# **Lecture 9b:** Bayesian Networks

CSCI 360

Introduction to Artificial Intelligence

USC

# Here is where we are…

| | | | | |
|---|---|---|---|---|
| | 3/1 | | Project 2 Out | |
| 9 | 3/4 | 3/5 | Quantifying Uncertainty | [Ch 13.1-13.6] |
| | 3/6 | 3/7 | Bayesian Networks | [Ch 14.1-14.2] |
| 10 | 3/11 | 3/12 | *(spring break, no class)* | |
| | 3/13 | 3/14 | *(spring break, no class)* | |
| 11 | 3/18 | 3/19 | Inference in Bayesian Networks | [Ch 14.3-14.4] |
| | 3/20 | 3/21 | Decision Theory | [Ch 16.1-16.3 and 16.5] |
| | 3/23 | | Project 2 Due | |
| 12 | 3/25 | 3/26 | Advanced topics (Chao traveling to NSF) | |
| | 3/27 | 3/28 | Advanced topics (Chao traveling to NSF) | |
| | 3/29 | | Homework 2 Out | |
| 13 | 4/1 | 4/2 | Markov Decision Processes | [Ch 17.1-17.2] |
| | 4/3 | 4/4 | Decision Tree Learning | [Ch 18.1-18.3] |
| | 4/5 | | Homework 2 Due | |
| | 4/5 | | Project 3 Out | |
| 14 | 4/8 | 4/9 | Perceptron Learning | [Ch 18.7.1-18.7.2] |
| | 4/10 | 4/11 | Neural Network Learning | [Ch 18.7.3-18.7.4] |
| 15 | 4/15 | 4/16 | Statistical Learning | [Ch 20.2.1-20.2.2] |
| | 4/17 | 4/18 | Reinforcement Learning | [Ch 21.1-21.2] |
| 16 | 4/22 | 4/23 | Artificial Intelligence Ethics | |
| | 4/24 | 4/25 | Wrap-Up and Final Review | |
| | 4/26 | | Project 3 Due | |
| | 5/3 | 5/2 | **Final Exam** (2pm-4pm) | |

# Outline

- What is AI?
- Problem-solving agent (search)
- Knowledge-based agent (logical reasoning)
- **Probabilistic reasoning**
  - Quantifying Uncertainty
  - **Bayesian Networks**
  - Inference in Bayesian Networks
  - Decision Theory
  - Markov Decision Processes
- Machine learning

# What we have learned so far…

- Early AI researchers largely rejected using probability in their systems
  - "People don't think that way…"

- However, neither **problem-solving** nor **logical reasoning** agents tolerate approximation well…
  - Need probabilistic modeling/reasoning
    - Represent KB as relationships among random variables
    - KB can be learned from data
    - Given the KB, make inference about variables of interest
  - Example applications
    - *Medical diagnosis*: symptoms and diseases as random variables
    - *Business decision making*, e.g., predicting customer's behavior
    - *Bio-informatics*, e.g., gene expression levels as random variables
    - *Computer science*: vision, speech recognition, spam filtering, …

# Recap: *Making decision*

**Rational decision** depends on

(1) The relative importance of various goals and

(2) likelihood that (and degree to which) they will be reached

**Decision theory** = Utility theory + Probability theory

Choose the action that yields the ***highest expected utility***, averaged over all the possible outcomes of the action

# Recap: *Probability axioms*

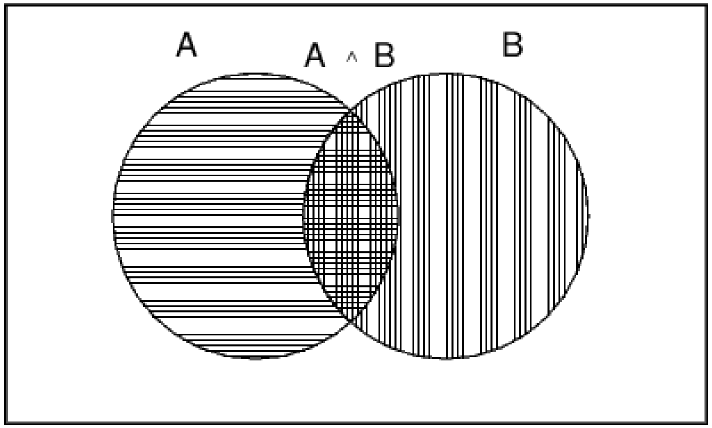
- A numerical probability *P(ω)* for each possible world

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega$$

$$\sum_{\omega \in \Omega} P(\omega) = 1$$

$$P(\neg a) \;=\; 1 - P(a)$$

$$\boxed{P(a \vee b) = P(a) + P(b) - P(a \wedge b)}$$



# Recap: *Conditional (or posterior) probability*

- For any propositions *a* and *b*, we have

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \qquad \text{whenever } P(b) > 0.$$

