W=t)	P(W=f)
P(S=t^R=t)	0.99	0.01
P(S=t^R=f)	0.9	0.1
P(S=f^R=t)	0.9	0.1
P(S=f^R=f)	0	1

To-do later: Calculate the exact probability for each.

Today

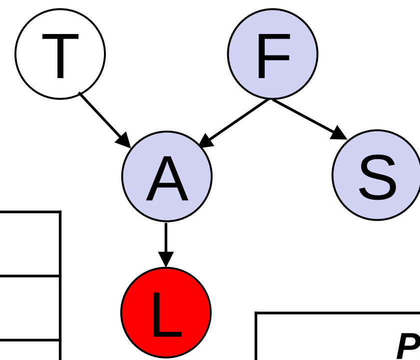
- How can we speed up (exact) inference with Bayes Nets?

1. Variable elimination algorithm
   - Sum up probabilities in an efficient way

2. Bayes Ball
   - Figure out which variables are conditionally independent given other variables

Inference in BNs: Leveraging structure

P(T)	
***T* = 1**	***T* = 0**
0.02	0.98



P(F)	
***F* = 1**	***F* = 0**
0.01	0.99

P(S | F)		
	***S* = 1**	***S* = 0**
***F* = 0**	0.01	0.99
***F* = 1**	0.90	0.10

P(A| T, F)		
	***A* = 1**	***A* = 0**
***T*=0, *F*=0**	0.0001	0.9999
***T*=0, *F*=1**	0.99	0.01
***T*=1, *F*=0**	0.85	0.15
***T*=1, *F*=1**	0.5	0.5

P(L | A)		
	***L* = 1**	***L* = 0**
***A* = 0**	0.001	0.999
***A* = 1**	0.88	0.12

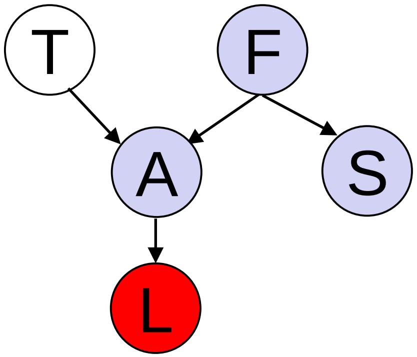
Recall: Entries in the full joint probability. e.g., ***P(~T, F, A, S, L)*** ?

P(~*T*, *F*, *A*, S, *L*) = *P(~T) P(F) P(A | ~T, F) P(S | F) P(L | A)*

= 0.98 x …

Say we want: P(*T* | *L*=1) = P(*T*, *L*=1) / P(*L*=1)

Inference in BNs: Leveraging structure



Naïve approach:

Start with

$P(T \mid L=1) = P(T, L=1) / P(L=1)$

$P(L=1) = P(T=1, L=1) + P(T=0, L=1)$

$P(T=1, L=1) = \sum_{a,s,f} P(A=a, S=s, F=f, T=1, L=1)$

$= \sum_{a,s,f} P(s \mid f)\, P(f)\, P(a \mid f, T=1)\, P(L=1 \mid a)\, P(T=1)$

$P(T=0, L=1) = \sum_{a,s,f} P(A=a, S=s, F=f, T=0, L=1)$

$= \sum_{a,s,f} P(s \mid f)\, P(f)\, P(a \mid f, T=0)\, P(L=1 \mid a)\, P(T=0)$

Inference in BNs: Leveraging structure

A better solution:

- Re-arrange the sums:

$$P(T, L=1) = \sum_{a,\, s,\, f} P(s \mid f)\, P(f)\, P(a \mid f, T)\, P(L=1 \mid a)\, P(T)$$

$$= \sum_{a,\, f} P(f)\, P(a \mid f, T)\, P(L=1 \mid a)\, P(T) \sum_{s} P(s \mid f)$$

Inference in BNs: Leveraging structure

A better solution:

- Re-arrange the sums:

$P(T, L=1) \quad = \sum_{a, s, f} P(s \mid f)\, P(f)\, P(a \mid f, T)\, P(L=1 \mid a)\, P(T)$

$\quad = \sum_{a, f} P(f)\, P(a \mid f, T)\, P(L=1 \mid a)\, P(T) \sum_{s} P(s \mid f)$

- Replace:

$\sum_{s} P(s \mid f) \quad = m_S(f)$ (Note that $m_S(f)=1$, but ignore for now.)

- Now we have:

$P(T, L=1) \quad = \sum_{a, f} P(f)\, P(a \mid f, T=1)\, P(L=1 \mid a)\, P(T=1)\, m_S(f)$

- Repeat with other hidden variables (A, F).

Instead of $O(2^n)$ factors, we have to sum over $O(2^k n)$ factors.

