A Primer on Graphical Models

Almost completely built from materials of
Prof Noah Smith (CMU)
Prof David Sontag (NYU)
Prof Eric Xing (CMU)

General ML Strategy

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
2. **Learn** the distribution from data
3. Perform "**inference**" (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)
4. Compute "**Likelihood**" of observed data/variables $p(X_{i_1}, \ldots, X_{i_k})$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Example:

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \mathrm{Val}(X_i) = \{0, 1\}$$

- Let outcome space Ω be the cross-product of their states:

$$\Omega = \mathrm{Val}(X_1) \times \mathrm{Val}(X_2) \times \mathrm{Val}(X_3)$$

- $X_i(\omega)$ is the value for X_i in the assignment $\omega \in \Omega$
- Specify $p(\omega)$ for each outcome $\omega \in \Omega$ by a big table:

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	⋮		
1	1	1	.05

- How many parameters do we need to specify?

$$2^3 - 1$$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, $\quad \text{Val}(X) = \{\text{“Very High”}, \text{“High”}\}$
  Y: Grade, $\quad \text{Val}(Y) = \{\text{“a”}, \text{“b”}\}$
- Joint distribution specified by:

X

		vh	h
Y	a	0.7	0.15
	b	0.1	0.05

- $p(Y = a) = ? = 0.85$
- More generally, suppose we have a joint distribution $p(X_1, \ldots, X_n)$. Then,

$$p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \ldots, x_n)$$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Conditioning

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, $\quad \text{Val}(X) = \{\text{"Very High"}, \text{"High"}\}$
  Y: Grade, $\quad \text{Val}(Y) = \{\text{"a"}, \text{"b"}\}$

		X	
		vh	h
Y	a	0.7	0.15
	b	0.1	0.05

- Can compute the conditional probability

$$
\begin{aligned}
p(Y = a \mid X = vh) &= \frac{p(Y = a, X = vh)}{p(X = vh)} \\
&= \frac{p(Y = a, X = vh)}{p(Y = a, X = vh) + p(Y = b, X = vh)} \\
&= \frac{0.7}{0.7 + 0.1} \quad = 0.875.
\end{aligned}
$$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Example: Medical Diagonsis

- Variable for each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "pneumonia", "flu", "common cold", "bronchitis", "tuberculosis")
- Diagnosis is performed by **inference** in the model:

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- **Estimation** of joint distribution would require a huge amount of data
- **Inference** of conditional probabilities, e.g.

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

would require summing over exponentially many variables' values

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Structure through independence

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

- 2^n entries can be described by just n numbers (if $|Val(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable X_i cannot influence our predictions of X_j
- If $X_1, \ldots, X_n$ are *conditionally independent* given Y, denoted as $X_i \perp \mathbf{X}_{-i} \mid Y$, then

$$\begin{aligned} p(y, x_1, \ldots, x_n) &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y) \\ &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid y). \end{aligned}$$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Revisit: naïve Bayes for spam

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index the words in our vocabulary (e.g., English)
  - $X_i = 1$ if word i appears in an e-mail, and 0 otherwise
  - E-mails are drawn according to some distribution $p(Y, X_1, \ldots, X_n)$
- Suppose that the words are conditionally independent given Y. Then,

$$p(y, x_1, \ldots x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y)$$

Estimate the model with maximum likelihood. **Predict** with:

$$p(Y = 1 \mid x_1, \ldots x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i \mid Y = y)}$$

As an Aside

- Are the independence assumptions made here reasonable?

- Philosophy: Nearly all probabilistic models are “wrong”, but many are nonetheless useful

Observation

Any probability distribution $p(X_1, ..., X_n)$ can always be expressed as follows:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i | X_{C_i})$$

for some set $C_i \subset [n]$.

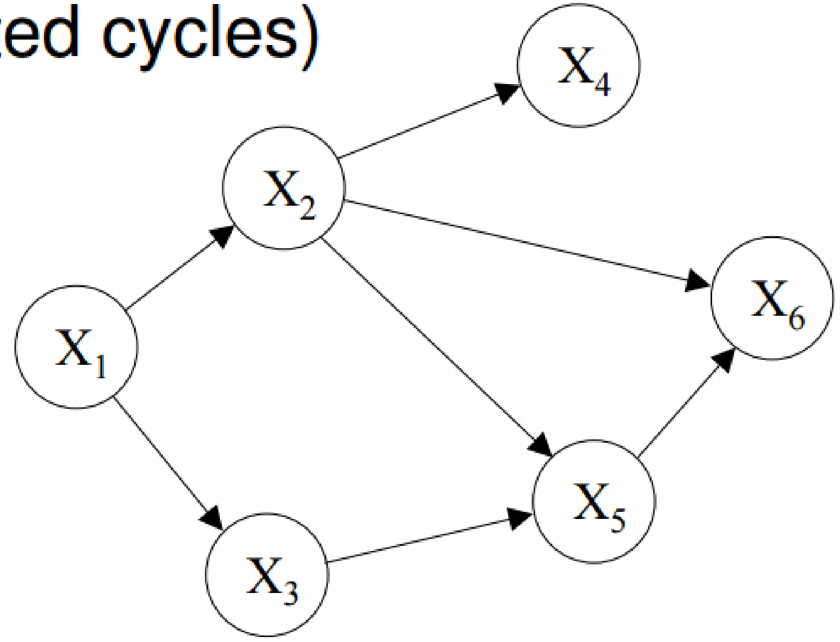
We call this set parent of i and henceforth denote by pa_i.

Visualization always helps!

Algebra is boring, so let's draw this

- Let's represent variables as circles
- Let's draw an arrow from j to i if $j \in pa_i$
- The resulting drawing will be a Directed Graph
- Moreover it will be Acyclic (no directed cycles)

Llatent variable / latent parameter	(var)
Observed variable	(obs)
Constant / hyper parameter	const



Bayesian Network

- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

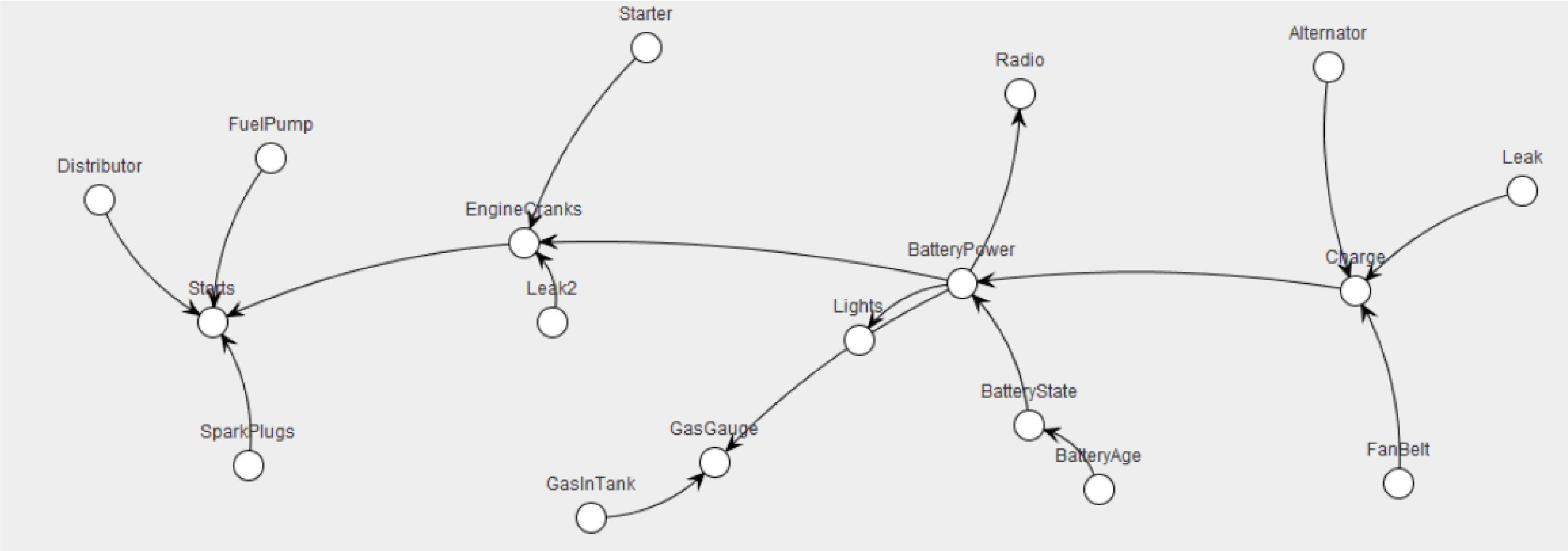
- Powerful framework for designing *algorithms* to perform probability computations

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

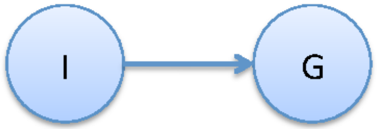
Will my car start this morning?



Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

Conditional Parameterization

- Grade is determined by Intelligence.



very high	0.85
high	0.15

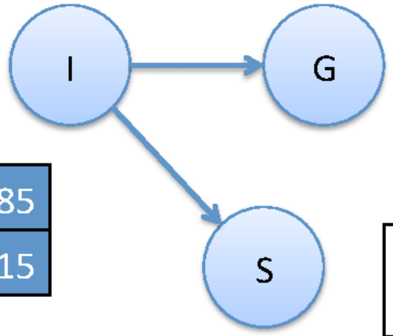
P(G|I)	very high	high
A	0.9	0.5
B	0.1	0.5

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Conditional Parameterization

- Grade and SAT score are determined by Intelligence: $G \perp S \mid I$

very high	0.85
high	0.15



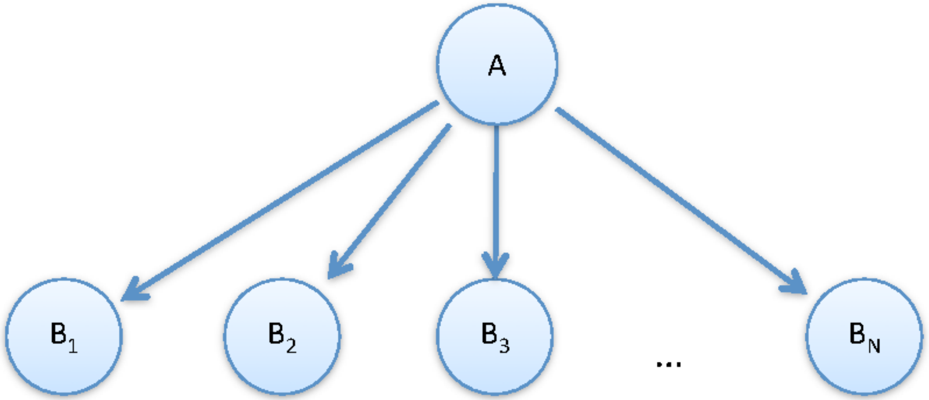
P(G|I)	very high	high
A	0.9	0.5
B	0.1	0.5

P(S|I)	very high	high
very high	0.7	0.4
high	0.3	0.6

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

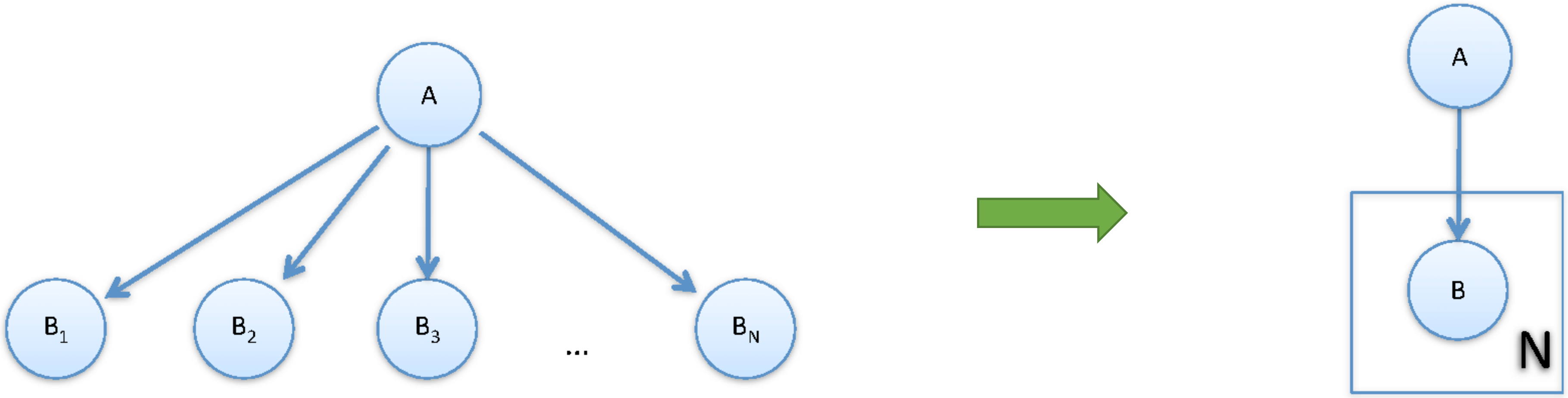
More drawing skills

- Drawing these figures can get messy for large models!
- How do we compactly represent repeated structure?



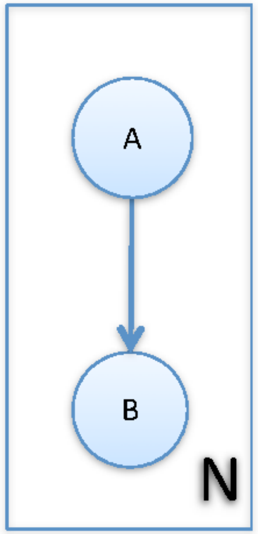
Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Plate Model



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Students and their Grades



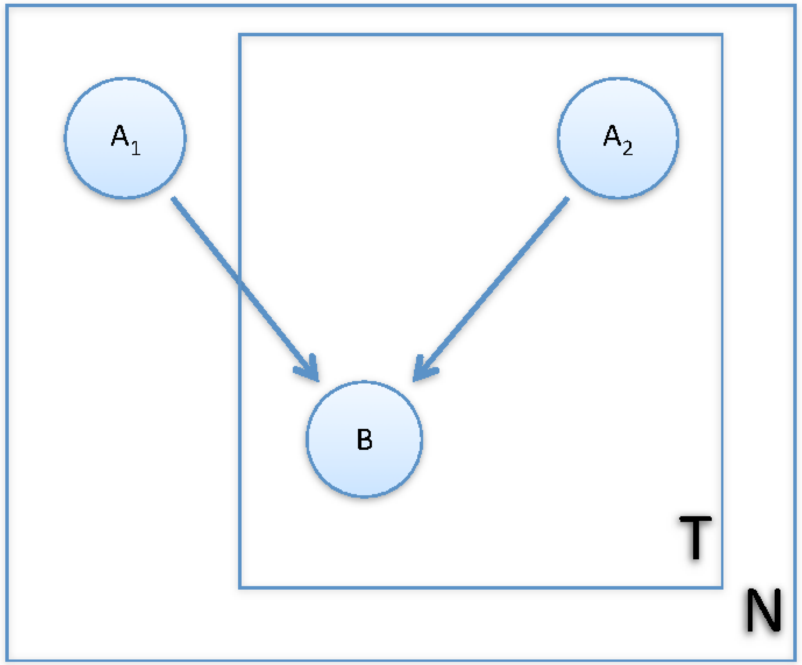
Example: A = student, B = grade

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Student, Course, Grade, Difficulty

Each student takes only one course

Nesting



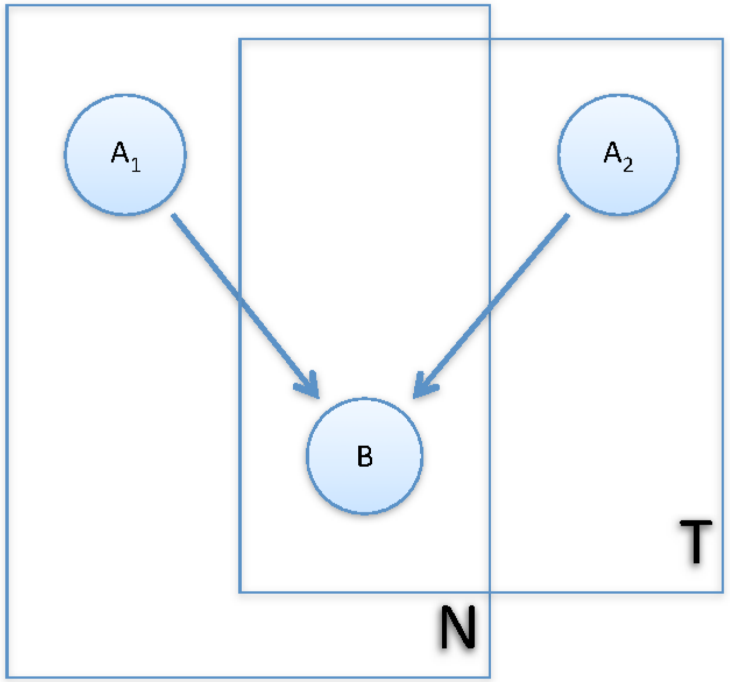
Example: A_1 = course difficulty, A_2 = student aptitude for the area, B = grade

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Student, Course, Grade, Difficulty

Multiple courses per student

Intersecting

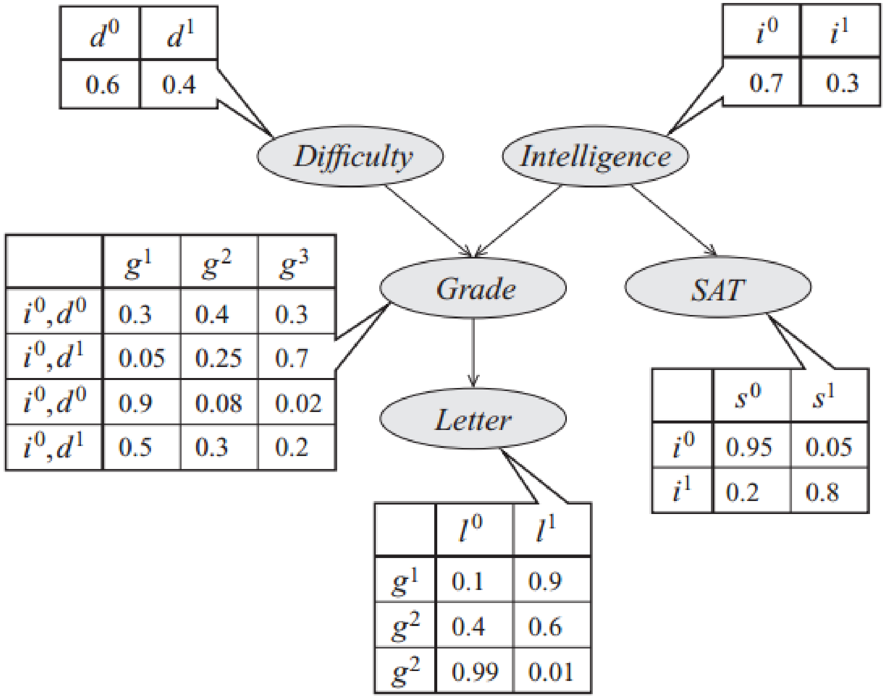


Example: A_1 = assignment difficulty, A_2 = intelligence, B = grade

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Detailed Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^0,d^0	0.9	0.08	0.02
i^0,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

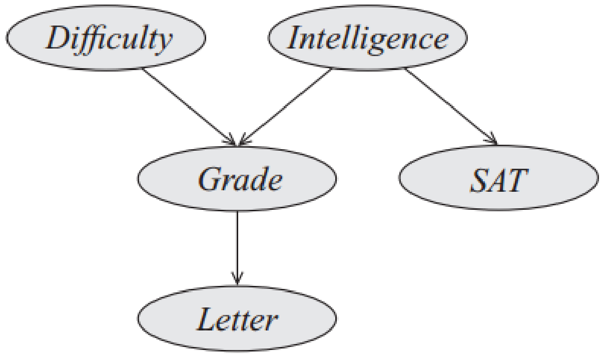
- What is its joint distribution?

$$p(x_1, \dots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Independencies



- The joint distribution corresponding to the above BN factors as

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

- However, by the chain rule, *any* distribution can be written as

$$p(d, i, g, s, l) = p(d)p(i \mid d)p(g \mid i, d)p(s \mid i, d, g)p(l \mid g, d, i, g, s)$$

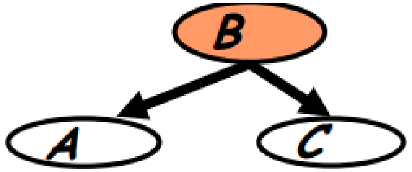
- Thus, we are assuming the following additional independencies:
  $D \perp I, \quad S \perp \{D, G\} \mid I, \quad L \perp \{I, D, S\} \mid G.$ What else?