- $\boxed{P(a \wedge b) = P(a \mid b)\, P(b)}$

*joint probability*

*conditional probability*

*prior probability*

# Recap: *Probability distribution*

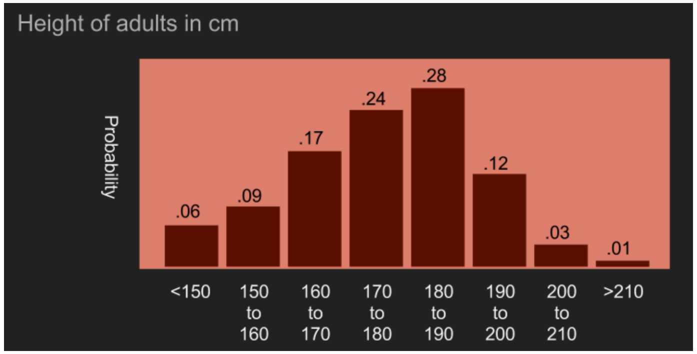
- Probabilities of all possible values of a random variable

$$
\begin{aligned}
&P(Weather = sunny) = 0.6 \\
&P(Weather = rain) = 0.1 \\
&P(Weather = cloudy) = 0.29 \\
&P(Weather = snow) = 0.01 ,
\end{aligned}
$$

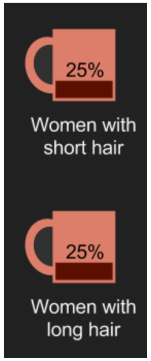
- In a vector format

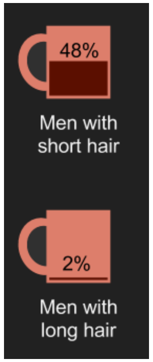
$$\boxed{\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle}$$

- Other examples



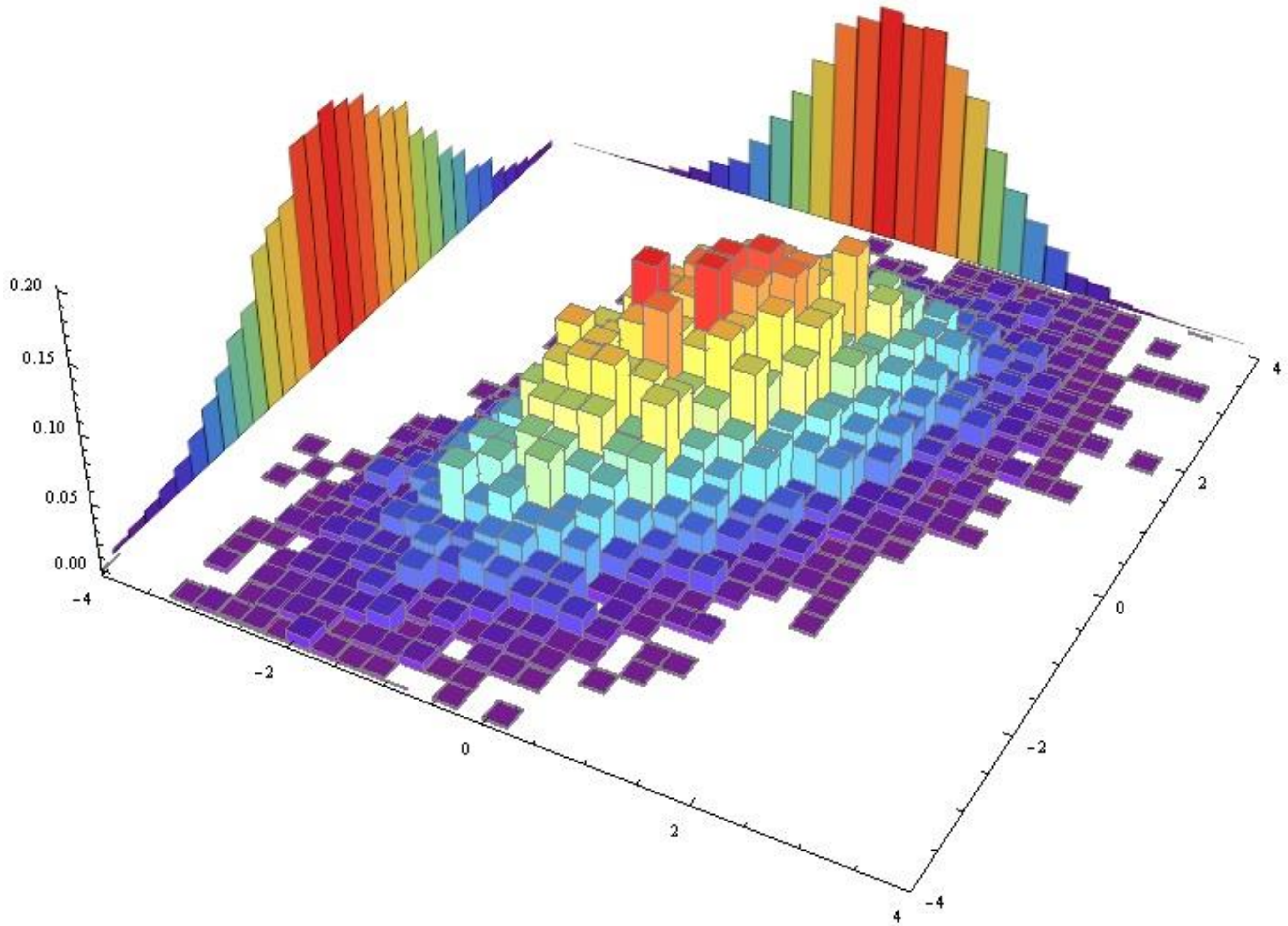

https://brohrer.github.io



# Recap: *Joint probability distribution*

- Probabilities of all possible values of **multiple** variables



# Recap: *Marginal probability*

- Extracting the distribution **over a subset** of variables from the full joint distribution