Basic idea of variable elimination

1. **Impose** an ordering over variables, with the **query variable** coming **last**.
2. **Maintain** a list of "factors", which depend on given variables.
3. **Sum** over variables in the **order** in which they appear in the list.
4. **Memorize** the result of **intermediate computations**.

This is a kind of dynamic programming.

A bit of notation

- Let X_i be an <u>evidence variable</u> with observed value x_i.
- Let the evidence potential be an indicator function:

  $\delta(x_j, x_i) = 1$ if and only if $x_j = x_i$.

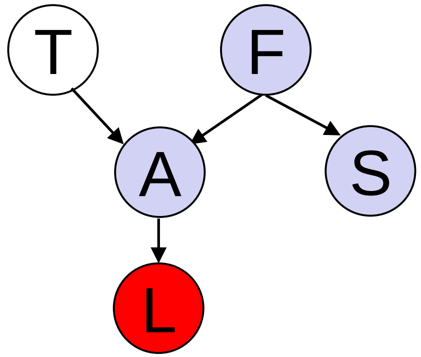
  This way we can turn conditionals into sums, e.g.

  $P(s \mid F=1) = \sum_f P(s \mid f)\, \delta(f, 1)$
- This is convenient for notation, but not effective as a practical implementation.

Variable elimination algorithm

1. Pick a variable ordering with Y (query variable) at the **end** of the list.
2. Initialize the active factor list with the conditional probability distributions (tables) in the Bayes net.
3. Add to the active factor list the evidence potentials $\delta(e_j, e_j)$ for all evidence variables E.
4. For $i=1..n$
   1. Take the next variable X_i from the ordering.
   2. Take all the factors that have X_i as an argument off the active factor list, and multiply them, then sum over all values of X_i, creating a new factor m_{Xi}
   3. Put m_{Xi} on the active factor list.

Example



To calculate:
P(*T* | *L*=1) =
P(*T*, *L*=1) / P(*L*=1)

1. Pick a variable ordering: *S, F, L, A, T*
2. Initialize the active factor list and introduce the evidence:

   List: *P(S|F), P(F), P(T), P(A|F,T), P(L|A),* $\boldsymbol{\delta(L,1)}$

3. Eliminate S: replace P(S|F) in list by computing

   $$m_s(F) = \sum_s P(s \mid F)$$

   List: ~~*P(S|F)*~~*, P(F), P(T), P(A|F,T), P(L|A),* $\delta(L,1)$, $\boldsymbol{m_s(F)}$

Example (cont'd)

List: $P(F), P(T), P(A|F,T), P(L|A), \delta(L,1), m_s(F)$

4. Eliminate F: $m_F(A,T) = \sum_f P(f)\, P(A|f,T)\, m_S(f)$

   List: $\cancel{P(F)}, P(T), \cancel{P(A|F,T)}, P(L|A), \delta(L,1), \cancel{m_S(F)}, \boldsymbol{m_F(A,T)}$

5. Eliminate L: $m_L(A) = \sum_l P(l|A)\, \delta(l,1)$

   List: $P(T), \cancel{P(L|A)}, \cancel{\delta(L,1)}, m_F(A,T), \boldsymbol{m_L(A)}$

6. Eliminate A: $m_A(T) = \sum_a m_F(a,T)\, m_L(a)$

   List: $P(T), \cancel{m_F(A,T)}, \cancel{m_L(A)}, \boldsymbol{m_A(T)}$

7. Compute answer for $T=0$ and $T=1$: $P(T=0)m_A(T=0)$

   $P(T=1)m_A(T=1)$

This is the answer we are looking for!

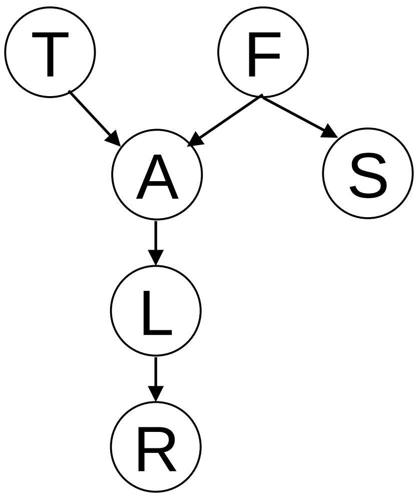
Complexity of variable elimination

- We need at most $O(n)$ multiplications to create an entry in a factor (where n is the total number of variables).
- If m is the maximum number of values that a variable can take, a factor depending on k variables will have $O(m^k)$ entries.

- So it is <u>important</u> to have <u>small factors</u>. But the size of the factors depends on the ordering of the variables!
- Choosing an optimal ordering is NP-complete.

Strategy #2: DAGs and independence

- Given a graph *G*, what independence assumptions are implied?