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Generalizing

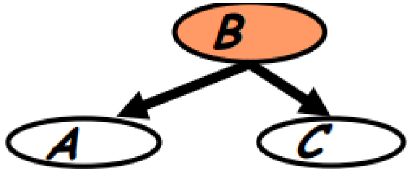
- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents

- **Common parent** – fixing B *decouples* A and C



Generalizing

- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents



- **Common parent** – fixing B *decouples* A and C

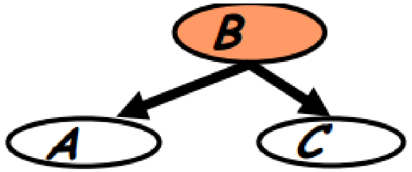
Proof: From the graph we have $p(A, B, C) = p(B)p(A|B)p(C|B)$. Now we can evaluate using Bayes rule as:

$$\begin{aligned} p(A, C|B) &= \frac{p(A, B, C)}{p(B)} \\ &= \frac{p(B)p(A|B)p(C|B)}{p(B)} \\ &= p(A|B)p(C|B) \end{aligned}$$

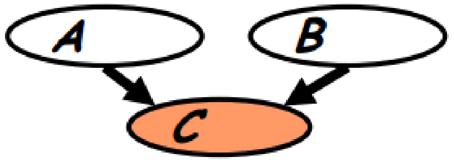
Thus showing the conditional independence.

Generalizing

- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents



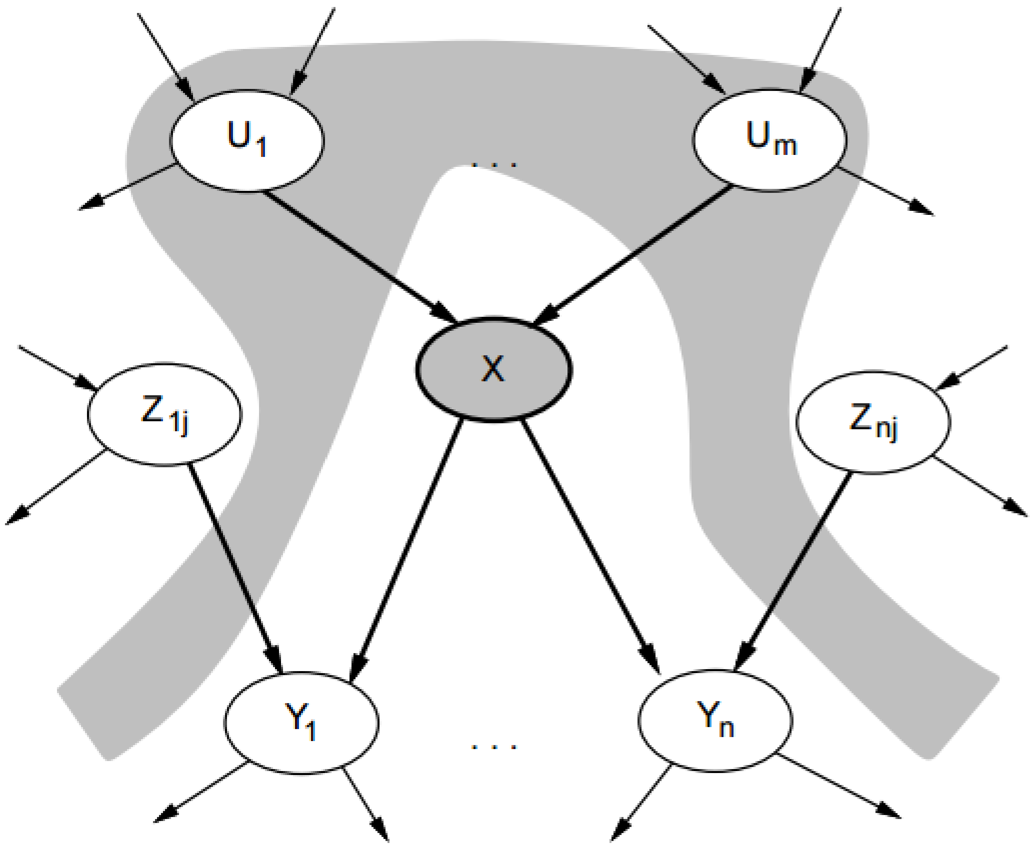
- **Common parent** – fixing B *decouples* A and C
- **Cascade** – knowing B *decouples* A and C



- **V-structure** – Knowing C *couples* A and B
  - This important phenomona is called **explaining away** and is what makes Bayesian networks so powerful

More properties

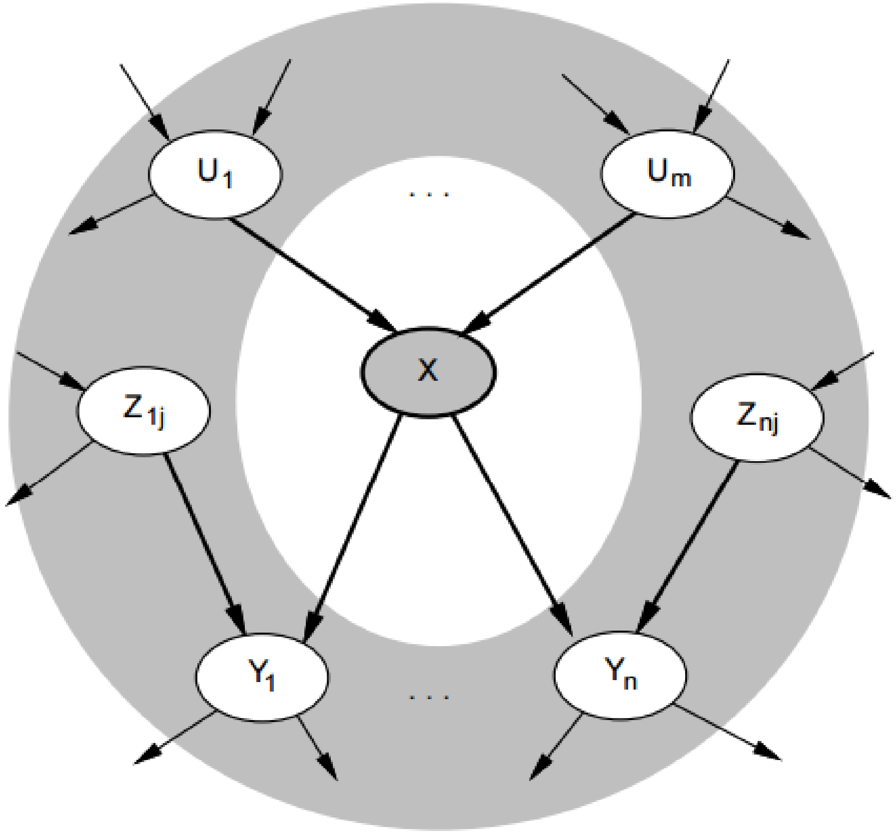
Local semantics: each node is conditionally independent of its nondescendants given its parents



Copied from: http://courses.cs.washington.edu/courses/cse515/15wi/slides/bnets.pdf

More Properties

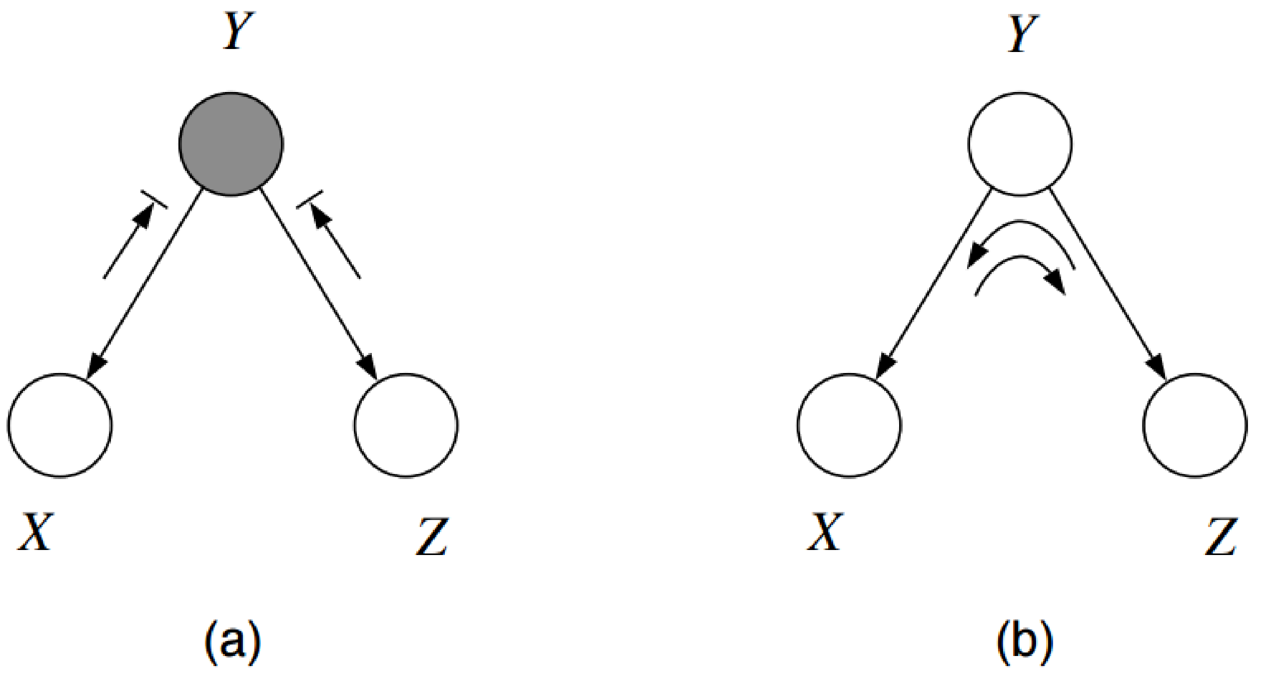
Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Copied from:
http://courses.cs.washington.edu/courses/cse515/15wi/slides/bnets.pdf

Bayes Ball

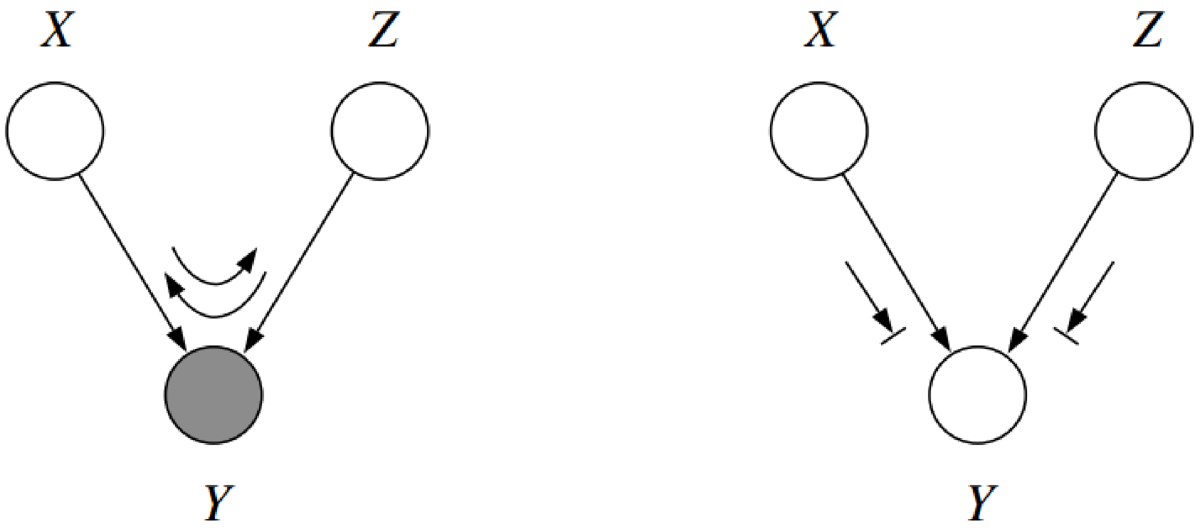
- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Bayes Ball

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:

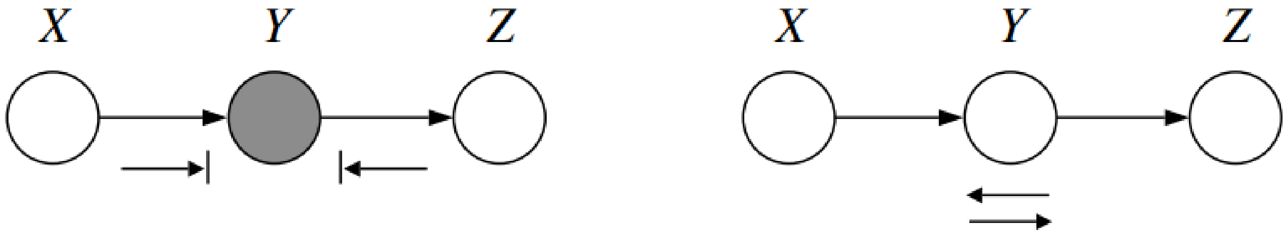


(a) (b)

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Bayes Ball

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



(a) (b)

Copied from: http://cs.nyu.edu/~dsontag/courses/pgm13/slides/lecture1.pdf

Bayes Ball

- If no such path, then X and Z are conditionally independent given Y
- This reduces statistical independencies (hard) queries to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

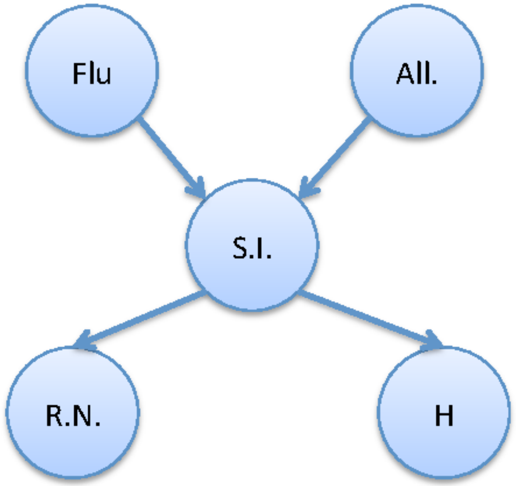
Causal Structure

- The flu causes sinus inflammation
- Allergies *also* cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Causal Structure

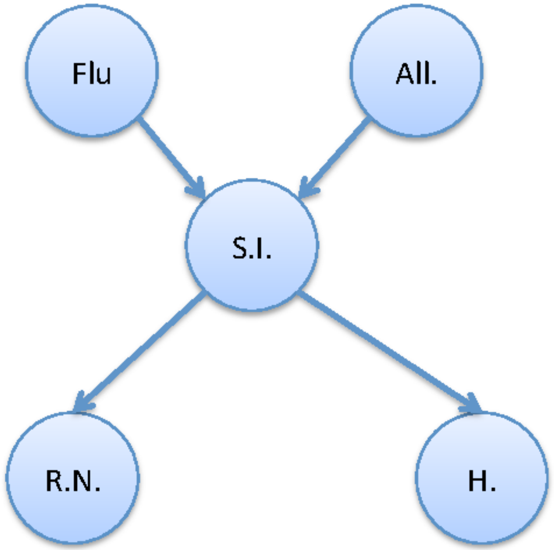
- The flu causes sinus inflammation
- Allergies *also* cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Querying the Model

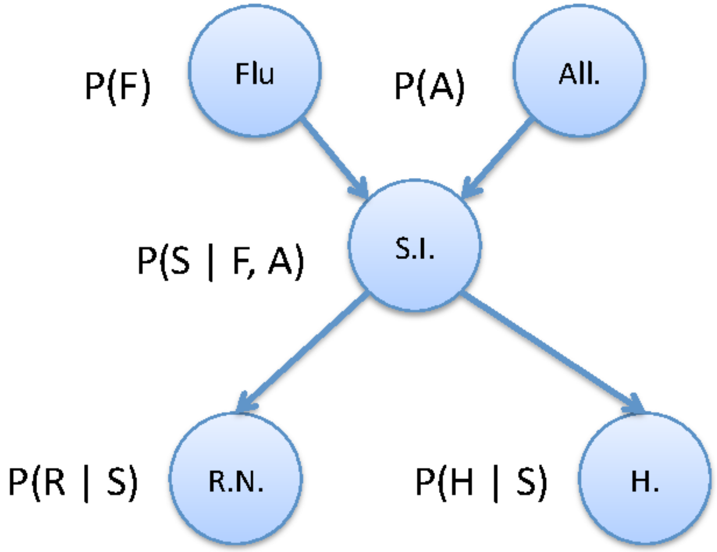
- Inference (e.g., do you have allergies?)
- What's the best explanation?
- Active data collection (what is the next best r.v. to observe?)



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Factored Joint Distribution

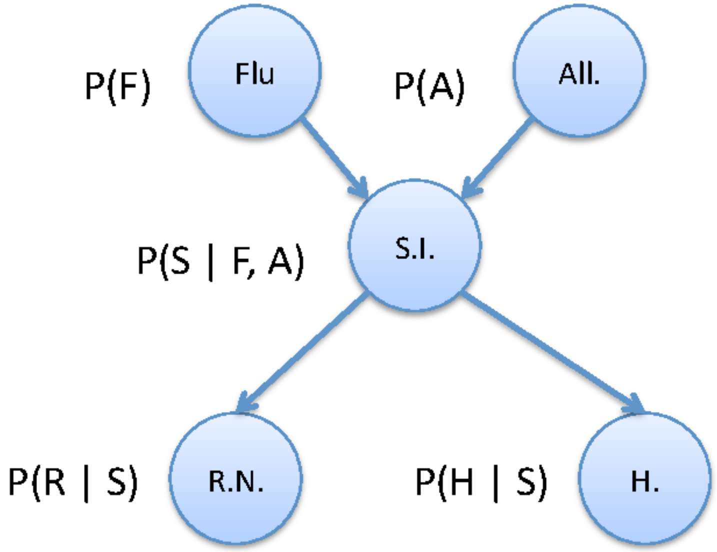
- Want:
  P(F, A, S, R, H)
  = P(F)
  P(A)
  P(S | F, A)
  P(R | S)
  P(H | S)



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Factored Joint Distribution

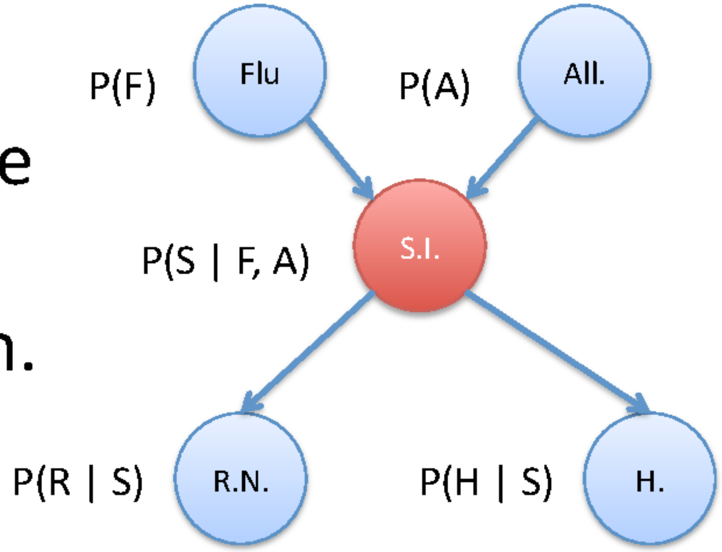
- Want:
  P(F, A, S, R, H)
  = P(F)
    P(A)
    P(S | F, A)
    P(R | S)
    P(H | S)
- How many parameters?