| | $toothache$ | | $\neg toothache$ | |
|---|---|---|---|---|
| | $catch$ | $\neg catch$ | $catch$ | $\neg catch$ |
| $cavity$ | 0.108 | 0.012 | 0.072 | 0.008 |
| $\neg cavity$ | 0.016 | 0.064 | 0.144 | 0.576 |

- Example

$$P(cavity) =$$

# Recap: *Marginal probability*

- Extracting the distribution **over a subset** of variables from the full joint distribution

| | $toothache$ | | $\neg toothache$ | |
|---|---|---|---|---|
| | $catch$ | $\neg catch$ | $catch$ | $\neg catch$ |
| $cavity$ | 0.108 | 0.012 | 0.072 | 0.008 |
| $\neg cavity$ | 0.016 | 0.064 | 0.144 | 0.576 |

- Example

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

# Recap: *Normalization*

- The probability of **cavity**, or **no cavity**, given **toothache**

| | $toothache$ | | $\neg toothache$ | |
|---|---|---|---|---|
| | $catch$ | $\neg catch$ | $catch$ | $\neg catch$ |
| $cavity$ | 0.108 | 0.012 | 0.072 | 0.008 |
| $\neg cavity$ | 0.016 | 0.064 | 0.144 | 0.576 |

- Example

$$P(cavity \mid toothache) =$$

$$P(\neg cavity \mid toothache) =$$

Sum of the two is always 1.0

# Recap: *Normalization*

- The probability of **cavity**, or **no cavity**, given **toothache**

| | *toothache* | | *¬toothache* | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

- Example

No need to compute *P (toothache)* any more

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{\cancel{P(toothache)}}$$

$$= \frac{0.108 + 0.012}{\cancel{0.108 + 0.012 + 0.016 + 0.064}} = 0.6$$

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{\cancel{P(toothache)}}$$

$$= \frac{0.016 + 0.064}{\cancel{0.108 + 0.012 + 0.016 + 0.064}} = 0.4$$

# Recap: *Normalization*

- The probability of **cavity**, or **no cavity**, given **toothache**

| | $toothache$ | | $\neg toothache$ | |
|---|---|---|---|---|
| | $catch$ | $\neg catch$ | $catch$ | $\neg catch$ |
| $cavity$ | 0.108 | 0.012 | 0.072 | 0.008 |
| $\neg cavity$ | 0.016 | 0.064 | 0.144 | 0.576 |

- Example

$$\mathbf{P}(Cavity \mid toothache) = \alpha\,\mathbf{P}(Cavity, toothache)$$

$$= \alpha\,\langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle\ .$$

Assume that
$\alpha = 1 / (0.12+0.08)$
$= 1/0.2$
$= 5$

# Recap: *Independence to reduce table size*

- Consider ***P****(Toothache, Catch, Cavity, Weather)*, which has 32 entries in the **full joint distribution** table

| | toothache | | ¬toothache | | toothache | | ¬toothache | | toothache | | ¬toothache | | toothache | | ¬toothache | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | catch | ¬catch | catch | ¬catch | catch | ¬catch | catch | ¬catch | catch | ¬catch | catch | ¬catch | catch | ¬catch | catch | ¬catch |
| cavity | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 |
| ¬cavity | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 |

- Applying the product rule

$$P(toothache, catch, cavity, cloudy)$$
$$= P(cloudy \mid toothache, catch, cavity) P(toothache, catch, cavity)$$

- But weather is not influenced by dentistry!

$$P(cloudy \mid toothache, catch, cavity) = P(cloudy)$$

$$\boxed{P(toothache, catch, cavity, cloudy) = P(cloudy) P(toothache, catch, cavity)}$$

# Recap: *Independence to reduce table size*

- Consider ***P****(Toothache, Catch, Cavity, Weather)*, which has 32 entries in the **full joint distribution** table

| | toothache | | ¬toothache | | toothache | | ¬toothache | | toothache | | ¬toothache | | toothache | | ¬toothache | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 | 0.016 | 0.064 | 0.144 | 0.576 |