DAGs and independence

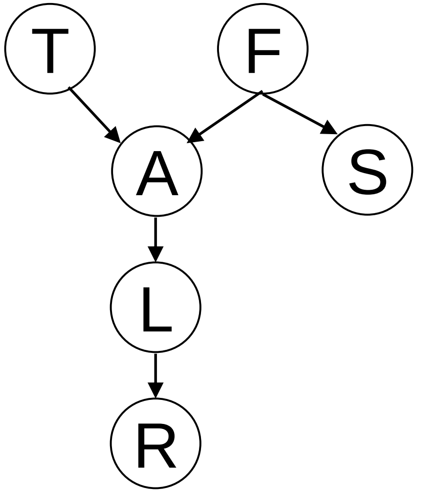
- Given a graph *G*, what independence assumptions are implied?
- In our example:
  - Variables were ordered: T, F, A, S, L, R.
  - Graph structure captures specific conditional independence relationships:

    $P(F|T) = P(F)$, i.e., $F \perp T$

    $P(S|T,F,A) = P(S|F)$, i.e., $S \perp \{T,A\} \mid F$

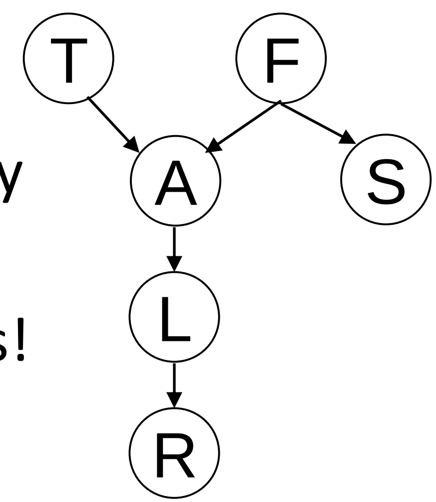
    $P(L|T,F,A,S) = P(L|A)$, i.e., $L \perp \{T,F,S\} \mid A$

    $P(R|T,F,A,S,L) = P(R|L)$, i.e., $R \perp \{T,F,A,S\} \mid L$
  - This notion of $\perp$ is called **d-separation** (directed separation)



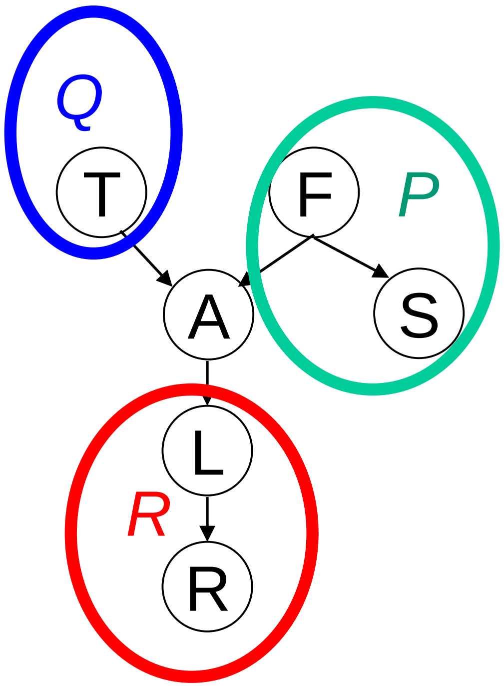
DAGs and independencies

- Absence of links implies certain conditional independence relationships.
  - Are there other independencies or conditional independencies between variables?
  - Are *A* and *S* independent?
  - Is *S* conditionally indep. of *L* and *R*, given *F*?
- What variables are independent, or conditionally independent, in general?
  - Can be read off the graph with the right tools!
  - Why do we care?   Faster inference!
    - E.g. Suppose we want to know *P(S | A)*
    - Do we need to sum over all values of *T, L, R*?
    - Skip variables that are conditionally indep. of *S* given *A*!