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Independence Assumptions

- Notice: knowing the value of S **separates** the other variables from each other in the graph.

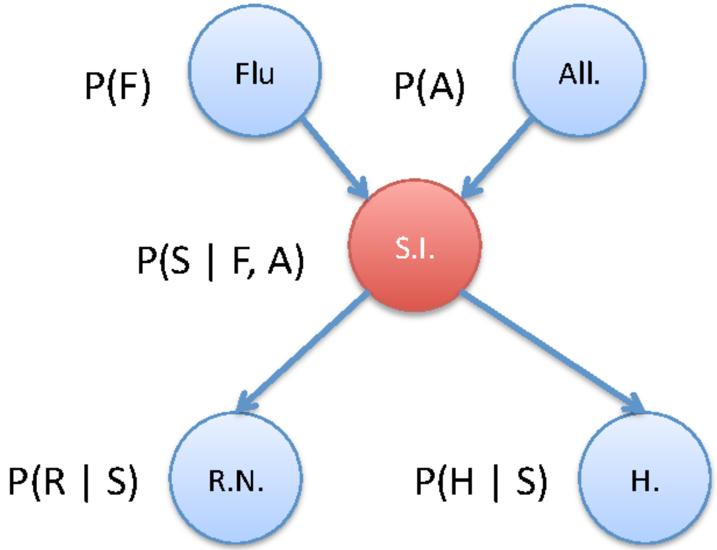


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Independence Assumptions

- In this model, ¬ R ⊥ H

1

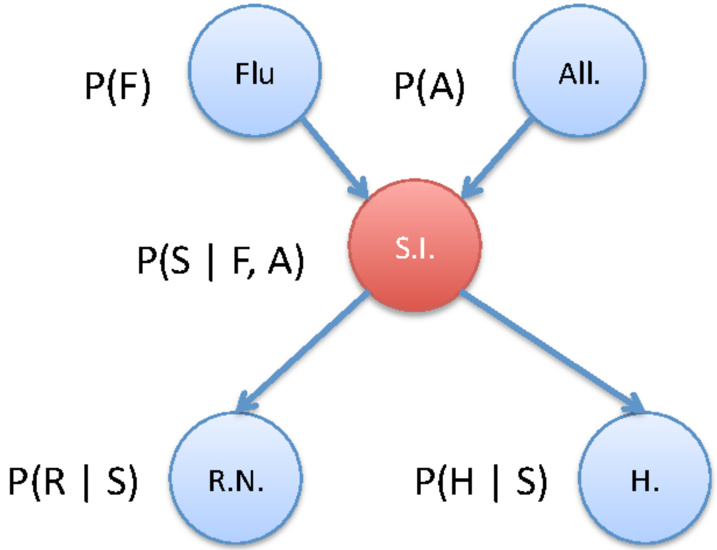


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Independence Assumptions

- In this model, $\neg\, R \perp H$
- But: $R \perp H \mid S$

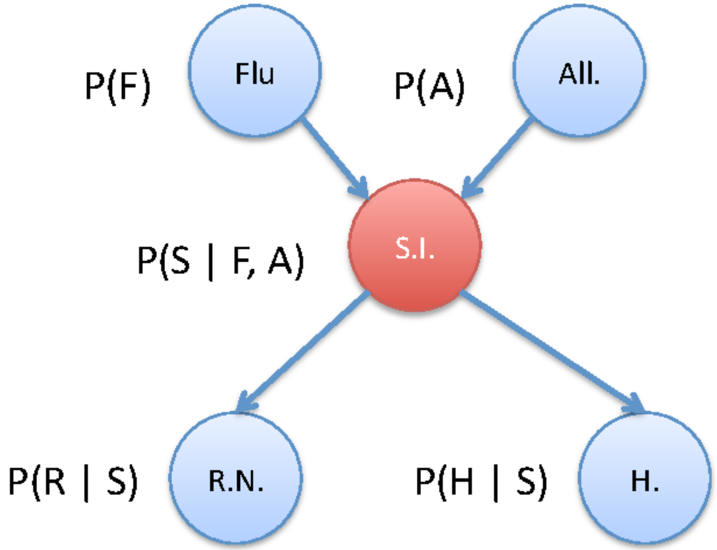
2



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Independence Assumptions

- In this model, $\neg\, R \perp H$
- But: $R \perp H \mid S$



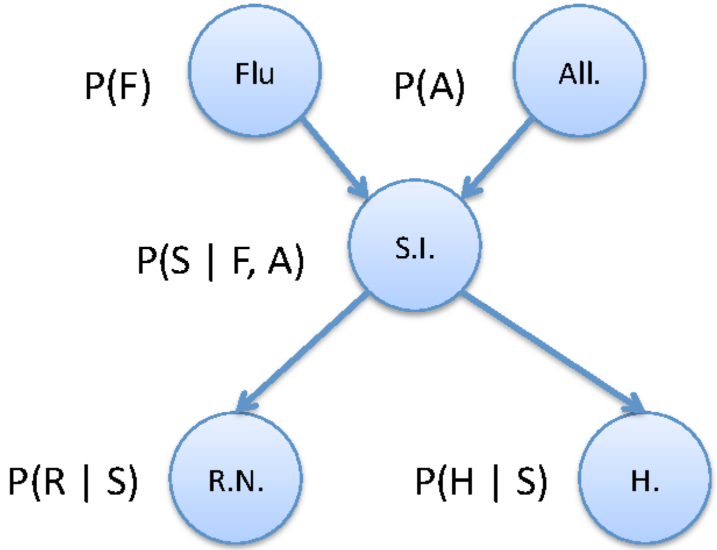
- Also: $\neg\, A \perp H$
- But: $A \perp H \mid S$

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Marginal Independence

- In this model, $F \perp A$
- $P(F, A) = P(F)\,P(A)$

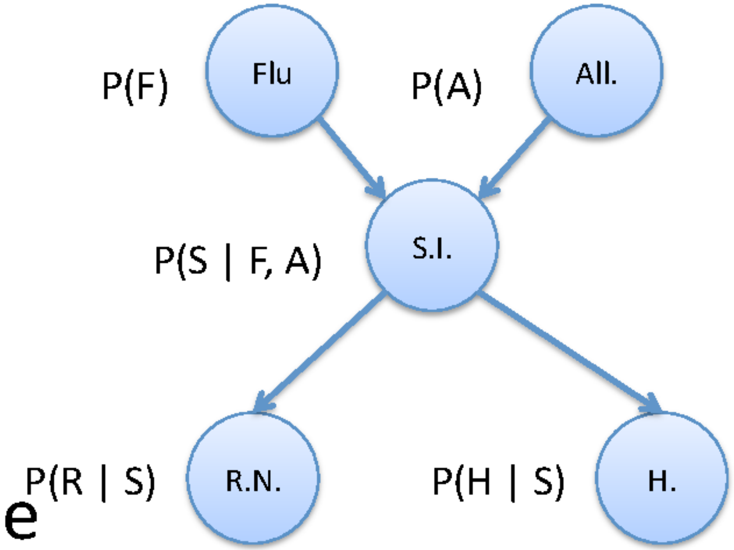
3



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Marginal Independence

- In this model, $F \perp A$
- $P(F, A) = P(F)\ P(A)$

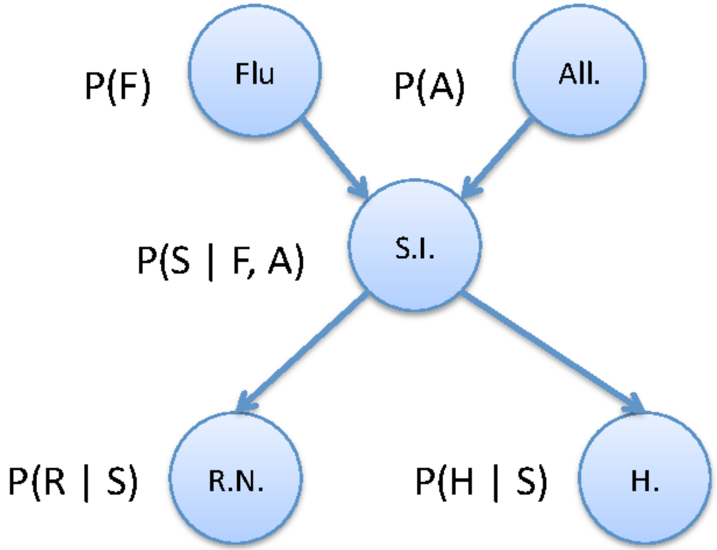


- Marginal independence of a set: $\forall \mathbf{Y} \subseteq \mathbf{X}, \mathbf{Z} \subseteq \mathbf{X}, \mathbf{Y} \perp \mathbf{Z}$
- Let $\mathbf{X} = \{A, F\}$

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Conditional Independence

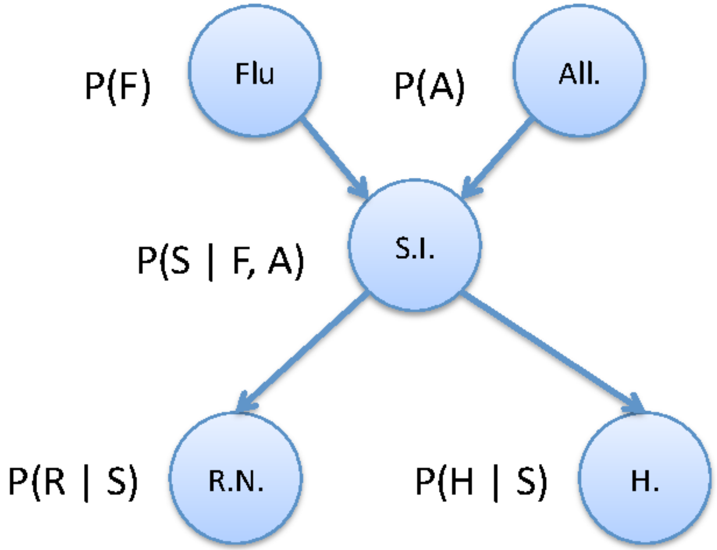
- In this model, $\neg F \perp H$
- $P(F, H) = P(F \mid H)\, P(H)$ $\neq P(F)\, P(H)$, in general



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Conditional Independence

- In this model, $\neg F \perp H$
- $P(F, H) = P(F \mid H)\, P(H)$
  $\neq P(F)\, P(H)$, in general



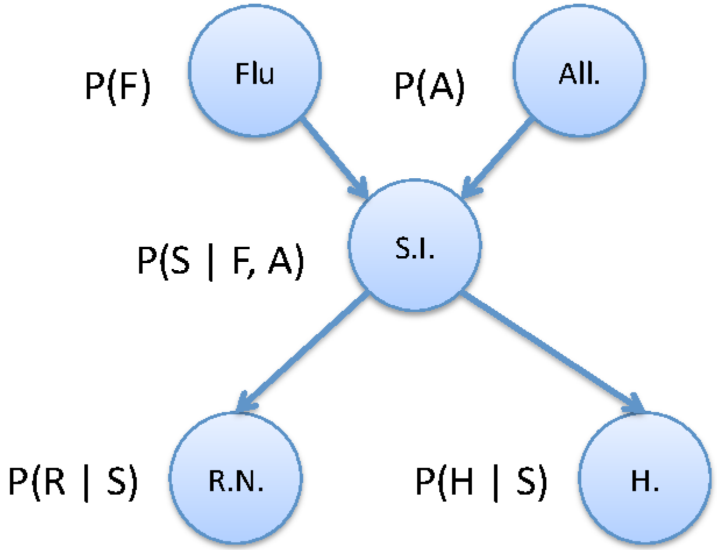
Given S, however …

- $F \perp H \mid S$
- $P(F, H \mid S) = P(F \mid S)\, P(H \mid S)$
  - How do we know this?

Copied from:
https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

What’s Independent?

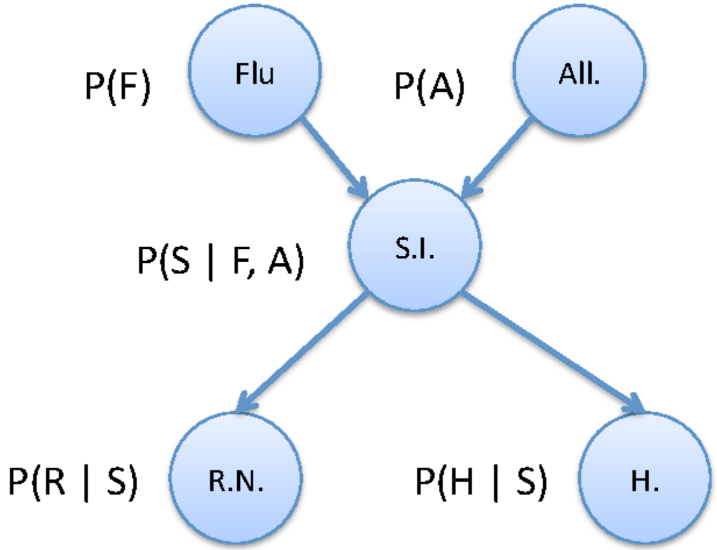
- $F \perp A \mid \varnothing$



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$

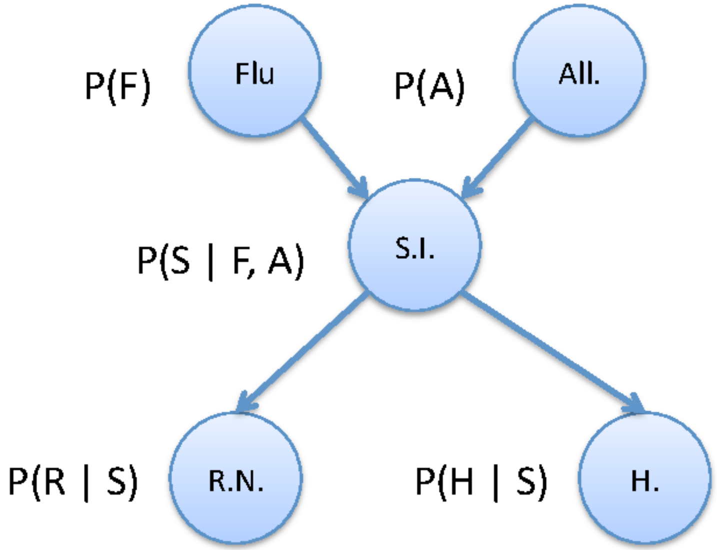


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$

- $R \perp \{F, A, H\} \mid S$

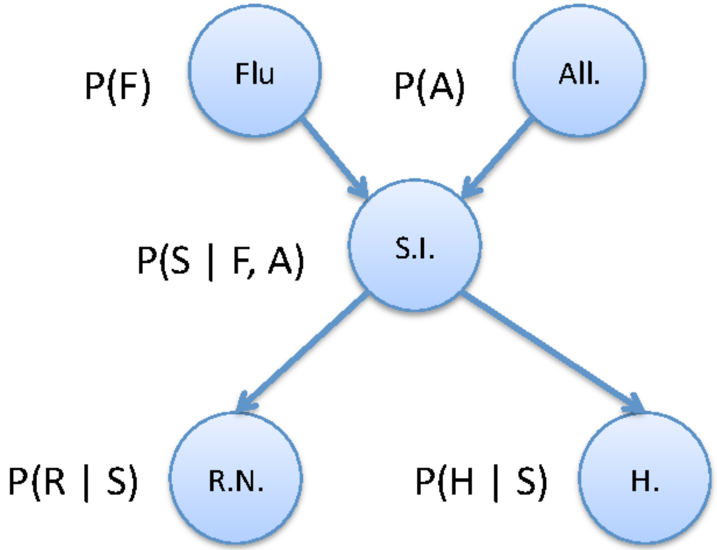


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

What’s Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$

- $R \perp \{F, A, H\} \mid S$
- $H \perp \{F, A, R\} \mid S$

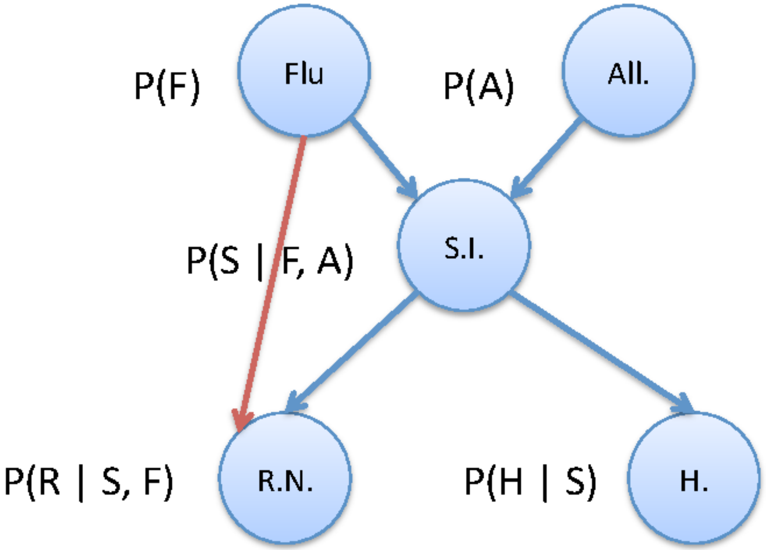


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

New Edge: What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$

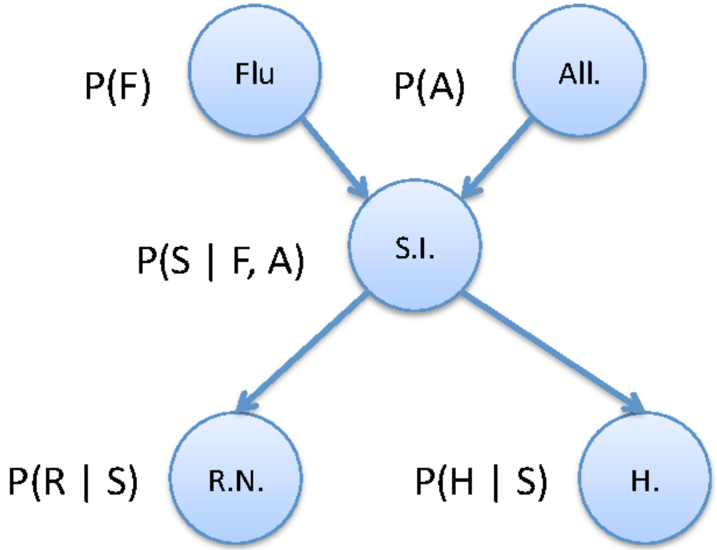
- $R \perp \{$~~F,~~ $A, H\} \mid S, F$
- $H \perp \{F, A, R\} \mid S$



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A Puzzle

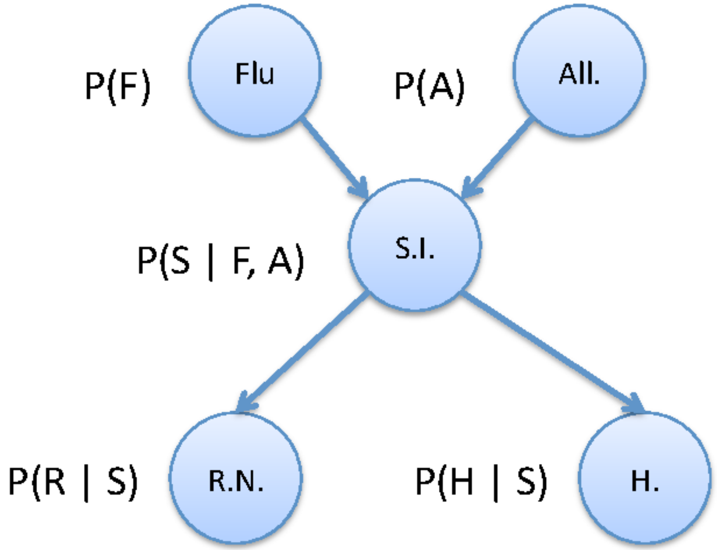
- $F \perp A \mid S$?