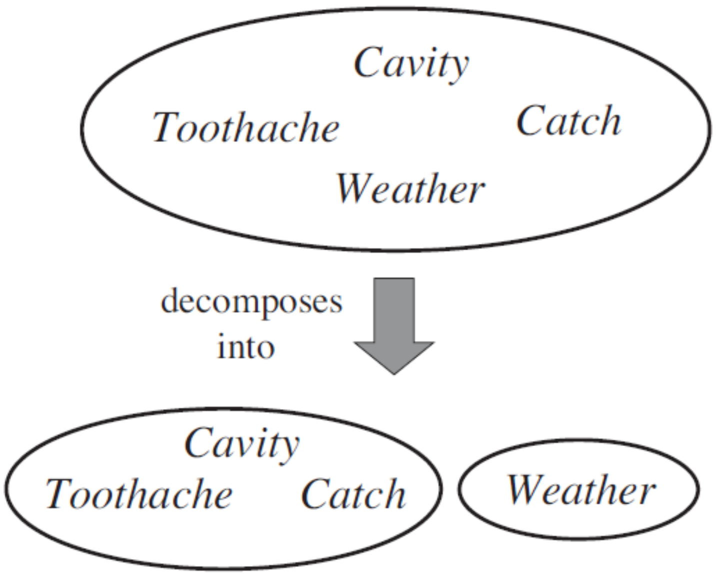
- The 32-element table can be reduced to a 8-element table and a 4-element table

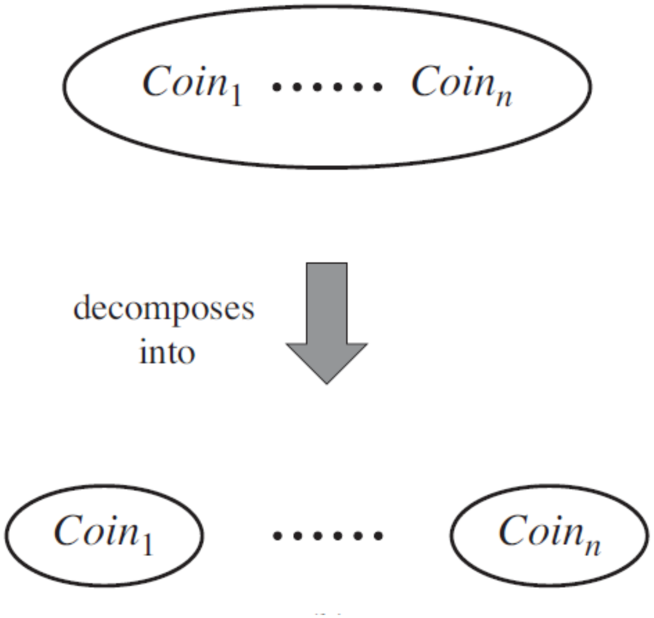


| | toothache | | ¬toothache | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

# Recap: *Independence to reduce table size*

*Leveraging the (absolute) independence*




Weather and dentistry are independent

Coin flips are independent

# Recap: *Conditional independence*

- Variables $X$ and $Y$ are **conditional independent**, given a third variable $Z$

$$\boxed{\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\mathbf{P}(Y \mid Z)}$$

- Alternatively, we have

$$\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z)$$

For **X**, variable **Y** doesn't provide any additional information, given **Z**

$$\mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$$

For **Y**, variable **X** doesn't provide any additional information, given **Z**

# Recap: *Conditional independence to reduce table size*

- For *n* effects that are conditionally independent given the cause, the **full joint distribution** table size grows as *O(n)* instead of *O($2^n$)*

$$\boxed{\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)}$$

# Recap: *Bayes' rule*

- Derive **Bayes' rule** from the **product rule** of conditional probability

$$P(a \wedge b) =$$

$$P(a \wedge b) =$$

- Equating the right-hand sides and dividing by $P(a)$

$$\boxed{P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}}$$

# Recap: *Bayes' rule*

- Derive **Bayes' rule** from the **product rule** of conditional probability

$$P(a \wedge b) = P(b \,|\, a) P(a)$$

$$P(a \wedge b) = P(a \,|\, b) P(b)$$

- Equating the right-hand sides and dividing by $P(a)$

$$\boxed{P(b \,|\, a) = \frac{P(a \,|\, b) P(b)}{P(a)}}$$

This equation underlies most **modern AI systems** for **probabilistic inference…**

# Recap: *Quiz 4 solution*

- Assume that the doctor knows some unconditional facts:
  - Prior probability that a patient has a disease $P(d) = 0.01$

    $P(\neg d) = 0.99$
  - Probability of test positive given no disease $P(\, tp \mid \neg d) = 0.096$
  - Probability of test positive given disease $P(\, tp \mid d\,) = 0.8$

- Now, a patient has test positive; what is the probability that this particular patient has the disease?

$$\boxed{P(d|tp) = P(tp|d)\, P(d) \,/\, P(tp)}$$

$$= P(tp|d)\, P(d) \,/\, (\, P(tp|d)\, P(d) + P(tp|\neg d)\, P(\neg d)\,)$$

$$= 0.8{*}0.01 \,/\, (\, 0.8{*}0.01 + 0.096{*}0.99)$$

$$= 0.008 \,/\, (0.008{+}0.09504)$$

$$= 0.0776$$

# Outline

- Representing knowledge in Bayesian networks
- Semantics of Bayesian networks
- Efficient representation of conditional distributions
- Exact inference in Bayesian networks