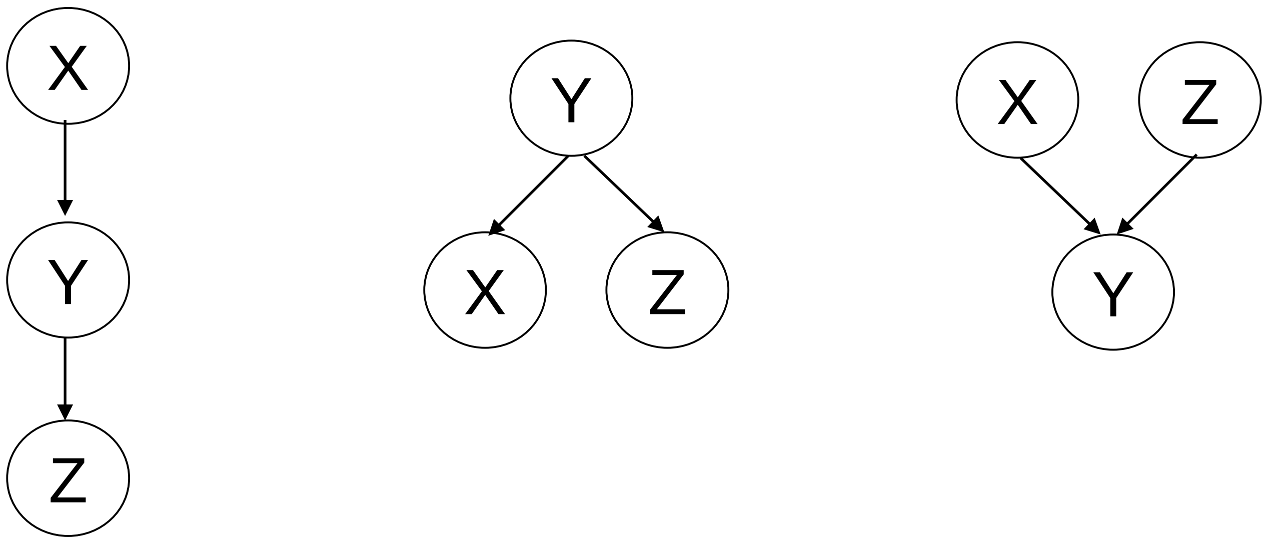
Implied independencies

- Given evidence for variable *Q*, what can we say about the *sets* of variables *P* and *R*?
  - If we get information about *P*, does this change our belief about *R*?



- If *P* and R are not directly connected, and having information about *P* gives information about *R*, then it must be because it gives ***information about the variables along the path*** from *P* to *R*.

Three types of connections



- The question to answer in each case:

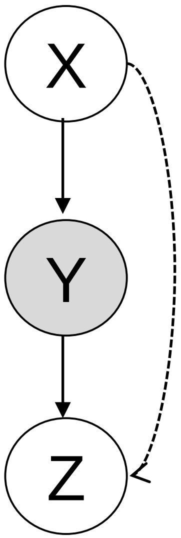
$X \perp Z$? X and Z independent?

$X \perp Z \mid Y$? X and Z conditionally independent given Y?

Indirect connection

- Interpret lack of edge between X and Z as <u>conditional independence</u>:

  $P(Z \mid X, Y) = P(Z \mid Y)$. Is this justified?



- Based on the graph structure, we have:

$$
\begin{aligned}
P(X, Y, Z) &= P(X)\, P(Y \mid X)\, P(Z \mid Y) \\
P(Z \mid X, Y) &= P(X, Y, Z) / P(X, Y) \\
&= P(X)\, P(Y \mid X)\, P(Z \mid Y) / P(X)\, P(Y \mid X) \\
&= P(Z \mid Y)
\end{aligned}
$$

- So Z is independent of X whenever the value of Y is known.

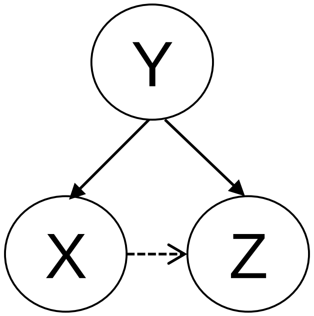
  E.g. X = Travel Y = Bird flu Z = Fever

- On the other hand, Z is not independent of X if Y not known. (Can check that $P(Z \mid X)$ does not simplify)

Common cause

- Again, interpret lack of edge between X and Z as conditional independence given Y. Is this true?

$$P(Z \mid X, Y) = P(X, Y, Z) / P(X, Y)$$
$$= P(Y)\, P(X \mid Y)\, P(Z \mid Y) / P(Y)\, P(X \mid Y)$$
$$= P(Z \mid Y)$$



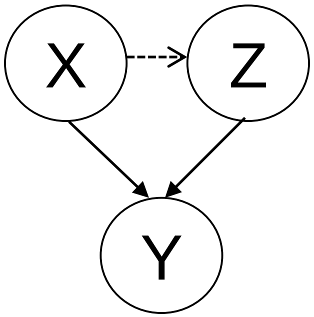
- This is a <u>hidden variable scenario</u>: if Y is unknown then X and Z can appear dependent on each other.

E.g. Y = Bronchitis X = Cough Z = Fever

- On the other hand, Z is not independent of X if Y not known. (can check that $P(Z \mid X)$ does not simplify)

V-structure

- Interpret lack of an edge between *X* and *Z* as statement of marginal independence.



- In this case, when given Y, X is not independent of *Z*.

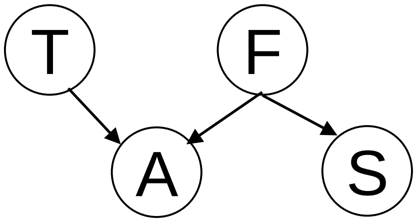
  (You can check that *P(Z | X, Y)* does not simplify.)

  - Same argument if *a child of Y* is given!

- This is a case of explaining away when there are multiple competing explanations.