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

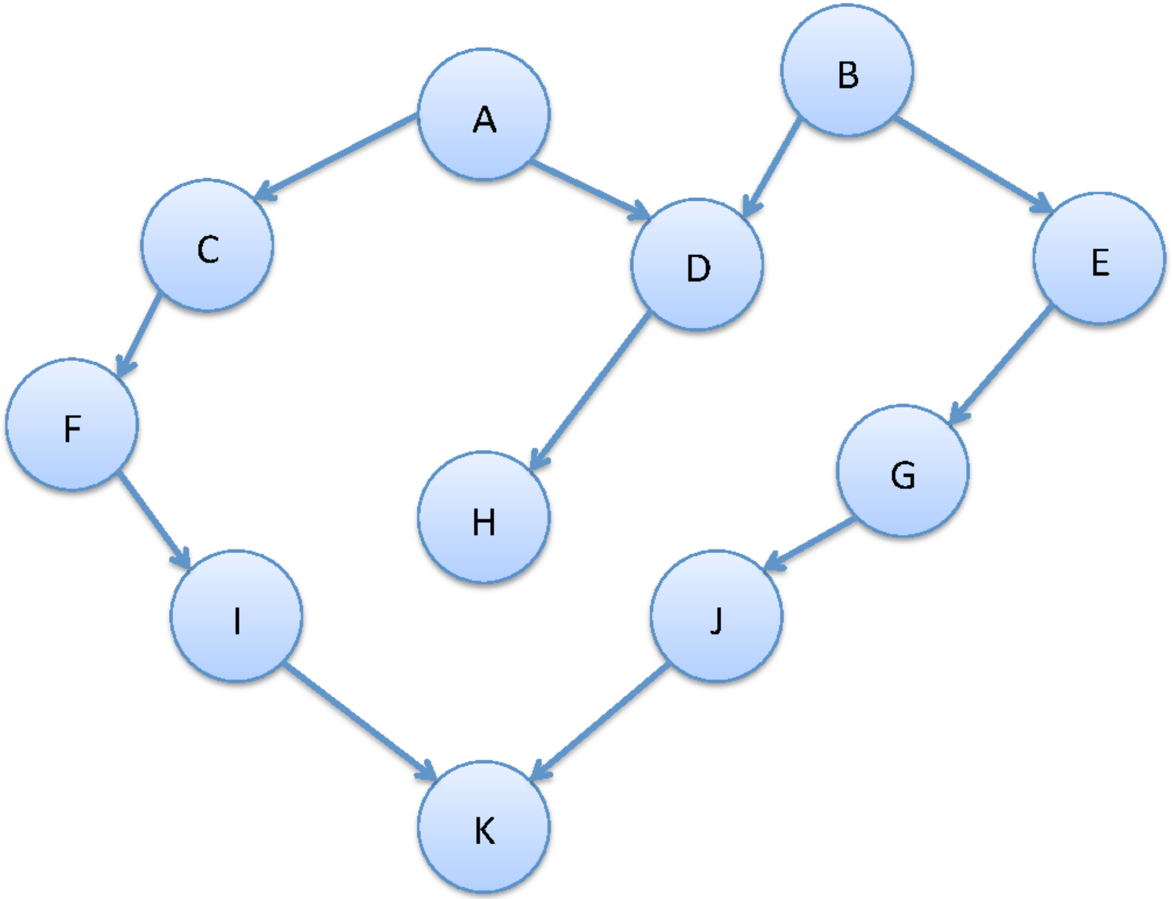
A Puzzle

- F ⊥ A | S ?
- In general, **no**.
  - This independence statement does not follow from the Local Markov assumption.
- ¬ (F ⊥ A | S)



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example

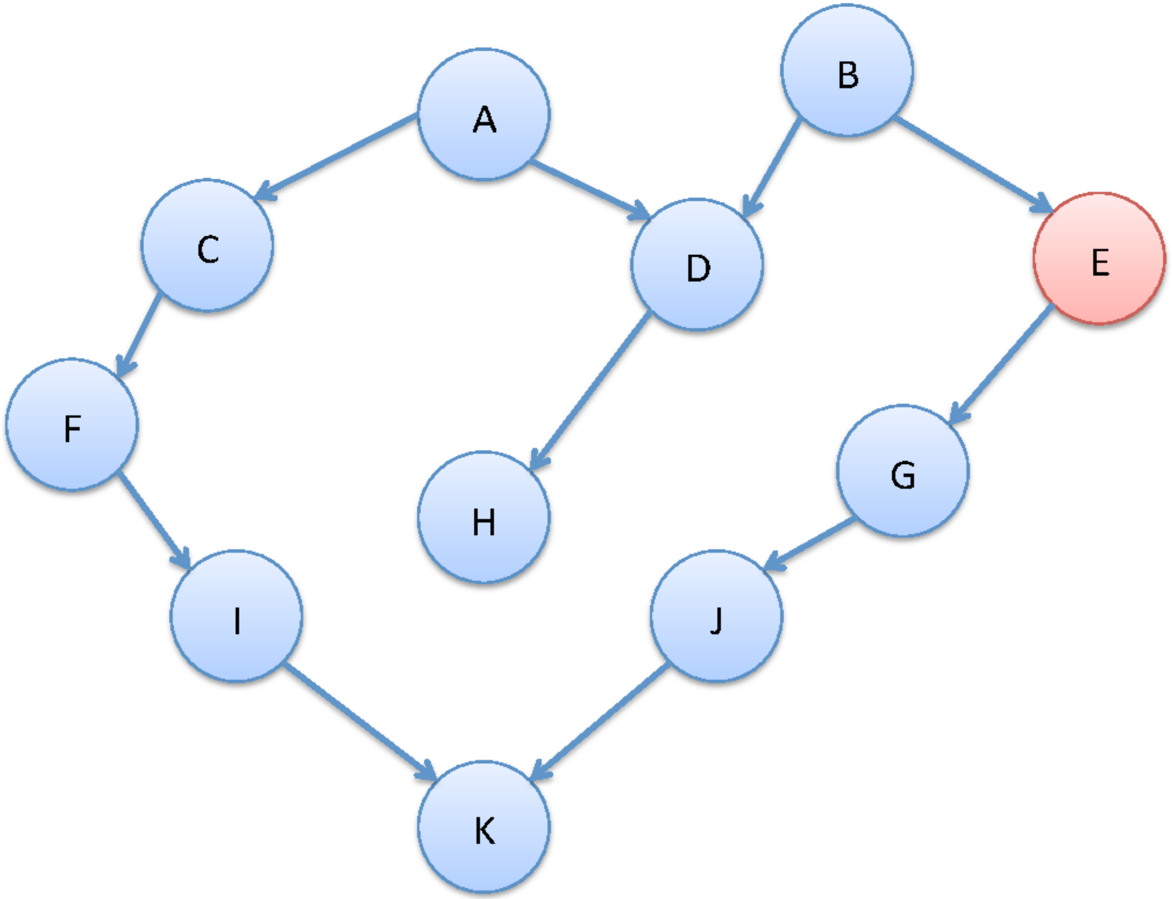


Easy:

- (A ⊥ B)

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



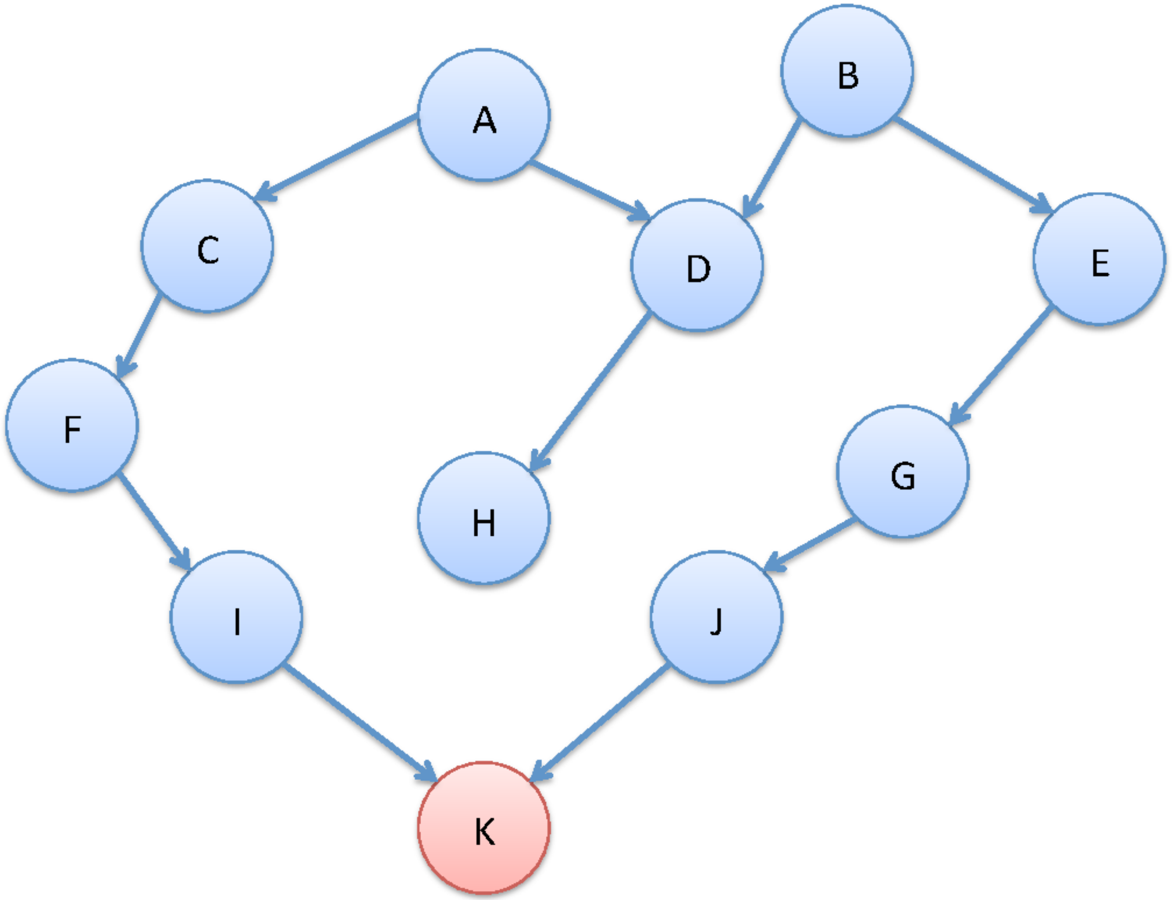
Easy:

- (B ⊥ G | E)

Local Markov Assumption for G!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



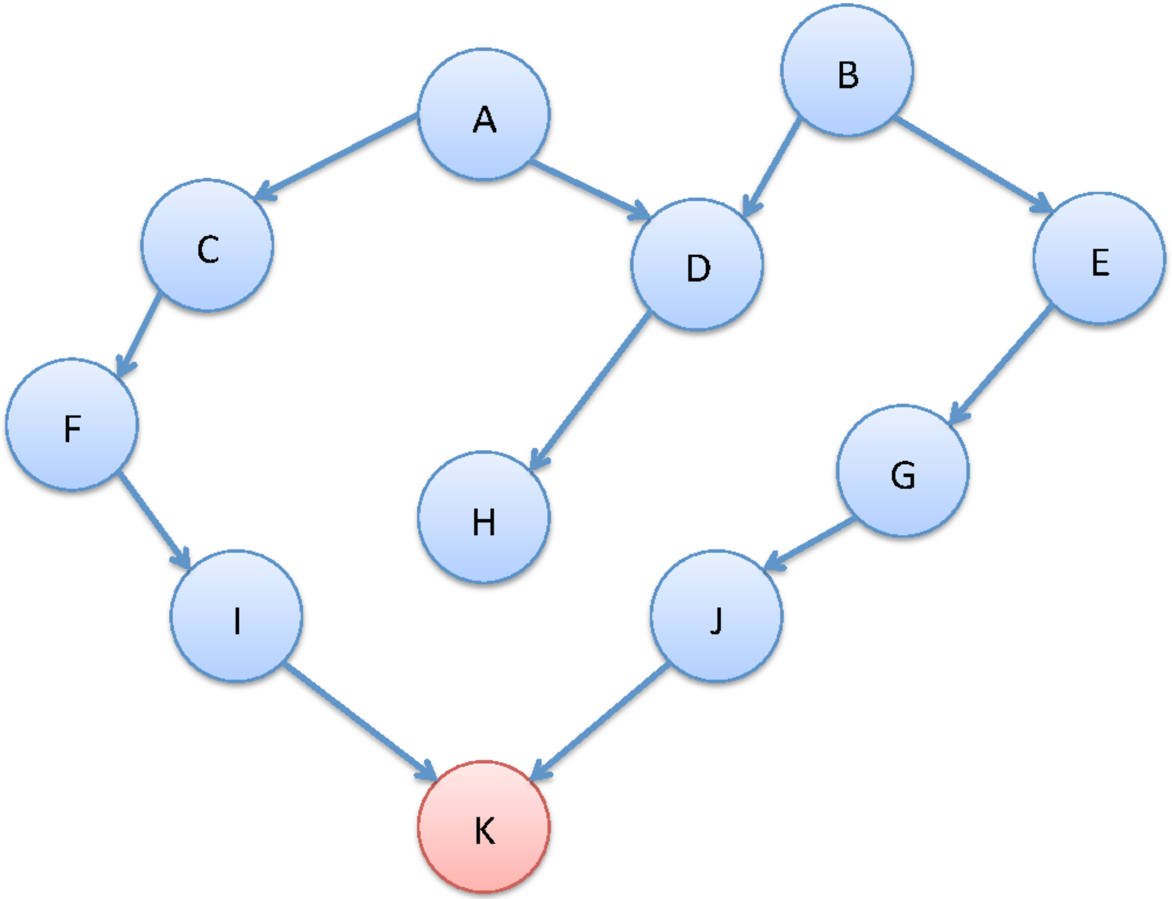
Easy:

- $\neg(I \perp J \mid K)$

V-Structure!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



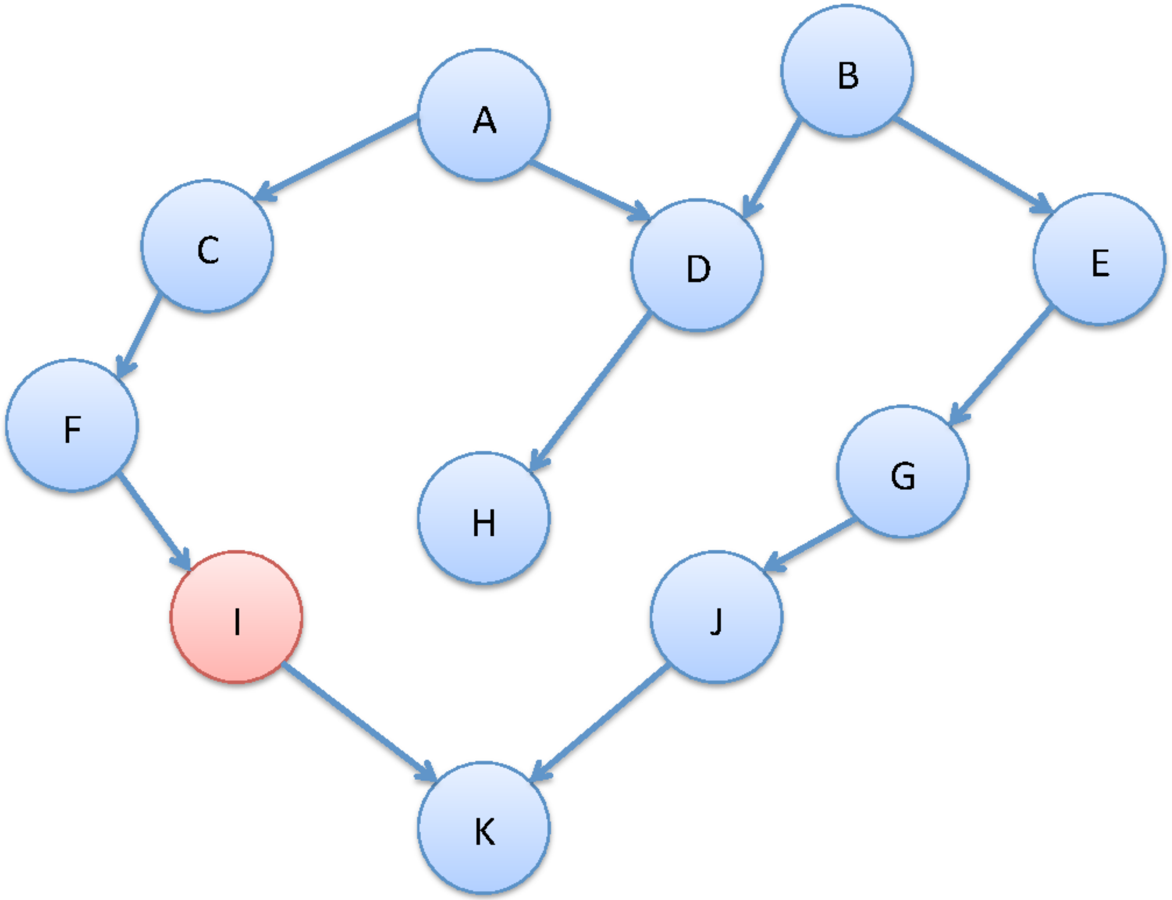
Harder:

- $\neg(E \perp F \mid K)$

"Flow of influence" along chains

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Harder:

- ¬(E ⊥ F | K)