# Why Bayesian networks?

- **Full joint distribution** can be used to answer any query about the world
  - But table size is exponential in the number of variables
- **Independence** and **conditional independence** relations are important in simplifying the table
  - But they are unnatural and tedious to specify
- **Bayesian networks** is a data structure to represent both joint distributions and the dependencies among variables

# Example

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

| S1 | S2 | S3 | … | D1 | D2 | D3 | … | P(S1, S2, S3, …, D1, D2, D3, …) |
|---|---|---|---|---|---|---|---|---|
| true | true | true | … | true | true | true | … | 0.0000001 |
| … | … | … | … | … | …. | … | … | |
| false | false | false | … | false | false | false | … | 0.0000002 |

# Example (cont'd)

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

| S1 | S2 | S3 | … | D1 | D2 | D3 | … | P(S1, S2, S3, …, D1, D2, D3, …) |
|---|---|---|---|---|---|---|---|---|
| true | true | true | … | true | true | true | … | 0.0000001 |
| … | … | … | … | … | …. | … | … | |
| false | false | false | … | false | false | false | … | 0.0000002 |

- When the doctor observes presence of S1 and absence of S3, calculate
  - P(D1 | S1, ¬ S3) = P(D1, S1, ¬ S3) / P(S1, ¬ S3) = …
  - P(D2 | S1, ¬ S3) = …
  - P(D3 | S1, ¬ S3) = …
  - …

# Example (cont'd)

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

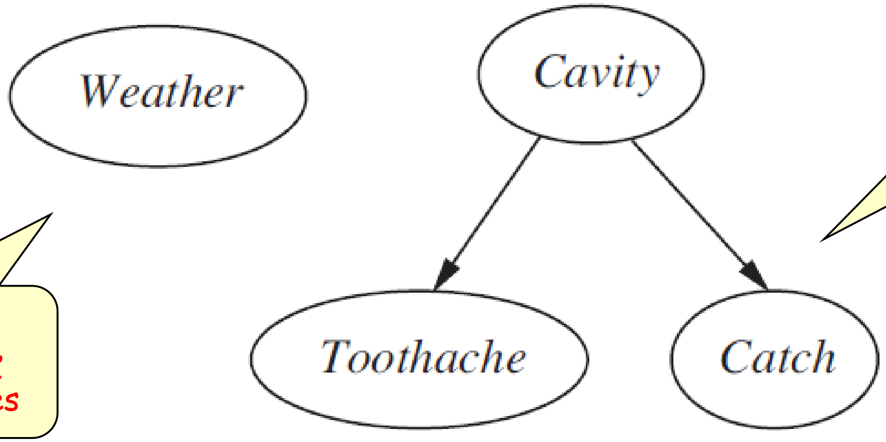
| **S1** | **S2** | **S3** | **…** | **D1** | **D2** | **D3** | **…** | **P(S1, S2, S3, …, D1, D2, D3, …)** |
|---|---|---|---|---|---|---|---|---|
| true | true | true | … | true | true | true | … | 0.0000001 |
| … | … | … | … | … | …. | … | … | |
| false | false | false | … | false | false | false | … | 0.0000002 |

- We need to acquire too many probabilities from the expert.
- Many of the probabilities are very close to zero and thus hard to specify by experts.

# Bayesian network

- A directed acyclic graph (DAG) where
  - Each node corresponds to a random variable,
  - Each edge from node *X* to node *Y* represents a direct influence of *X* on *Y*,
  - Each node $X_i$ has a conditional probability distribution $P( X_i \mid Parents(X_i) )$ that quantifies the effect of the parents on the node.