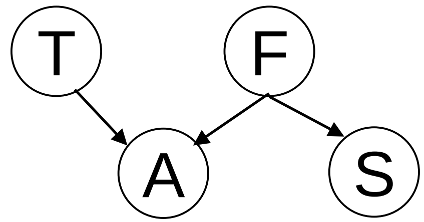
  E.g. X = Bird flu   Z = Cold   Y = Fever

Explaining away (Cont.)



- *Tampering* and *Fire* are independent variables (not the same as mutually exclusive! They can both be True, for example).

Explaining away (Cont.)

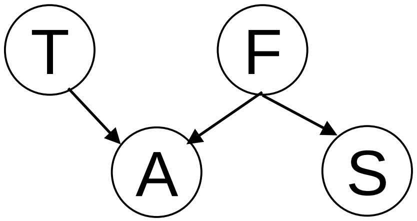


- *Tampering* and *Fire* are independent variables (not the same as mutually exclusive! They can both be True, for example).
- Agent hears the *Alarm* sound. It could be due to *Tampering* and/or *Fire*.

  $P(F \mid A) > P(F)$; $P(T \mid A) > P(T)$
- The agent gets some other piece of evidence, it sees *Smoke*.

  $P(F \mid A, S)$? $P(F \mid A)$

Explaining away (Cont.)



- *Tampering* and *Fire* are independent variables (not the same as mutually exclusive! They can both be True, for example).
- Agent hears the *Alarm* sound. It could be due to *Tampering* and/or *Fire*.

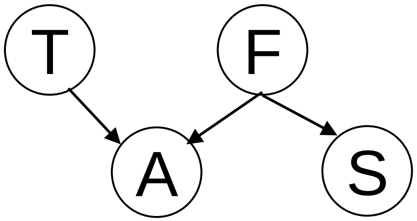
  $$P(F \mid A) > P(F) \; ; \; P(T \mid A) > P(T)$$

- The agent gets some other piece of evidence, it sees *Smoke*.

  $$P(F \mid A, S) \; > \; P(F \mid A)$$

  $$P(T \mid A, S) \; ? \; P(T \mid A)$$

Explaining away (Cont.)



- *Tampering* and *Fire* are independent variables (not the same as mutually exclusive! They can both be True, for example).
- Agent hears the *Alarm* sound. It could be due to *Tampering* and/or *Fire*.

  $P(F \mid A) > P(F)$; $P(T \mid A) > P(T)$

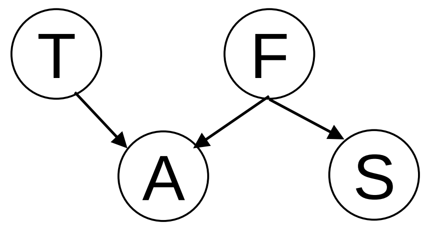
- The agent gets some other piece of evidence, it sees *Smoke*.

  $P(F \mid A, S) > P(F \mid A)$

  $P(T \mid A, S) < P(T \mid A)$

  $P(T \mid A, S)$? $P(T)$

Explaining away (Cont.)



- *Tampering* and *Fire* are independent variables (not the same as mutually exclusive! They can both be True, for example).
- Agent hears the *Alarm* sound. It could be due to *Tampering* and/or *Fire*.

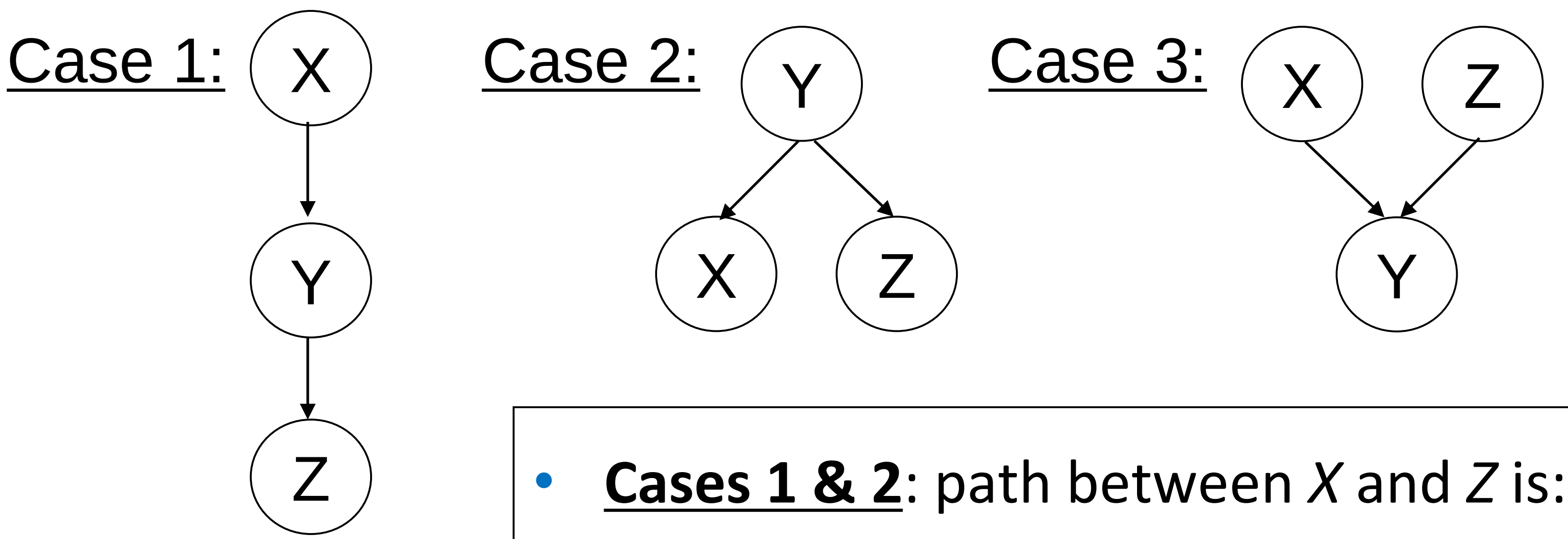
  $$P(F \mid A) > P(F) \; ; \; P(T \mid A) > P(T)$$