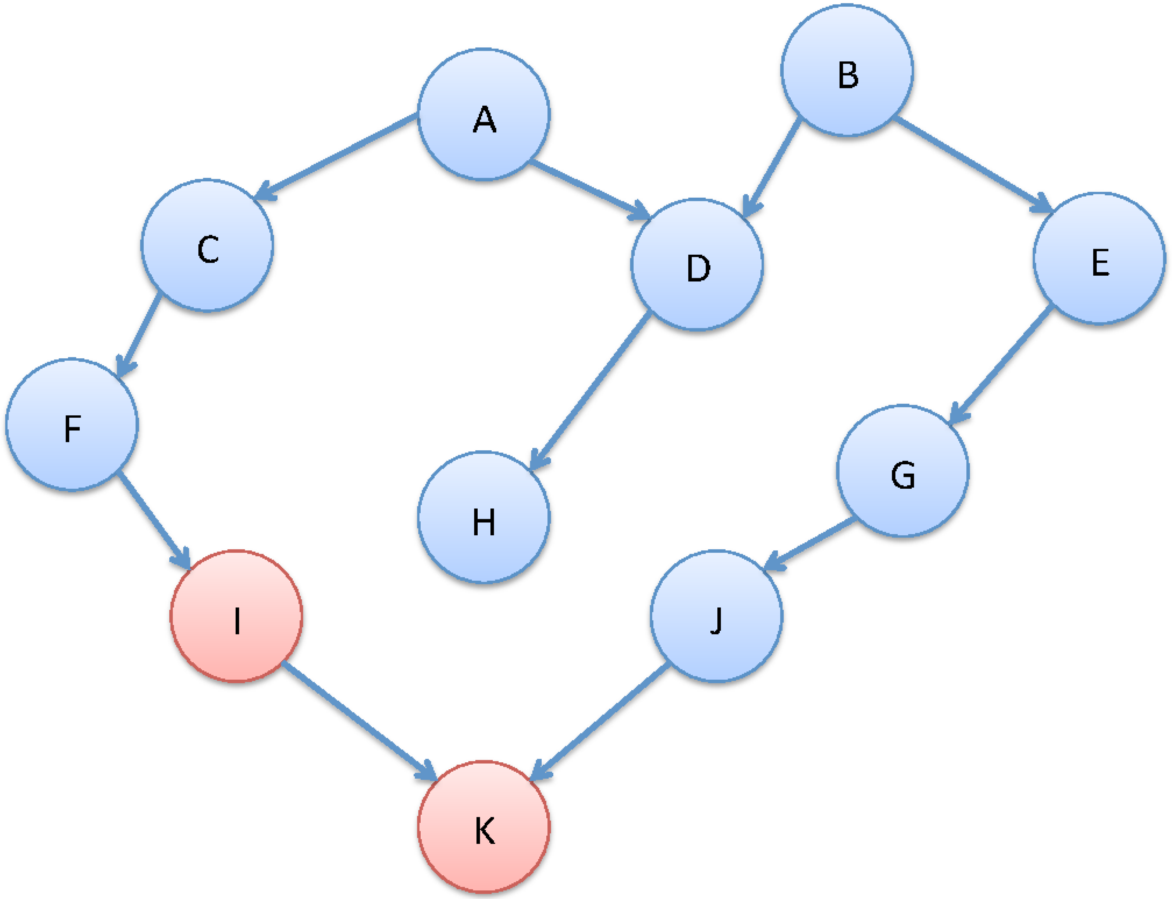
Easy:

- (F ⊥ K | I)

Local Markov Assumption for K!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Harder:

- $\neg(E \perp F \mid K)$

Easy:

- $(F \perp K \mid I)$

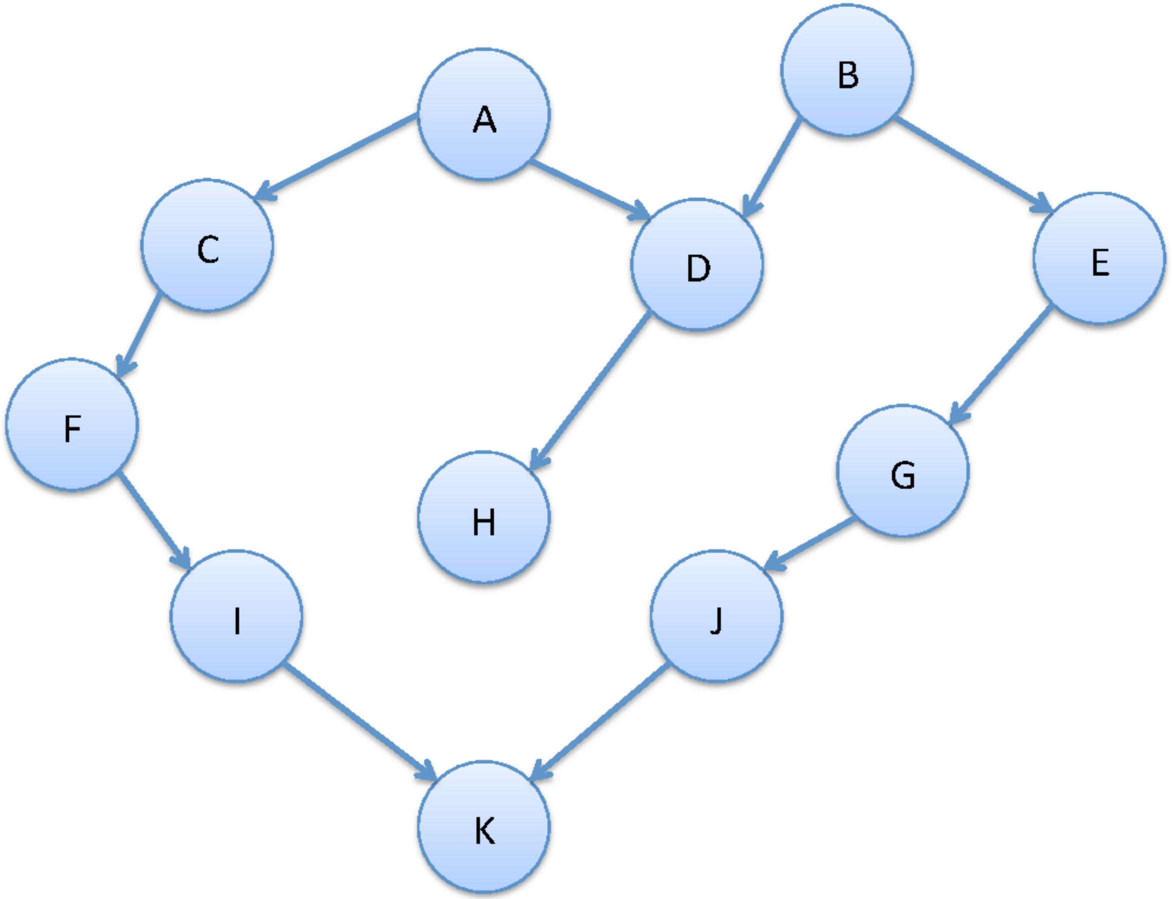
Claim:

- $(E \perp F \mid I, K)$

Interaction?

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



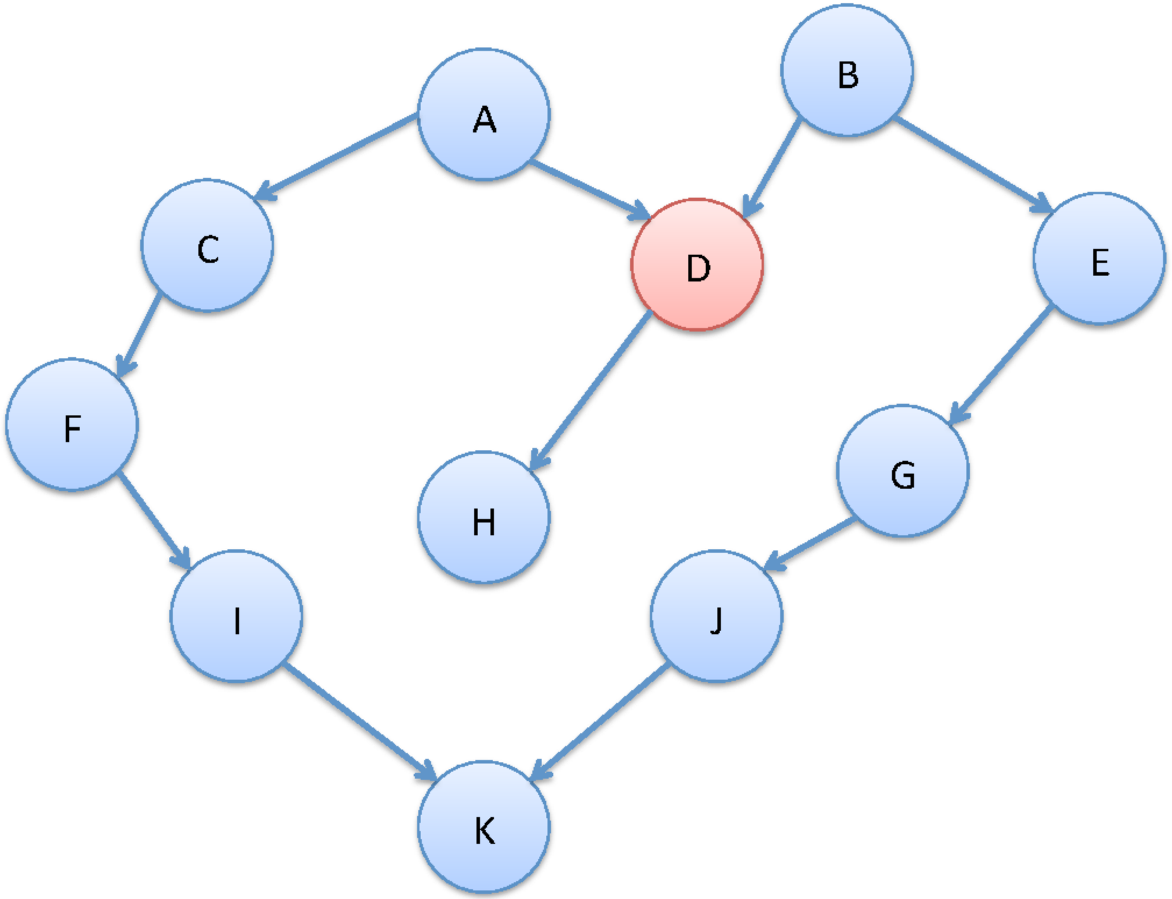
Easy:

- {A, C, F, I} ⊥ {B, E, G, J}

Local Markov Assumption (A, B)!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



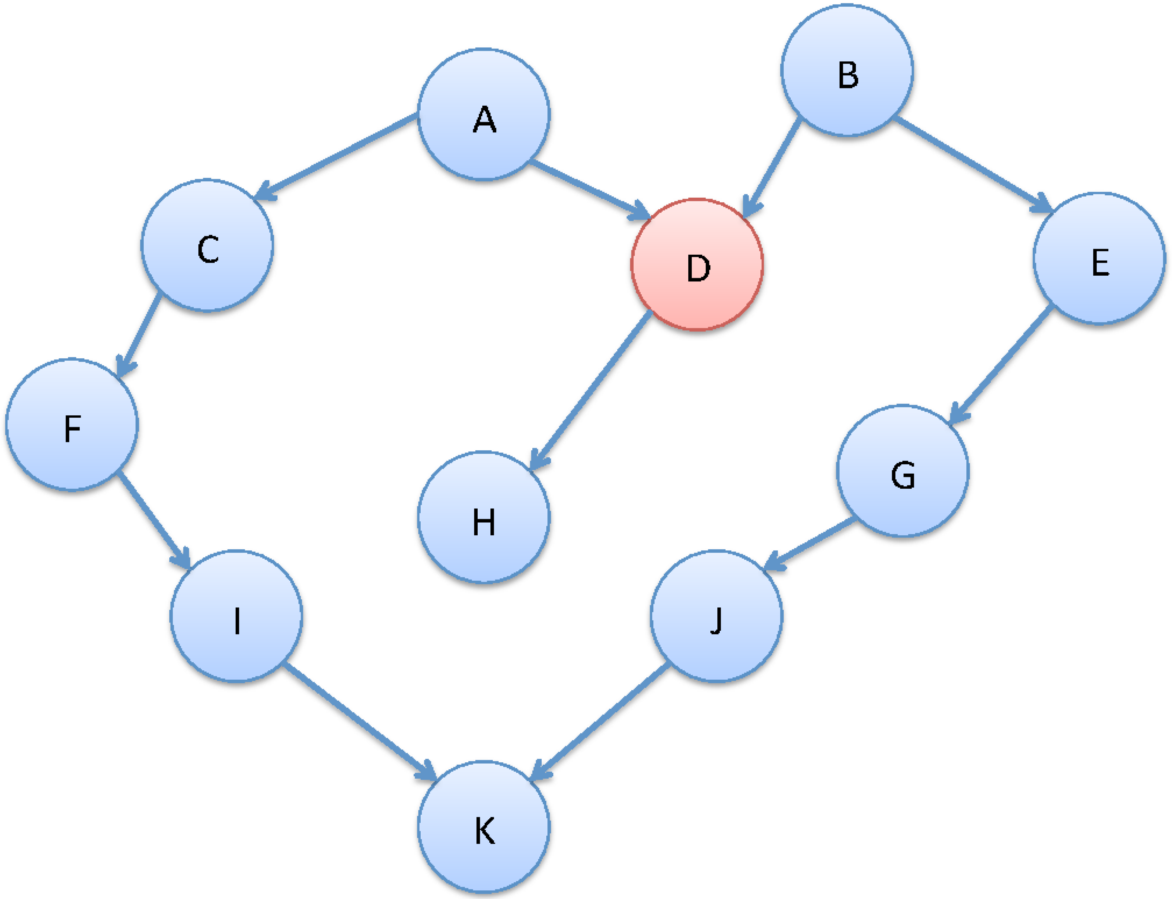
Easy:

- ¬(A ⊥ B | D)

V-Structure!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Easy:

- $\neg(A \perp B \mid D)$

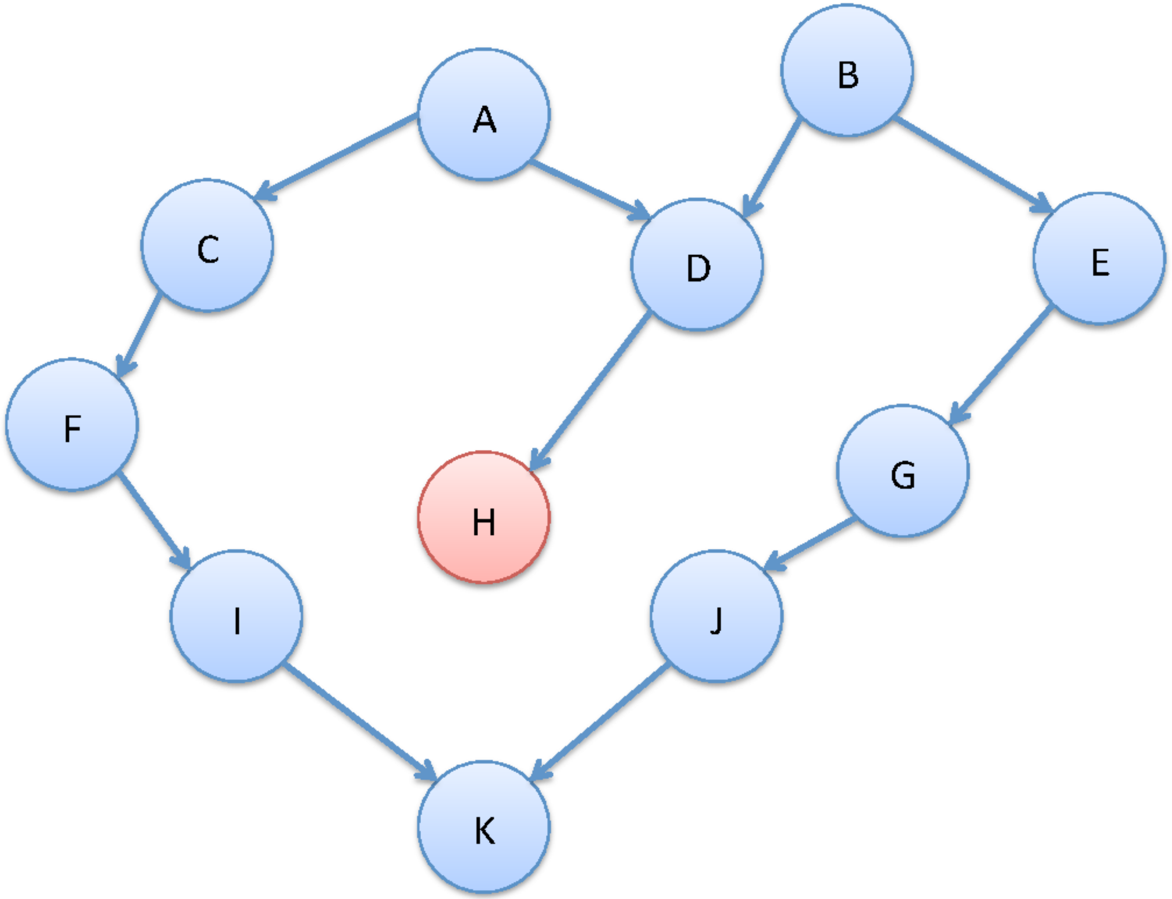
Harder:

- $\neg(F \perp G \mid D)$

“Flow of influence” again.

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Easy:

- $\neg(A \perp B \mid D)$

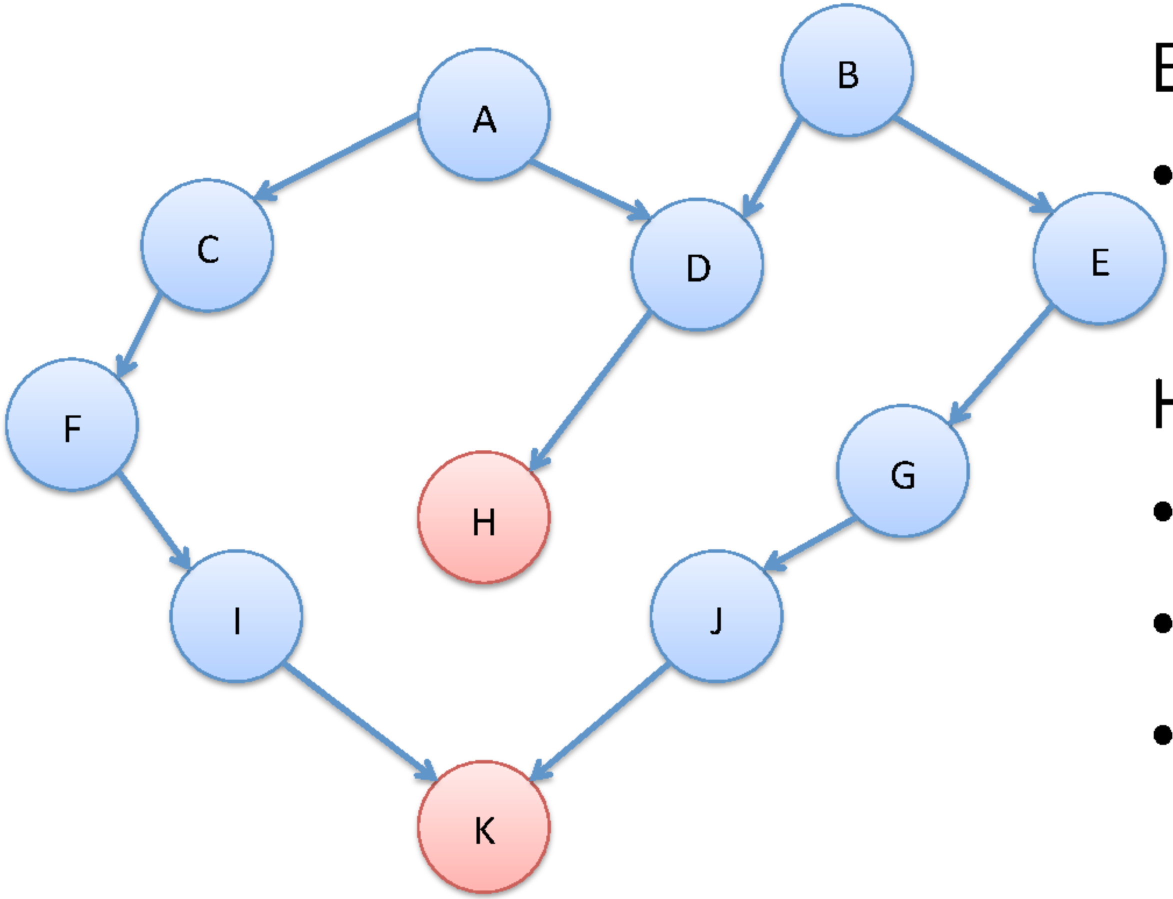
Harder:

- $\neg(F \perp G \mid D)$
- $\neg(F \perp G \mid H)$

Flow of influence, again?