- Example

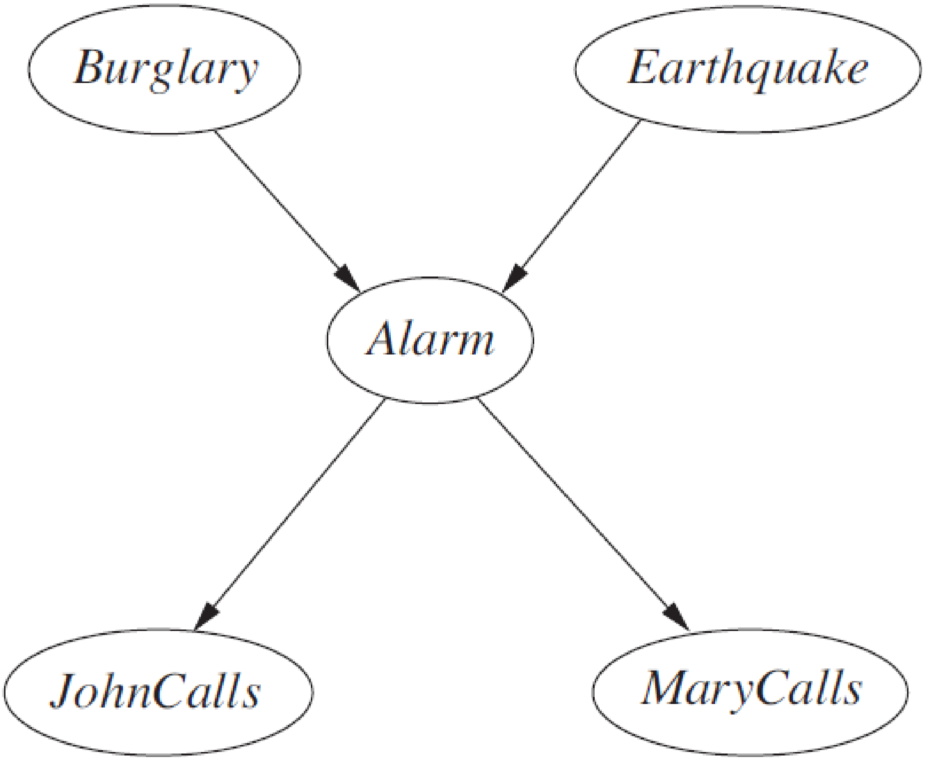



**Independence** between *Weather* and other variables

**Conditional independence** between *Toothache* and *Catch*, given *Cavity*

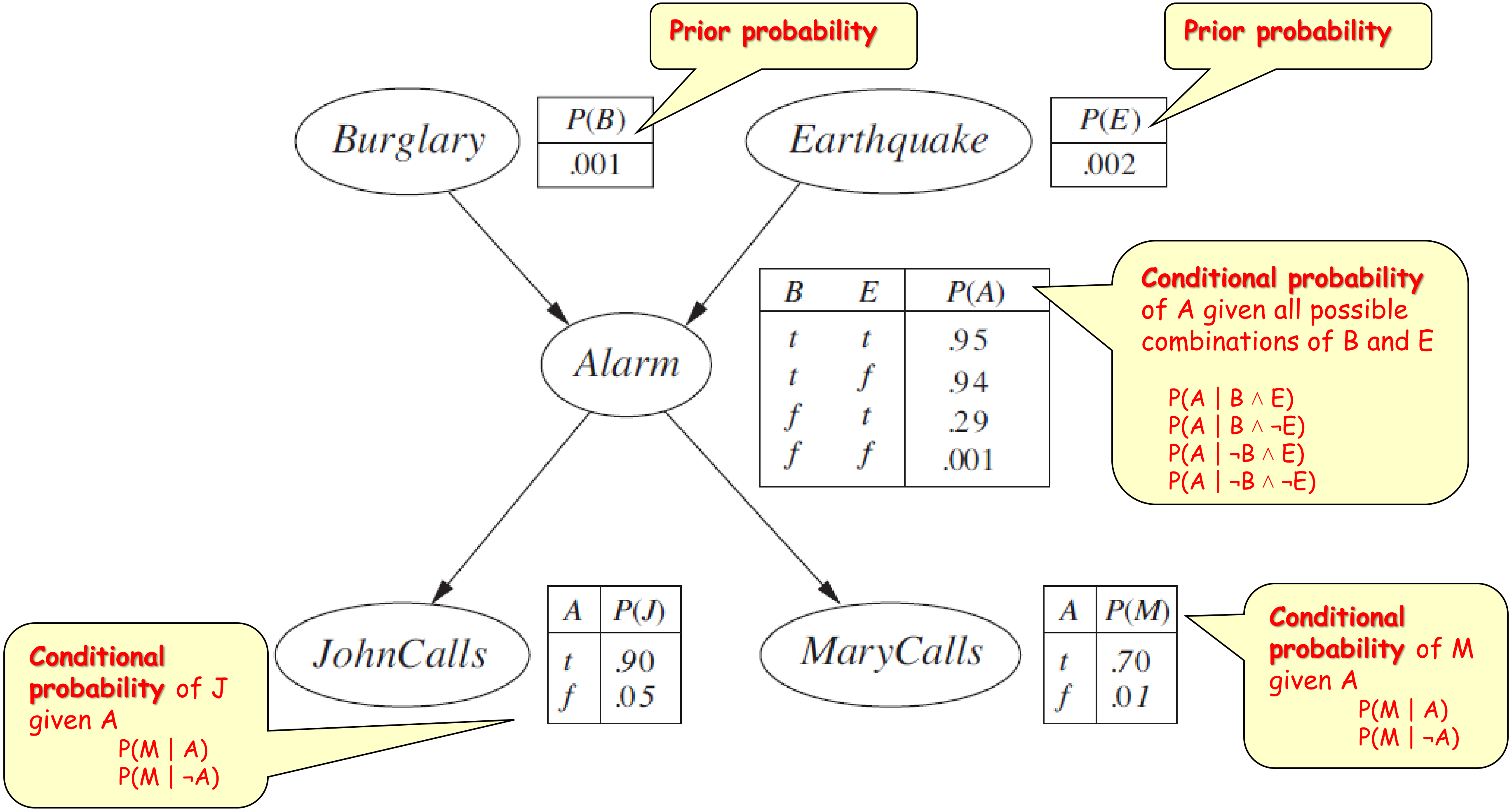
# Bayesian network *(another example)*

- Both the **topology** and the **conditional probability tables** (CPTs)