- The agent gets some other piece of evidence, it sees *Smoke*.

  $$P(F \mid A, S) > P(F \mid A)$$
  $$P(T \mid A, S) < P(T \mid A)$$
  $$P(T \mid A, S) > P(T)$$

- If instead there was some additional evidence directly related to *Tampering*, what happens to our belief about *Fire*, as compared to just having heard the *Alarm*?

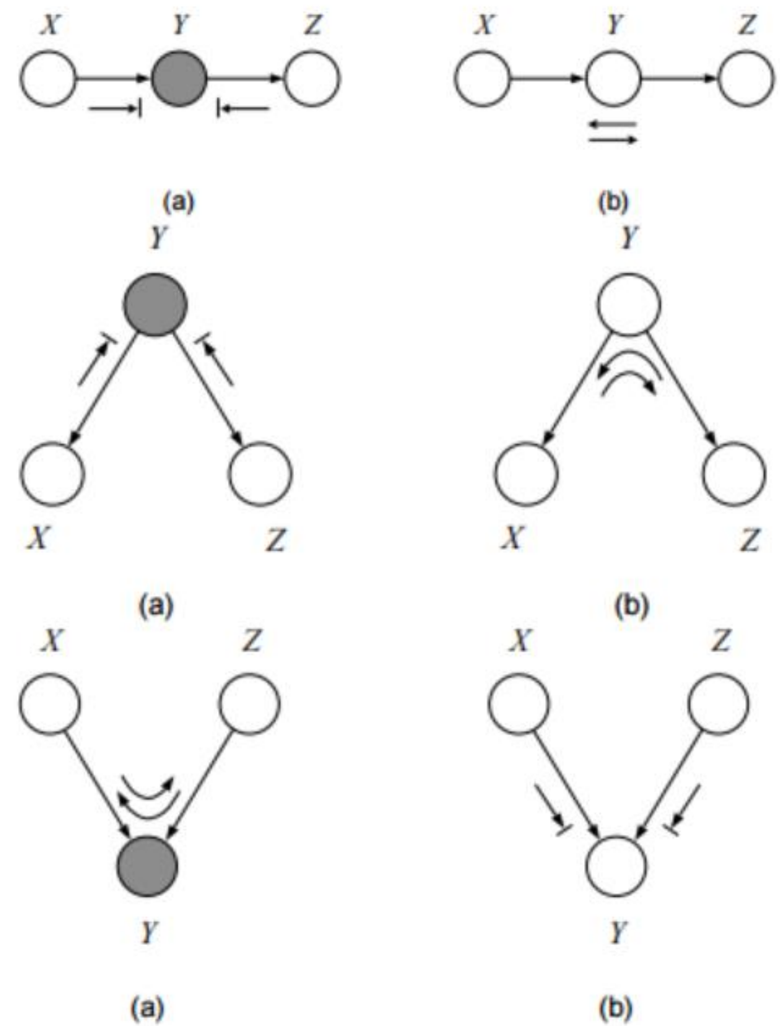
Summary of the three cases



- **Cases 1 & 2**: path between *X* and *Z* is:
  - open if *Y* is unknown.
  - blocked if *Y* is known.
- **Case 3**: path between *X* and *Z* is:
  - blocked if *Y* is unknown.
  - open if *Y* is known.