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Easy:

- ¬(A ⊥ B | D)

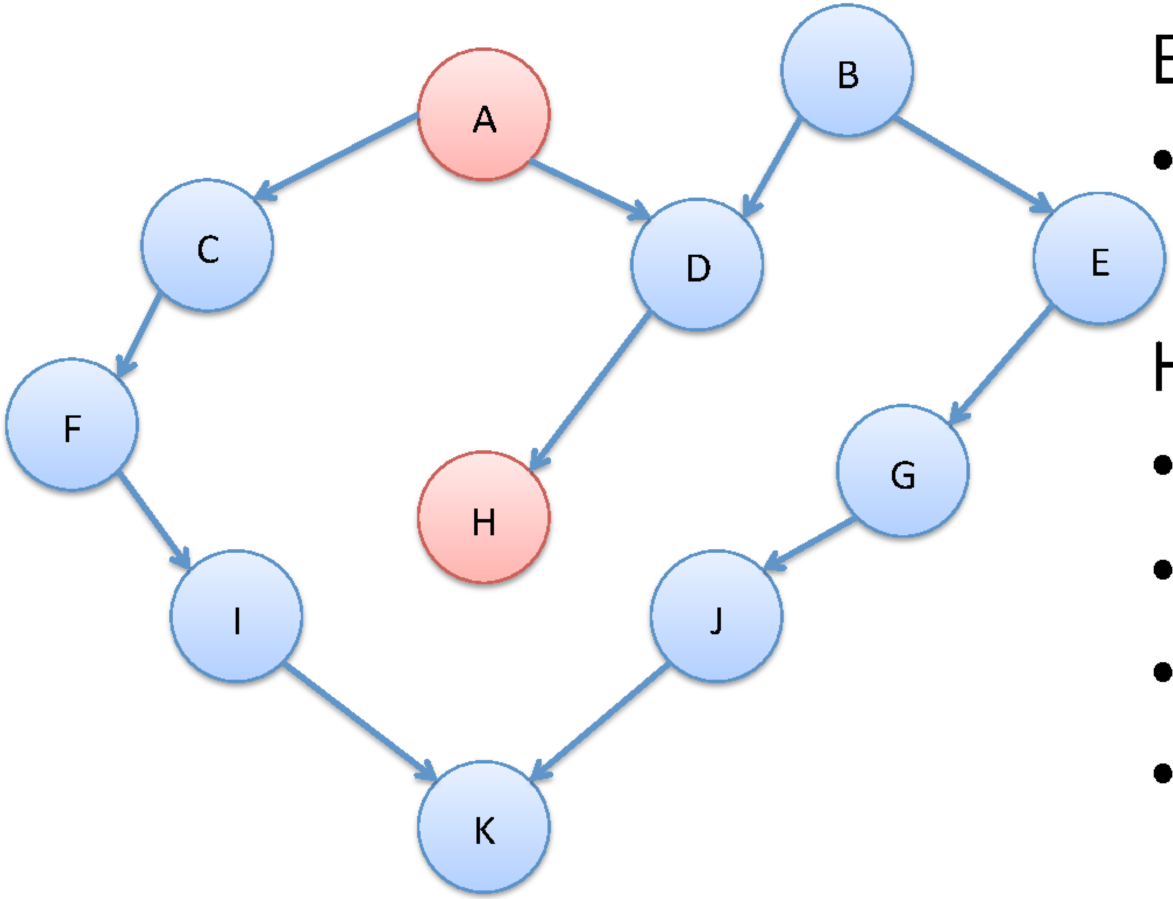
Harder:

- ¬(F ⊥ G | D)
- ¬(F ⊥ G | H)
- ¬(F ⊥ G | H, K)

More flow of influence!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

A More Complex Example



Easy:

- $\neg(A \perp B \mid D)$

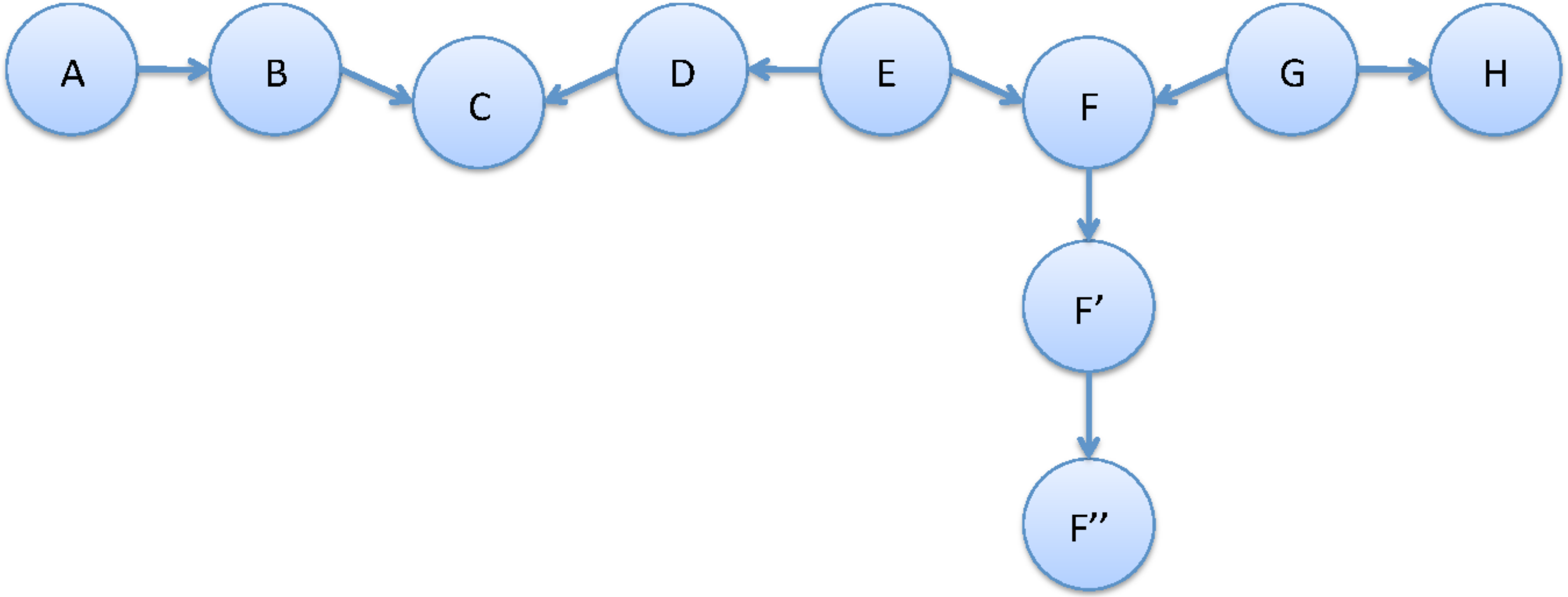
Harder:

- $\neg(F \perp G \mid D)$
- $\neg(F \perp G \mid H)$
- $\neg(F \perp G \mid H, K)$
- $(F \perp G \mid H, A)$

Observing A can “block” flow!

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

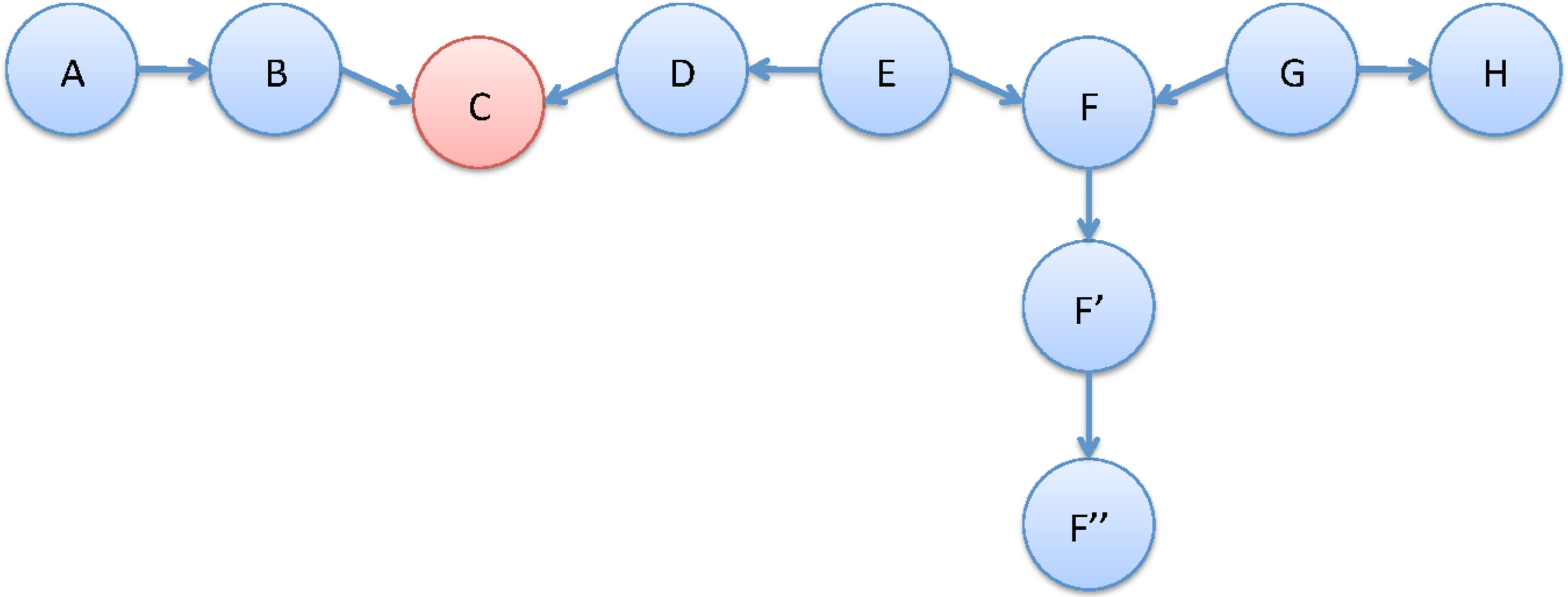
Another Example



- If I observe nothing, then A ⊥ H.

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

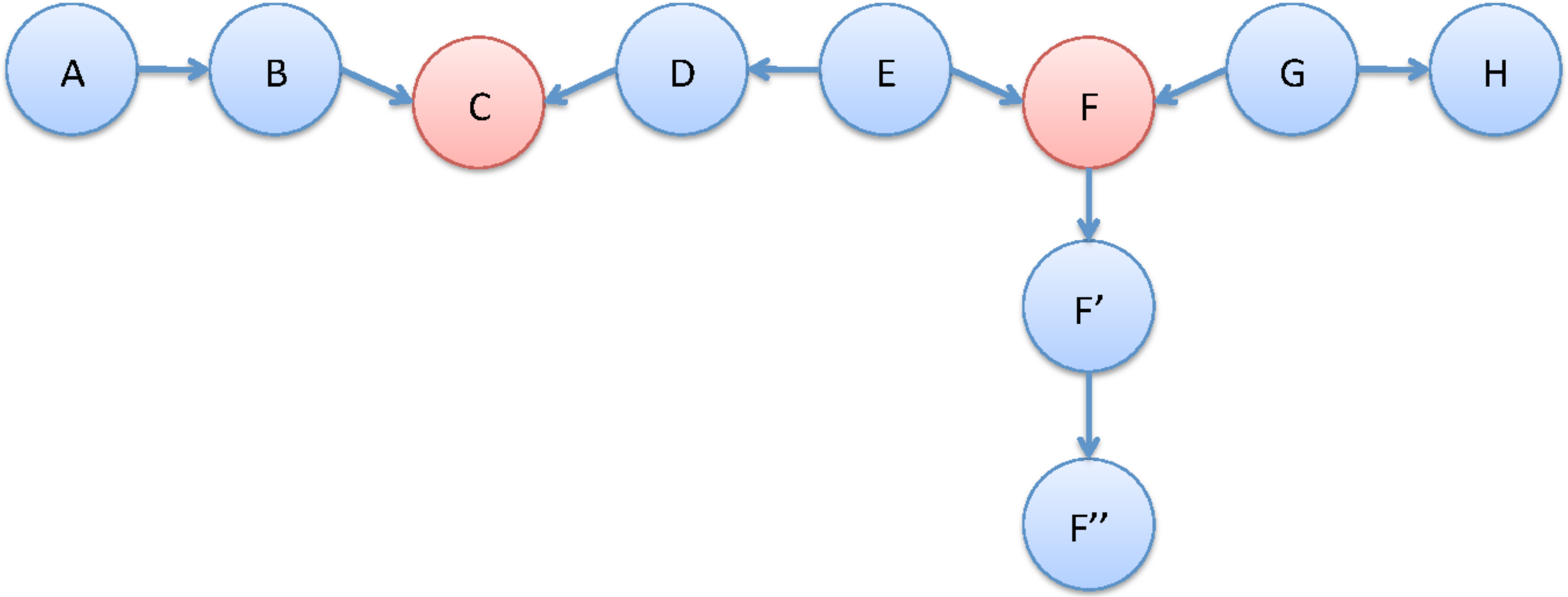
Another Example



- If I observe C, then A ⊥ H.

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

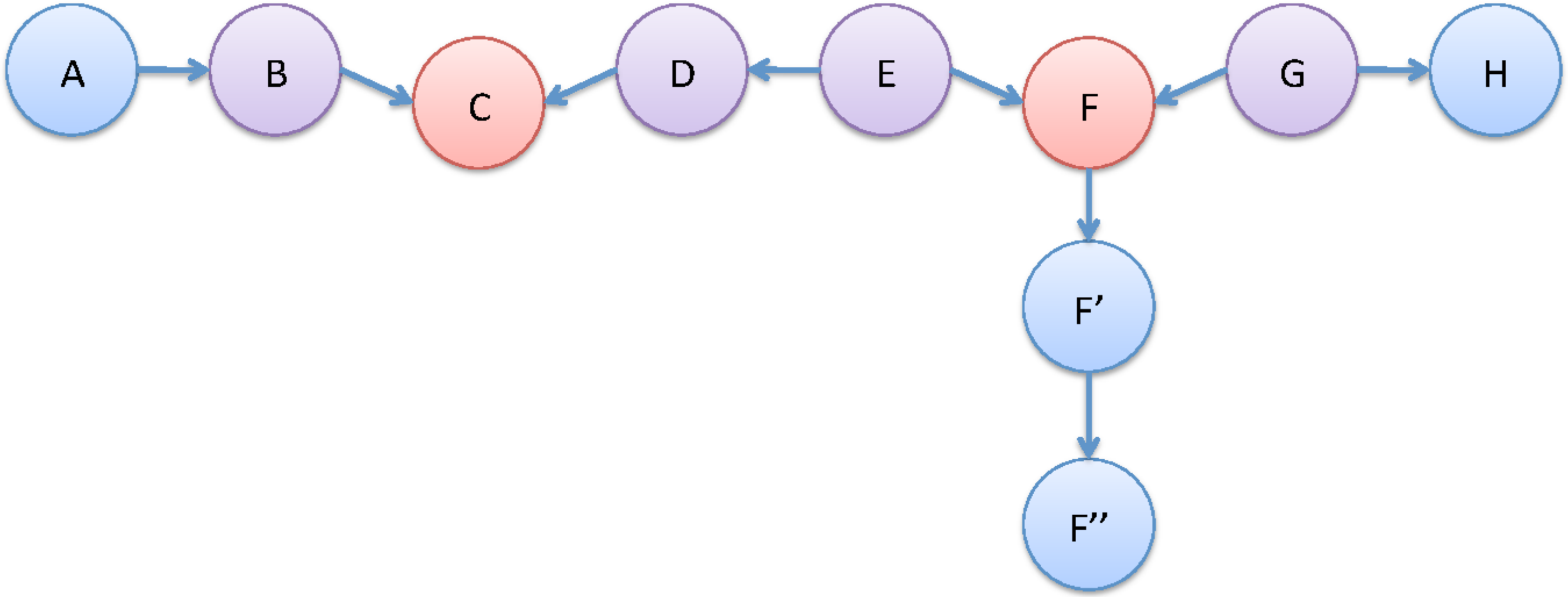
Another Example



- If I observe C and F, then ¬(A ⊥ H).

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

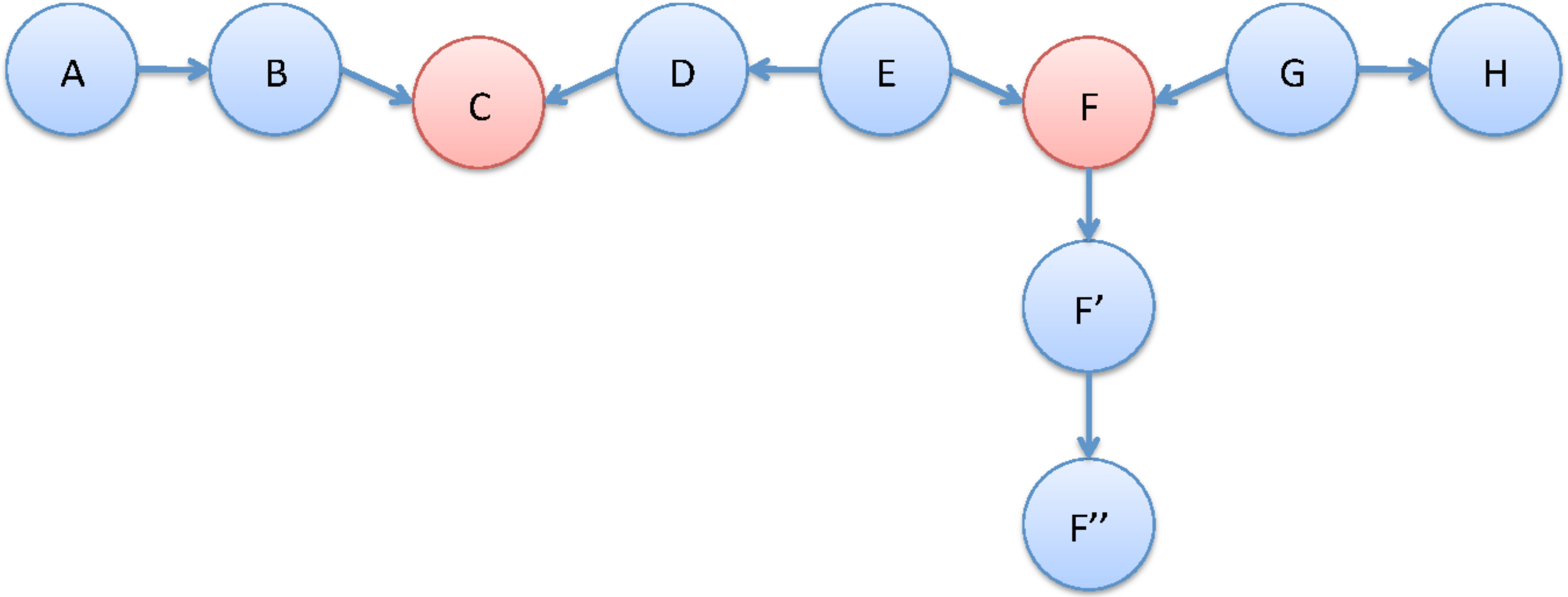
Another Example



- If I observe C and F, then ¬(A ⊥ H).
  - But if I observe B, D, E, and/or G, then A ⊥ H.