# Bayesian network *(another example)*

- Both the **topology** and the **conditional probability tables** (CPTs)




| $P(B)$ |
|---|
| .001 |

| $P(E)$ |
|---|
| .002 |

| $B$ | $E$ | $P(A)$ |
|---|---|---|
| $t$ | $t$ | .95 |
| $t$ | $f$ | .94 |
| $f$ | $t$ | .29 |
| $f$ | $f$ | .001 |

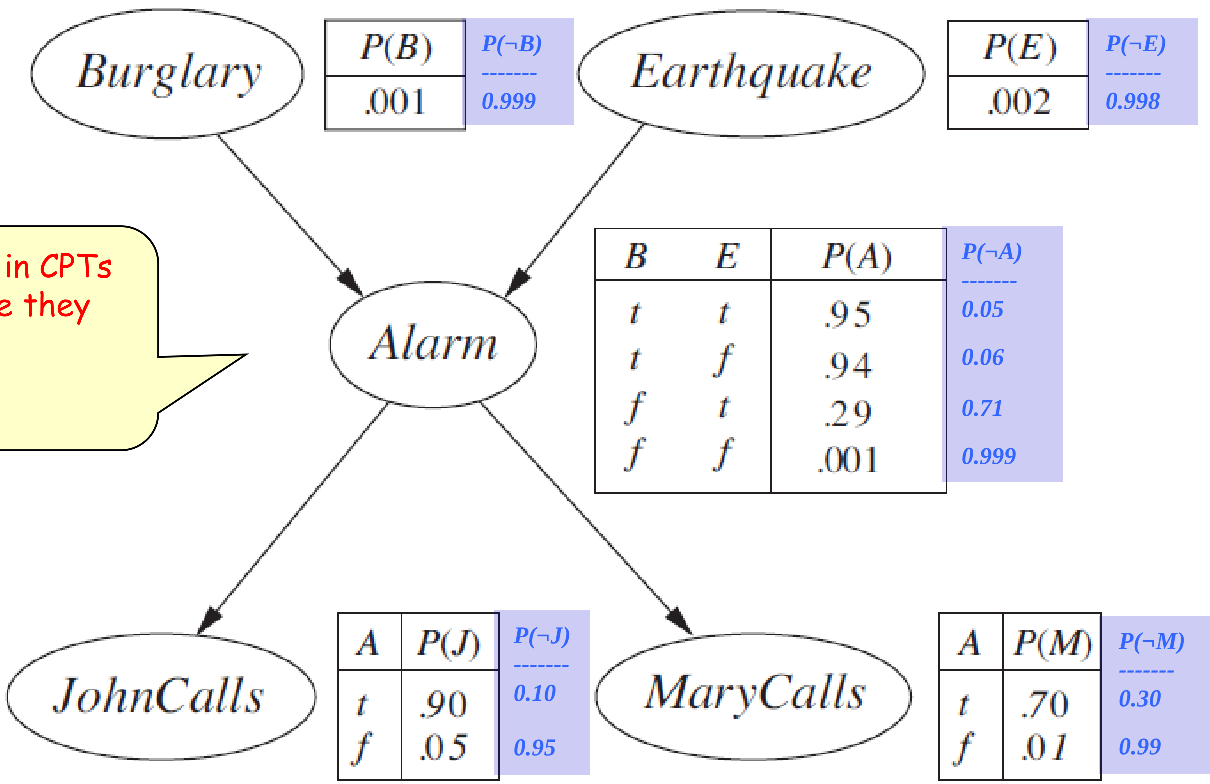
| $A$ | $P(J)$ |
|---|---|
| $t$ | .90 |
| $f$ | .05 |

| $A$ | $P(M)$ |
|---|---|
| $t$ | .70 |
| $f$ | .01 |

# Bayesian network *(another example)*

- Both the **topology** and the **conditional probability tables** (CPTs)

Half of the entries in CPTs are omitted because they can be inferred




| $P(B)$ | $P(\neg B)$ |
|---|---|
| .001 | 0.999 |

| $P(E)$ | $P(\neg E)$ |
|---|---|
| .002 | 0.998 |

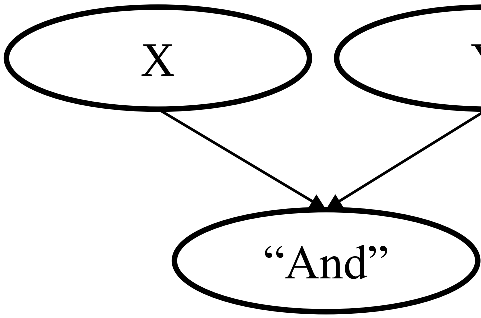
| $B$ | $E$ | $P(A)$ | $P(\neg A)$ |
|---|---|---|---|
| $t$ | $t$ | .95 | 0.05 |
| $t$ | $f$ | .94 | 0.06 |
| $f$ | $t$ | .29 | 0.71 |
| $f$ | $f$ | .001 | 0.999 |

| $A$ | $P(J)$ | $P(\neg J)$ |
|---|---|---|
| $t$ | .90 | 0.10 |
| $f$ | .05 | 0.95 |