Bayes Ball algorithm

- Determine if $x_A \perp x_B \mid X_C$ by looking at structure of graph
  - Can we find a path of influence from x_A to x_B?
- Use this handy guide →
  - Shaded variable: the observed variables in X_C



https://www.cs.nyu.edu/~roweis/csc412-2004/notes/lec2x.pdf

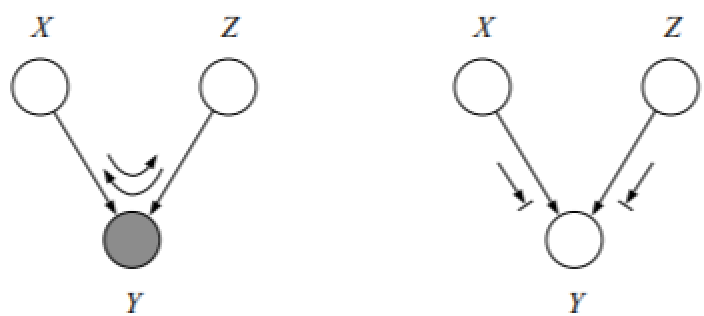
Boundary Cases and Explaining Away

- Boundary cases

(a) (b)

(a) (b)

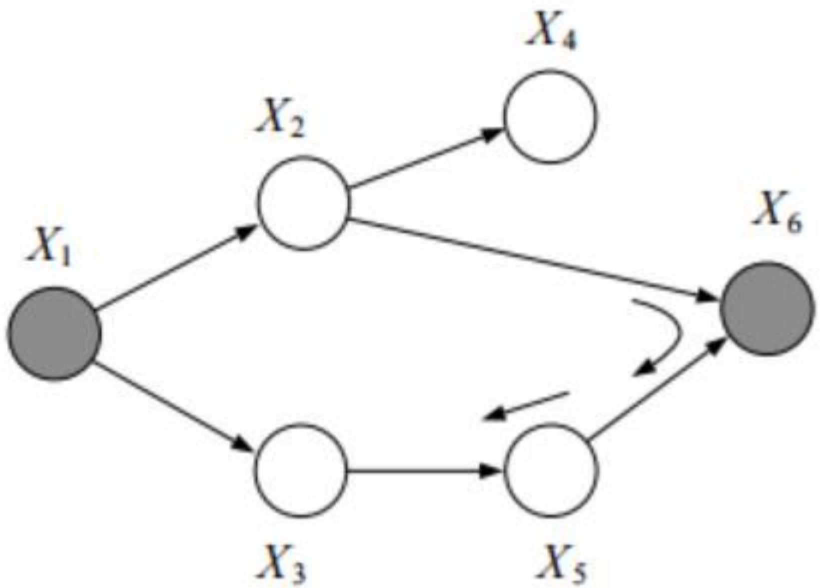
- Case b) above means that explaining away happens if *Y or any of its descendants is shaded*.



(a) (b)

Bayes Ball Example 1

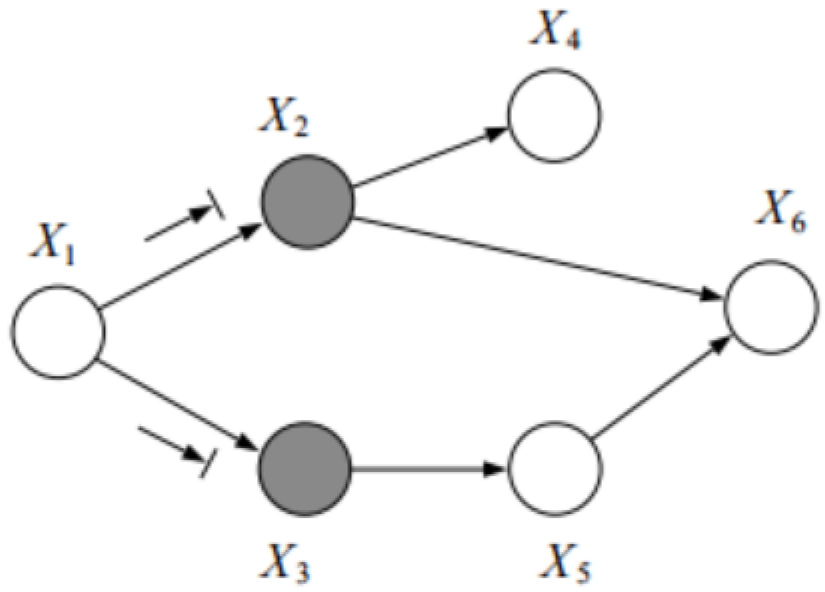
$$\mathbf{x}_2 \perp \mathbf{x}_3 | \{\mathbf{x}_1, \mathbf{x}_6\} \quad ?$$



Notice: balls can travel opposite to edge directions.