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

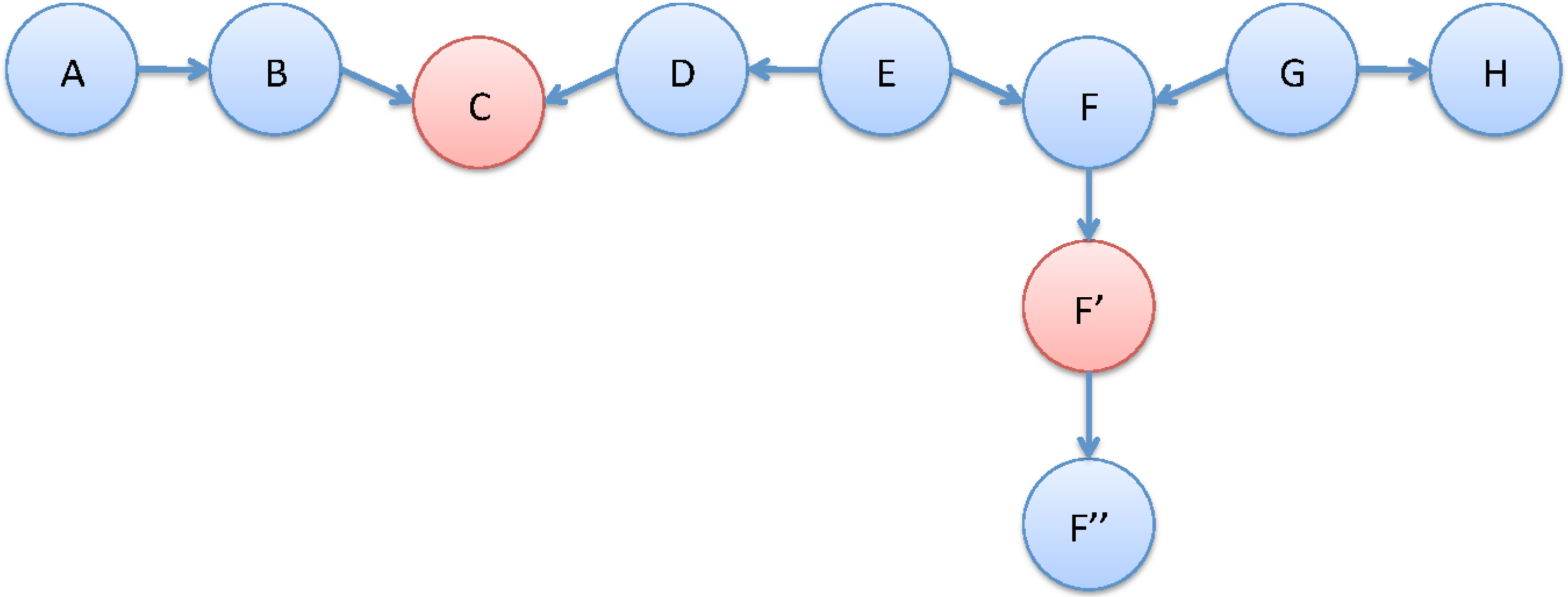
Another Example



- If I observe C and F, then ¬(A ⊥ H).

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

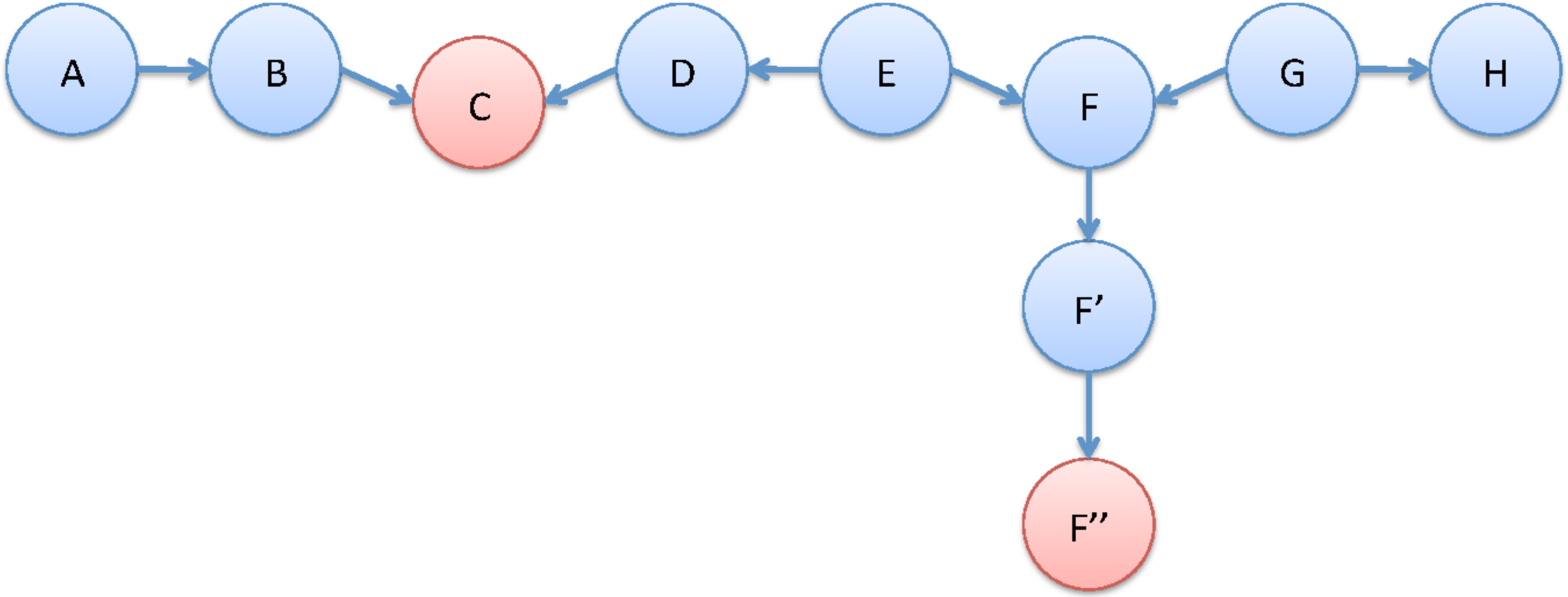
Another Example



- If I observe C and F’, then ¬(A ⊥ H).

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Another Example



- If I observe C and F'', then ¬(A ⊥ H).

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

The Real Inference Problem

- Given a Bayesian network over **X**, and a value $x \in \text{Val}(X_i)$, compute $P(X_i = x)$.

$$P(X_i = x) = \sum_{x_{-i} \in \text{Val}(X_{-i})} P(X_1 = x_1, ..., X_i = x, ..., X_n = x_n)$$

- Assume we are given a graphical model.
- Want:

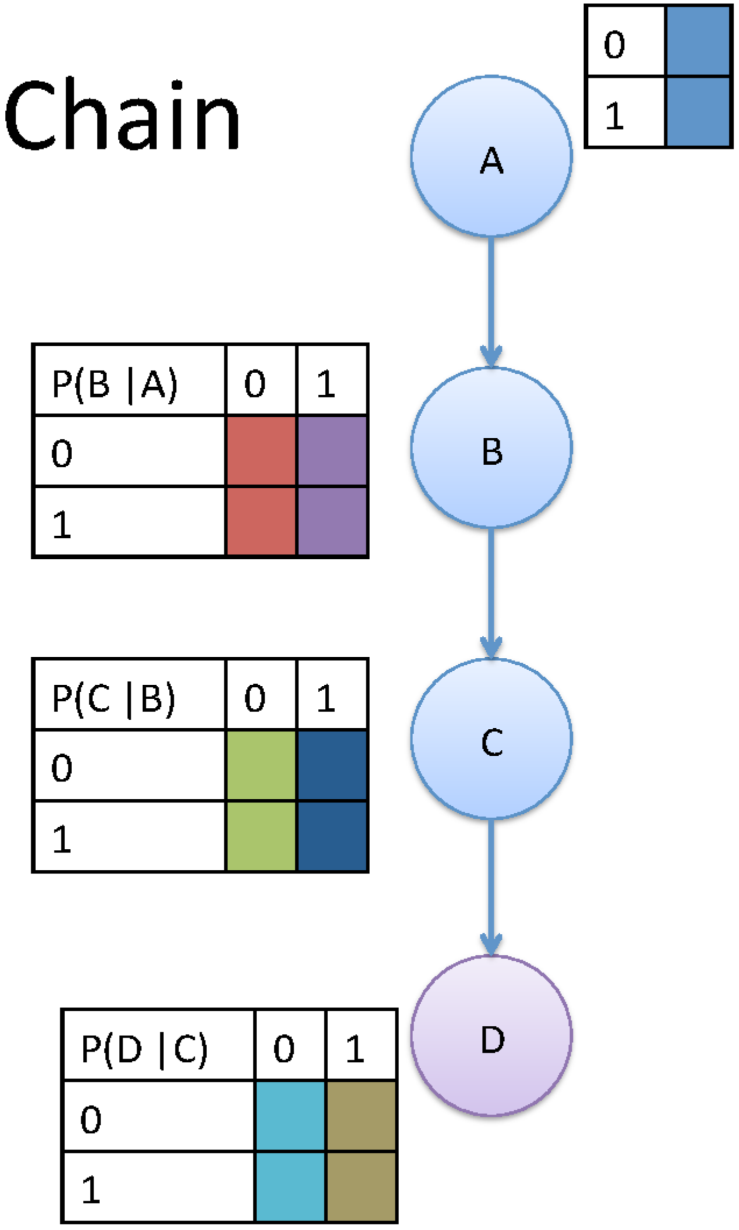
$$\begin{aligned} P(\boldsymbol{X} \mid \boldsymbol{E} = \boldsymbol{e}) &= \frac{P(\boldsymbol{X}, \boldsymbol{E} = \boldsymbol{e})}{P(\boldsymbol{E} = \boldsymbol{e})} \\ &\propto P(\boldsymbol{X}, \boldsymbol{E} = \boldsymbol{e}) \\ &= \sum_{\boldsymbol{y} \in \text{Val}(\boldsymbol{Y})} P(\boldsymbol{X}, \boldsymbol{E} = \boldsymbol{e}, \boldsymbol{Y} = \boldsymbol{y}) \end{aligned}$$

Such exact inference is hopeless in general.

Let’s just try it anyway.

Markov Chain

- Let's calculate P(B) from things we have.

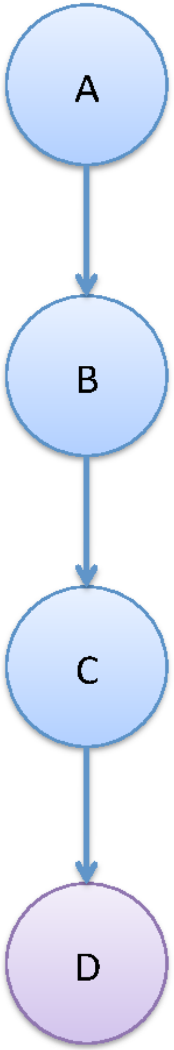


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- Let's calculate P(B) from things we have.

$$P(B) = \sum_{a \in \mathrm{Val}(A)} P(A = a) P(B \mid A = a)$$



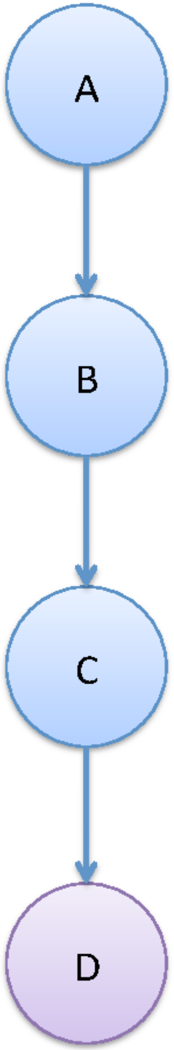
Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- Let's calculate P(B) from things we have.

$$P(B) = \sum_{a \in \text{Val}(A)} \underbrace{P(A=a)P(B \mid A=a)}_{P(A=a,\,B)}$$

- Note that C and D do not matter.



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- Let's calculate P(B) from things we have.

$$P(B) = \sum_{a \in \mathrm{Val}(A)} P(A = a) P(B \mid A = a)$$

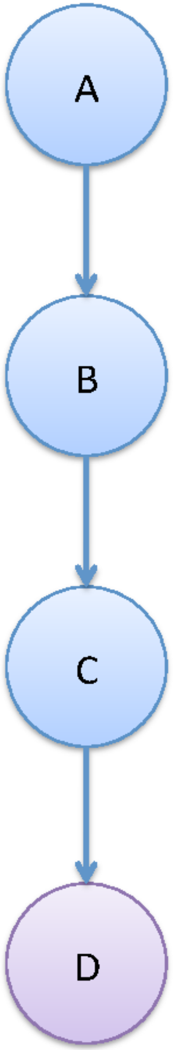
0	
1	

T

P(B | A)	0	1
0		
1		

=

0	
1	

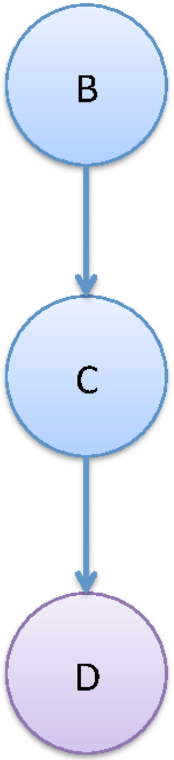


Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- We now have a Bayesian network for the marginal distribution P(B, C, D).

0	
1	



P(C |B)	0	1
0		
1		

P(D |C)	0	1
0		
1		

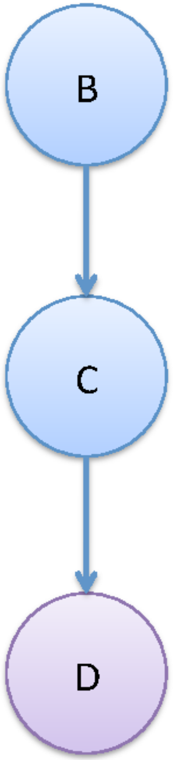
Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- We can repeat the same process to calculate P(C).

$$P(C) = \sum_{b \in \mathrm{Val}(B)} P(B = b) P(C \mid B = b)$$

- We already have P(B)!



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- We can repeat the same process to calculate P(C).

$$P(C) = \sum_{b \in \mathrm{Val}(B)} P(B=b)P(C \mid B=b)$$

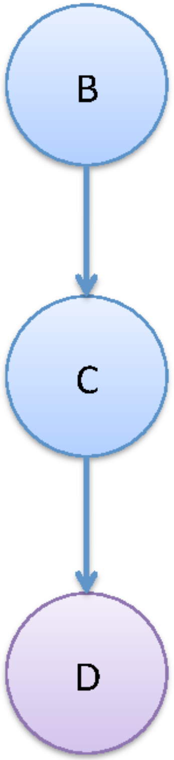
0	
1	

T

P(C | B)	0	1
0		
1		

=

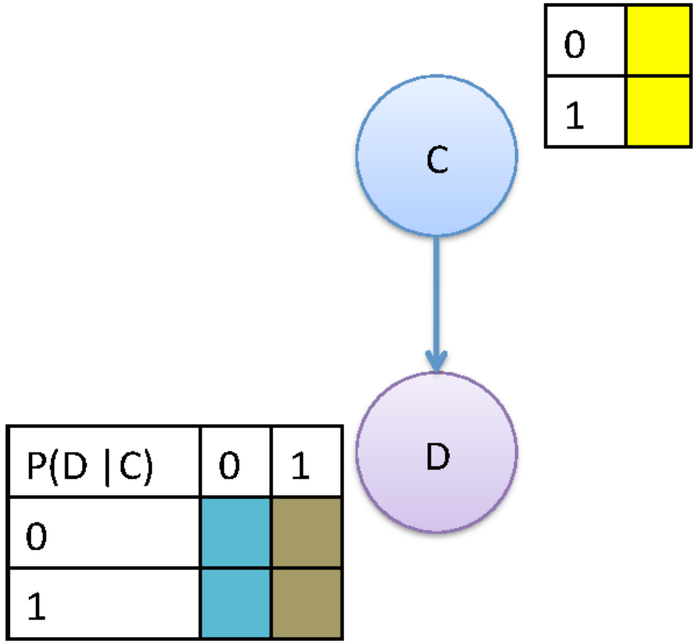
0	
1	



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- We now have P(C, D).
- Marginalizing out A and B happened in two steps, and we seem to be exploiting the Bayesian network structure.



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- Last step to get P(D):

$$P(D) \;=\; \sum_{c \in \mathrm{Val}(C)} P(C=c)P(D \mid C=c)$$

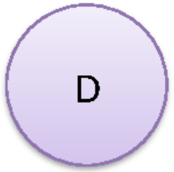
0	
1	

T

P(D | C)	0	1
0		
1		

=

0	
1	



Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

Markov Chain

- Notice that the same step happened for each random variable:
  - We created a new CPD over the variable and its "successor"
  - We summed out (marginalized) the variable.

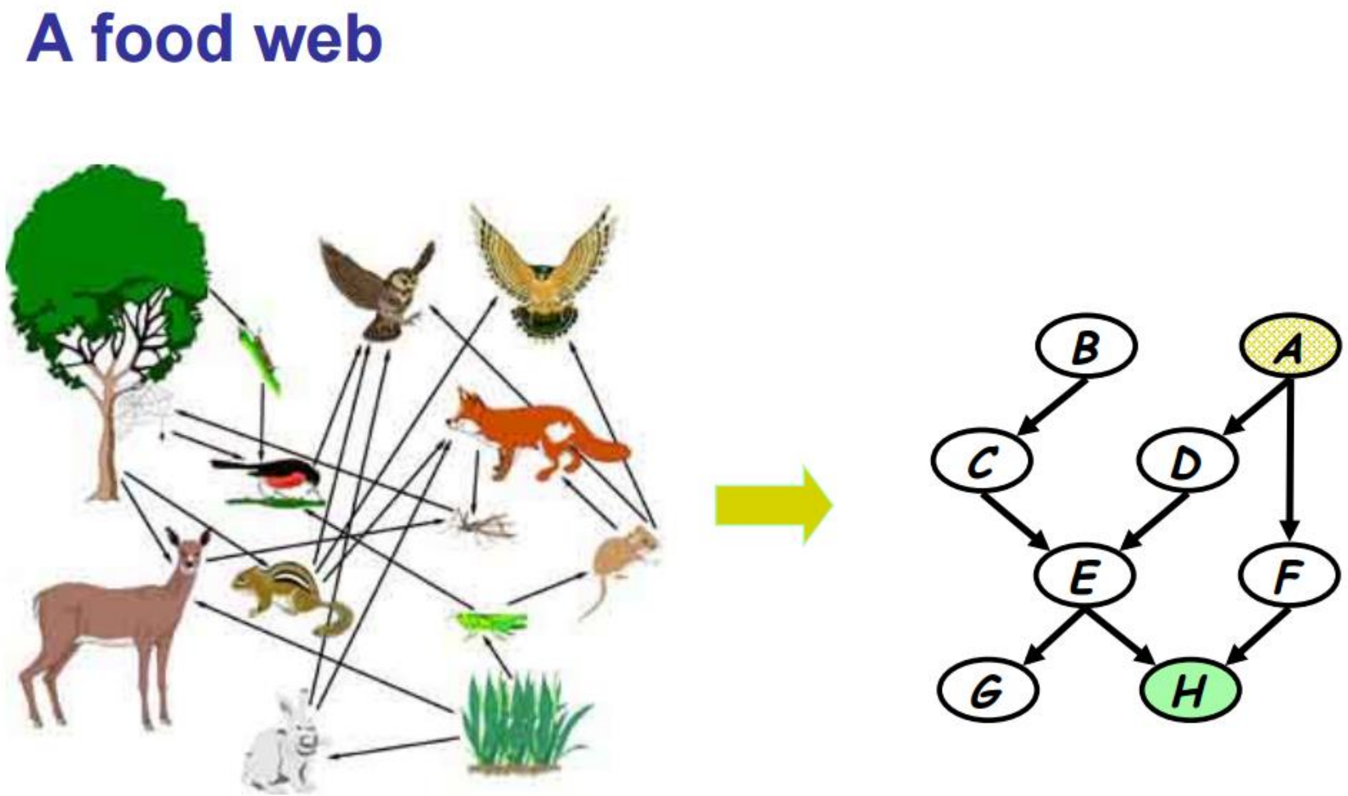
$$
\begin{aligned}
P(D) &= \sum_{a \in \mathrm{Val}(A)} \sum_{b \in \mathrm{Val}(B)} \sum_{c \in \mathrm{Val}(C)} P(A=a)P(B=b \mid A=a)P(C=c \mid B=b)P(D \mid C=c) \\
&= \sum_{c \in \mathrm{Val}(C)} P(D \mid C=c) \sum_{b \in \mathrm{Val}(B)} P(C=c \mid B=b) \sum_{a \in \mathrm{Val}(A)} P(A=a)P(B=b \mid A=a)
\end{aligned}
$$

Copied from: https://www.ark.cs.cmu.edu/PGM/index.php/Current_events_(2010)

That Was Variable Elimination

- We reused computation from previous steps and avoided doing the same work more than once.
  - Dynamic programming!
- We exploited the Bayesian network structure (each subexpression only depends on a small number of variables).
- Exponential blowup avoided!
- But: is there a general technique for any graphical model?