Bayes Ball Example 2

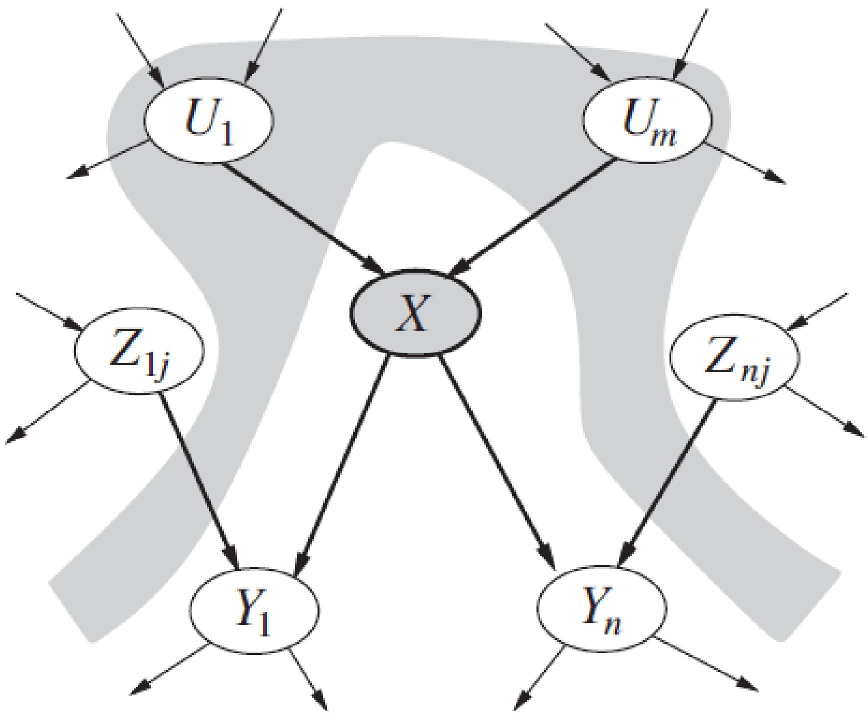
$$\mathbf{x}_1 \perp \mathbf{x}_6 | \{\mathbf{x}_2, \mathbf{x}_3\} \quad ?$$



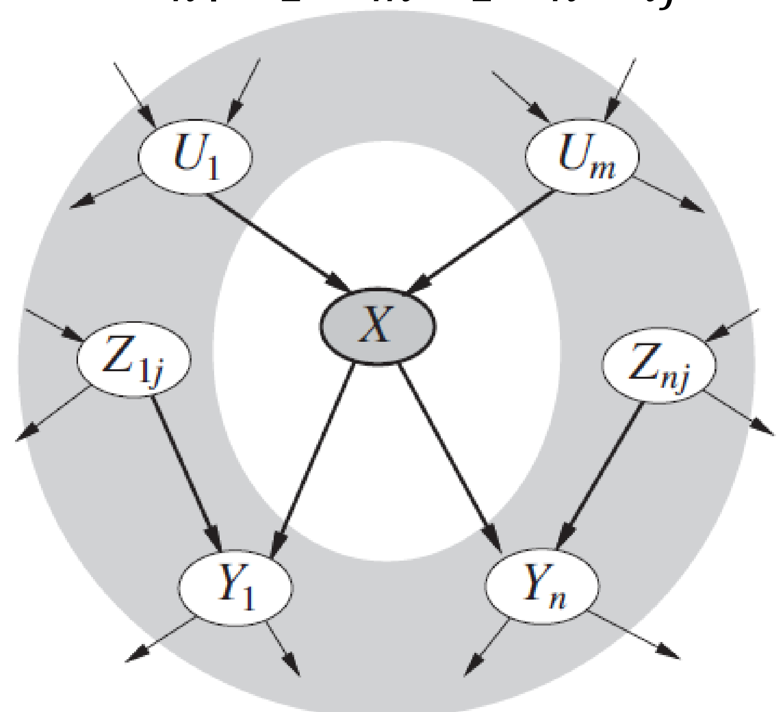
Bayes Ball

- Bayes Ball implies the two properties we discussed before

$X \perp Z_{ij} | U_1, U_m \quad \forall i$



$X \perp W_k | U_1, U_m, Y_1, Y_n, Z_{ij} \; \forall k$



Summary of Inference with Speed-Ups

Given query, $P(X_A|X_C)$ (where X_A, X_C are sets)

1. Use Bayes Ball algorithm to determine set X_B s.t. $X_A \perp X_B|X_c$
   - Look at graph structure
   - Shade in nodes X_C
   - Put a ball in each node X_A
   - See where the balls don't reach
2. Prune nodes X_B from graph
3. Start variable elimination algorithm

Summary of inference in Bayes nets

- Complexity of inference depends a lot on network's structure.
  - Inference is efficient (poly-time) for tree-structured networks.
  - In worse-case, inference is NP-complete.

- Best exact inference algorithm converts network to a tree, then does exact inference on the tree.

- In practice, for large nets, approximate inference methods work much better.

What you should know

- Basic rules of probabilistic inference.
- Variable elimination
  - Algorithm, complexity
- Independence in a Bayes net
  - Three base cases and the boundary leaf case.
  - Bayes ball algorithm