A more complex example



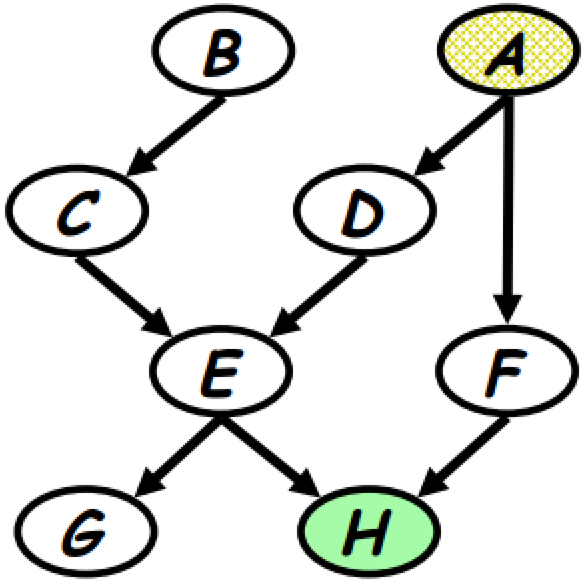
What is the probability that hawks are leaving given that the grass condition is poor?

Copied from Prof Eric Xing at http://www.cs.cmu.edu/~epxing/Class/10708/slides/lecture7-ExactInference.pdf

- Query: *P(A | h)*
  - Need to elimimate: *B,C,D,E,F,G,H*
- Initial factors:

$$P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$

- Choose an elimination order: *H,G,F,E,D,C,B*

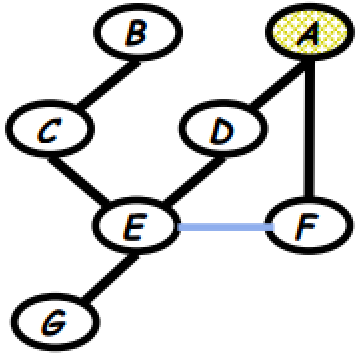


- Step 1:
  - **Conditioning** (fix the evidence node (i.e., *h*) on its observed value (i.e., $\tilde{h}$)):

$$m_h(e,f) = p(h = \tilde{h} \mid e,f)$$

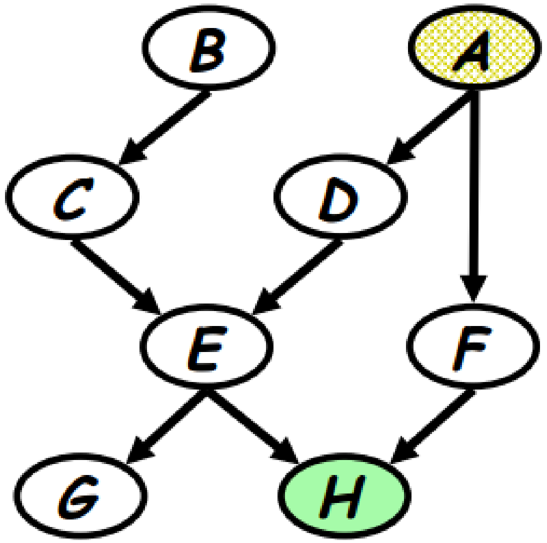
  - This step is isomorphic to a marginalization step:

$$m_h(e,f) = \sum_h p(h|e,f)\delta(h = \tilde{h})$$



Copied from Prof Eric Xing at http://www.cs.cmu.edu/~epxing/Class/10708/slides/lecture7-ExactInference.pdf

- Query: $P(B \mid h)$
  - Need to eliminate: *B,C,D,E,F,G*

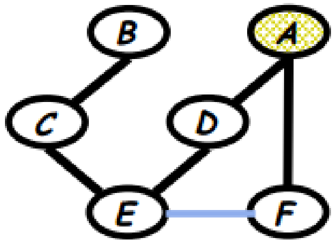


- Initial factors:

$$P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)P(h \mid e,f)$$
$$\Rightarrow P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)m_h(e,f)$$

- Step 2: Eliminate *G*
- compute

$$m_g(e) = \sum_g p(g \mid e) = 1$$



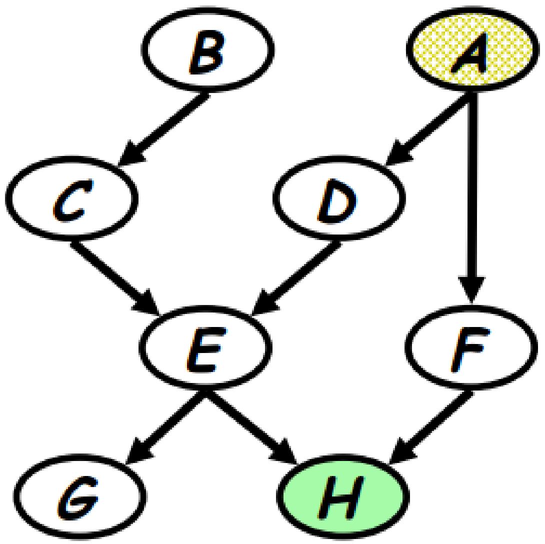
$$\Rightarrow P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)m_g(e)m_h(e,f)$$
$$= P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)m_h(e,f)$$

- Keep eliminating *F,E,D,C,B* in order

- Query: $P(B \mid h)$
  - Need to elimimate: B

- Initial factors:

$$P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)P(g|e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)m_f(a,e)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)m_e(a,c,d)$$
$$\Rightarrow P(a)P(b)P(c|d)m_d(a,c)$$
$$\Rightarrow P(a)P(b)m_c(a,b)$$
$$\Rightarrow P(a)m_b(a)$$

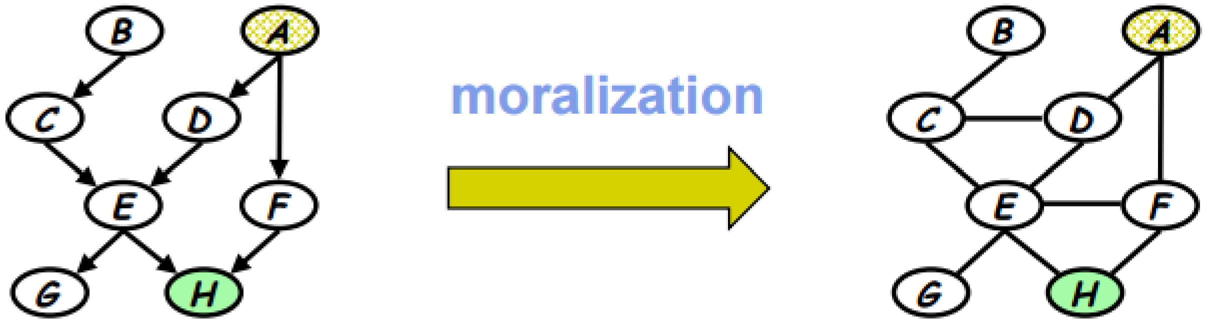


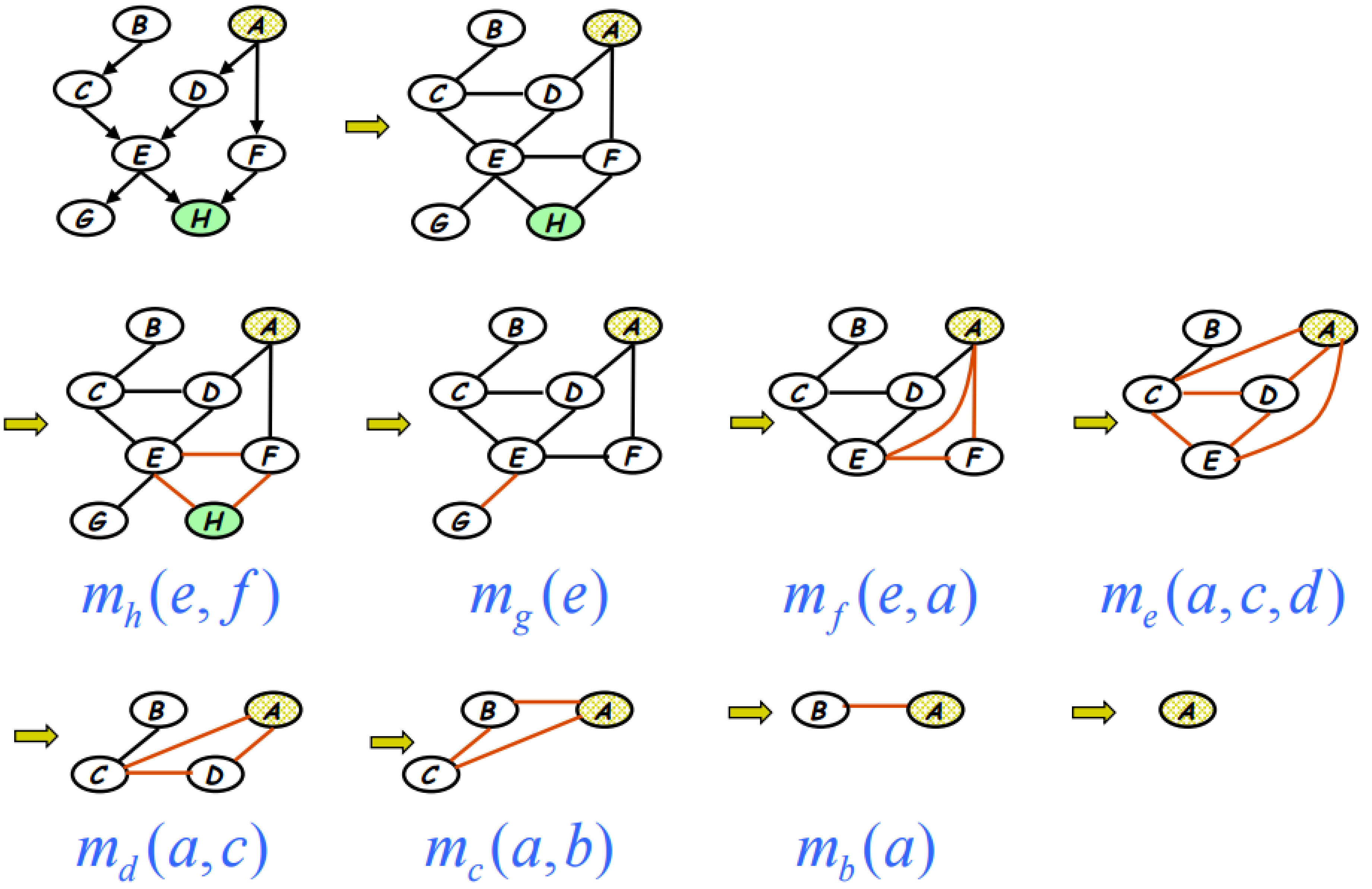
- Final Step: Wrap-up $p(a,\tilde{h}) = p(a)m_b(a)$, $\quad p(\tilde{h}) = \sum_a p(a)m_b(a)$

$$\Rightarrow P(a \mid \tilde{h}) = \frac{p(a)m_b(a)}{\sum p(a)m_b(a)}$$

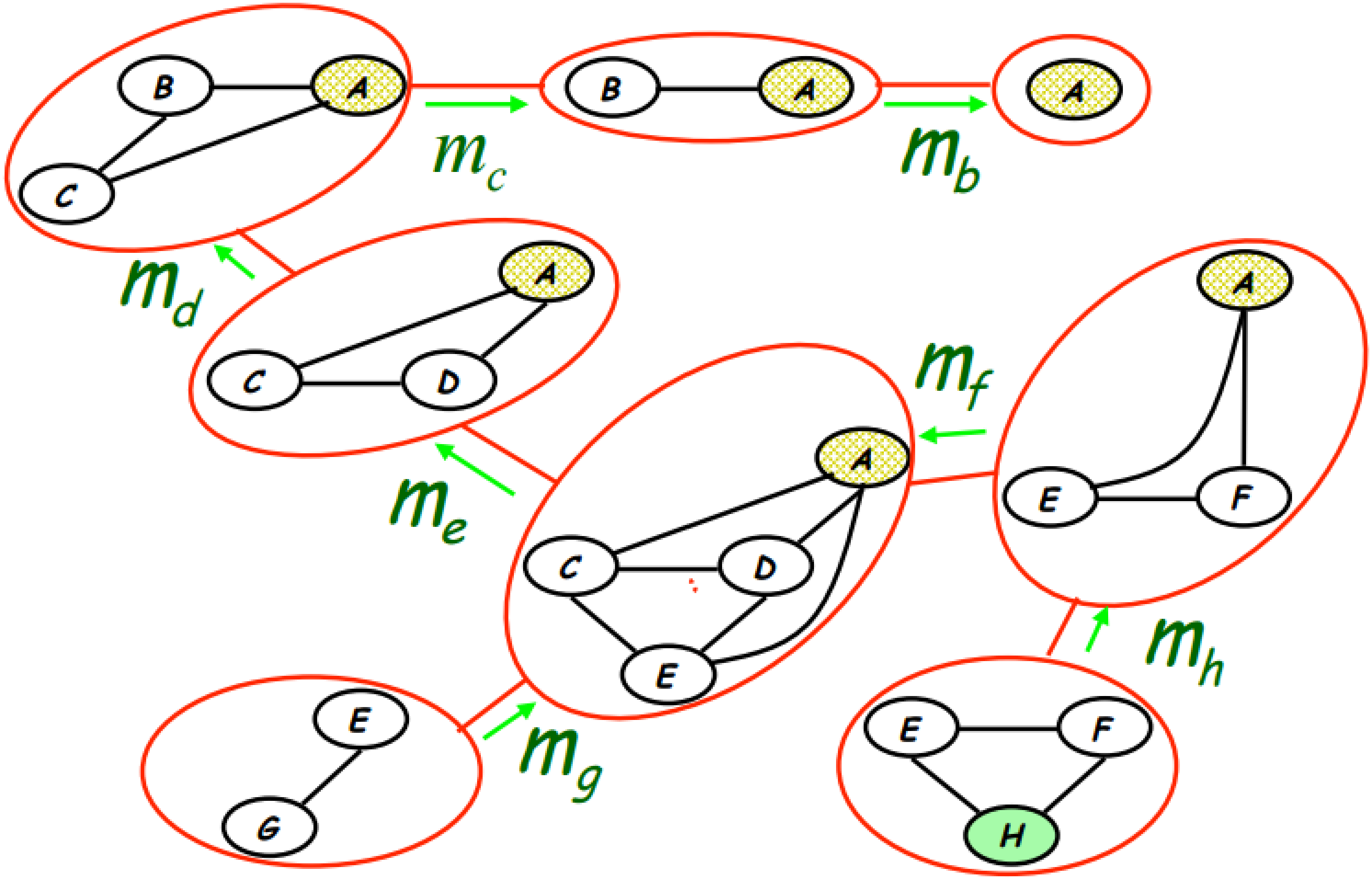
Cumbersome

- Maybe graph way could be easier
- Begin by moralizing the Bayesian network
  - Get parent nodes *married* if they have a common child
  - Ignore directedness of the graph





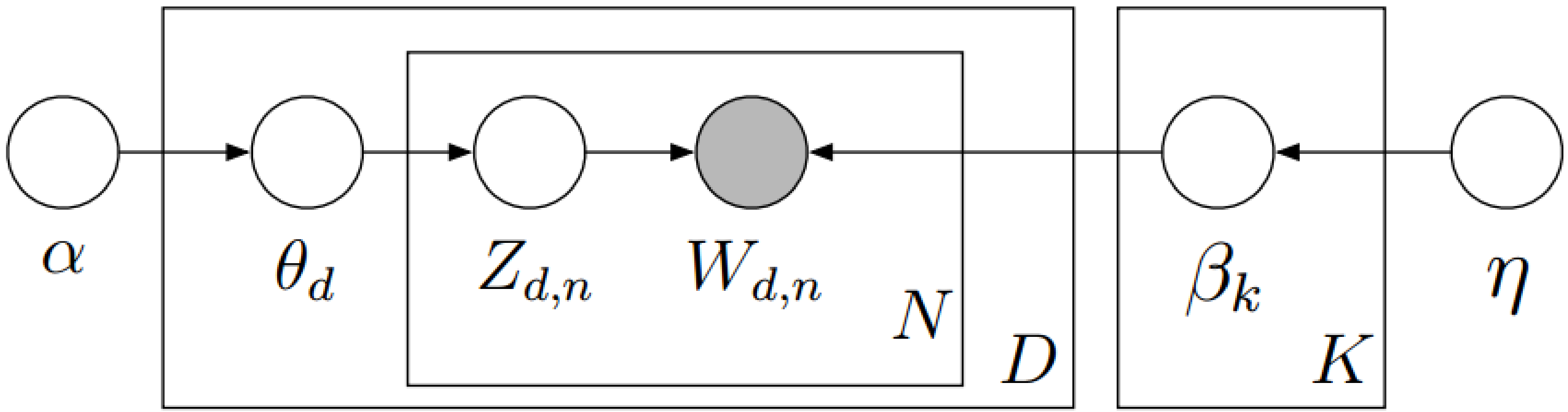
Copied from Prof Eric Xing at http://www.cs.cmu.edu/~epxing/Class/10708/slides/lecture7-ExactInference.pdf



$$m_e(a,c,d) = \sum_e p(e|c,d)m_g(e)m_f(a,e)$$

Copied from Prof Eric Xing at http://www.cs.cmu.edu/~epxing/Class/10708/slides/lecture7-ExactInference.pdf

Hopefully by now you can decipher



And maybe even …