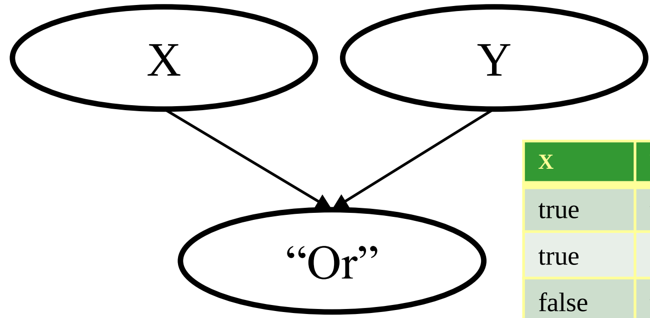
| $A$ | $P(M)$ | $P(\neg M)$ |
|---|---|---|
| $t$ | .70 | 0.30 |
| $f$ | .01 | 0.99 |

# Bayesian networks for Boolean functions

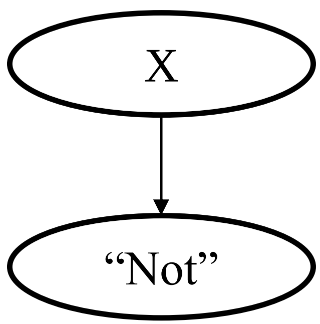
- Can Bayesian networks represent all **Boolean functions**? – Yes.
  *f(Feature_1, …, Feature_n) ≡ some propositional sentence*




| X | Y | P("And" \| X, Y) |
|---|---|---|
| true | true | 1.0 |
| true | false | 0.0 |
| false | true | 0.0 |
| false | false | 0.0 |



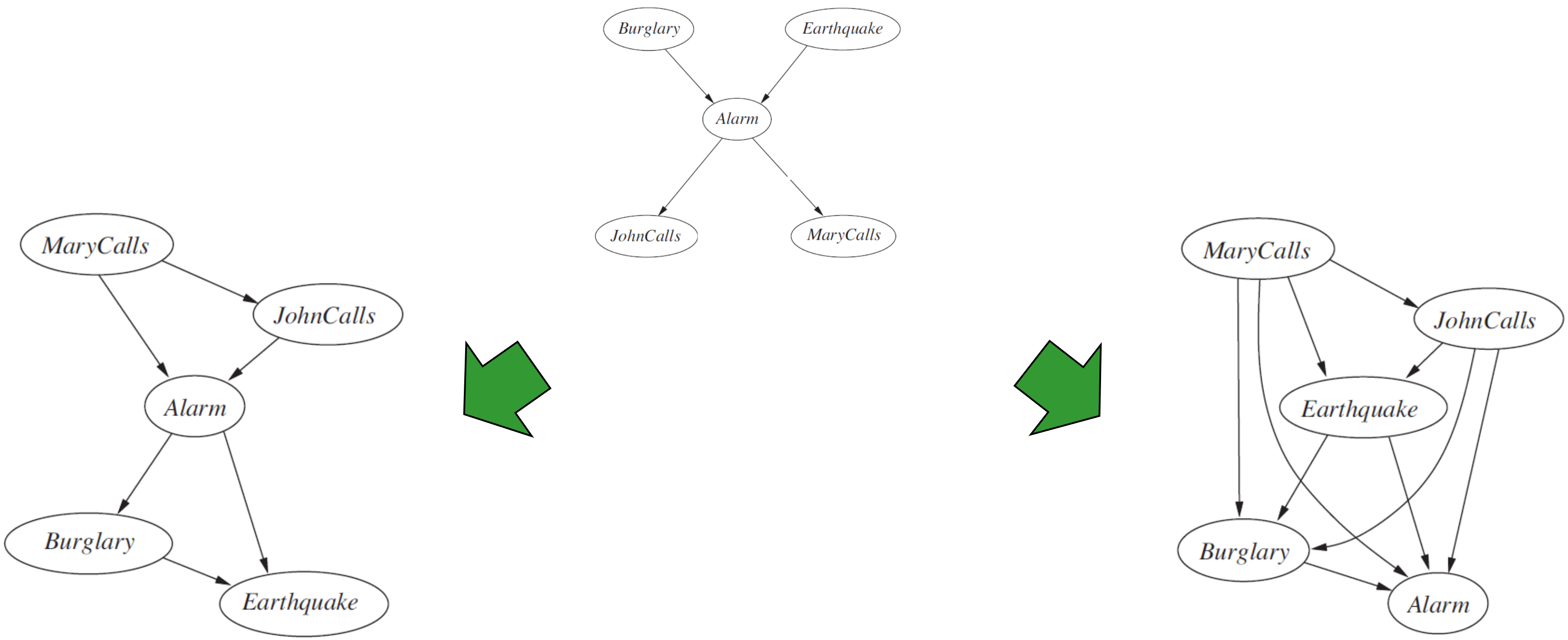

| X | Y | P("Or" \| X, Y) |
|---|---|---|
| true | true | 1.0 |
| true | false | 1.0 |
| false | true | 1.0 |
| false | false | 0.0 |




| X | P("Not" \| X) |
|---|---|
| true | 0.0 |
| false | 1.0 |

# Compactness and node ordering

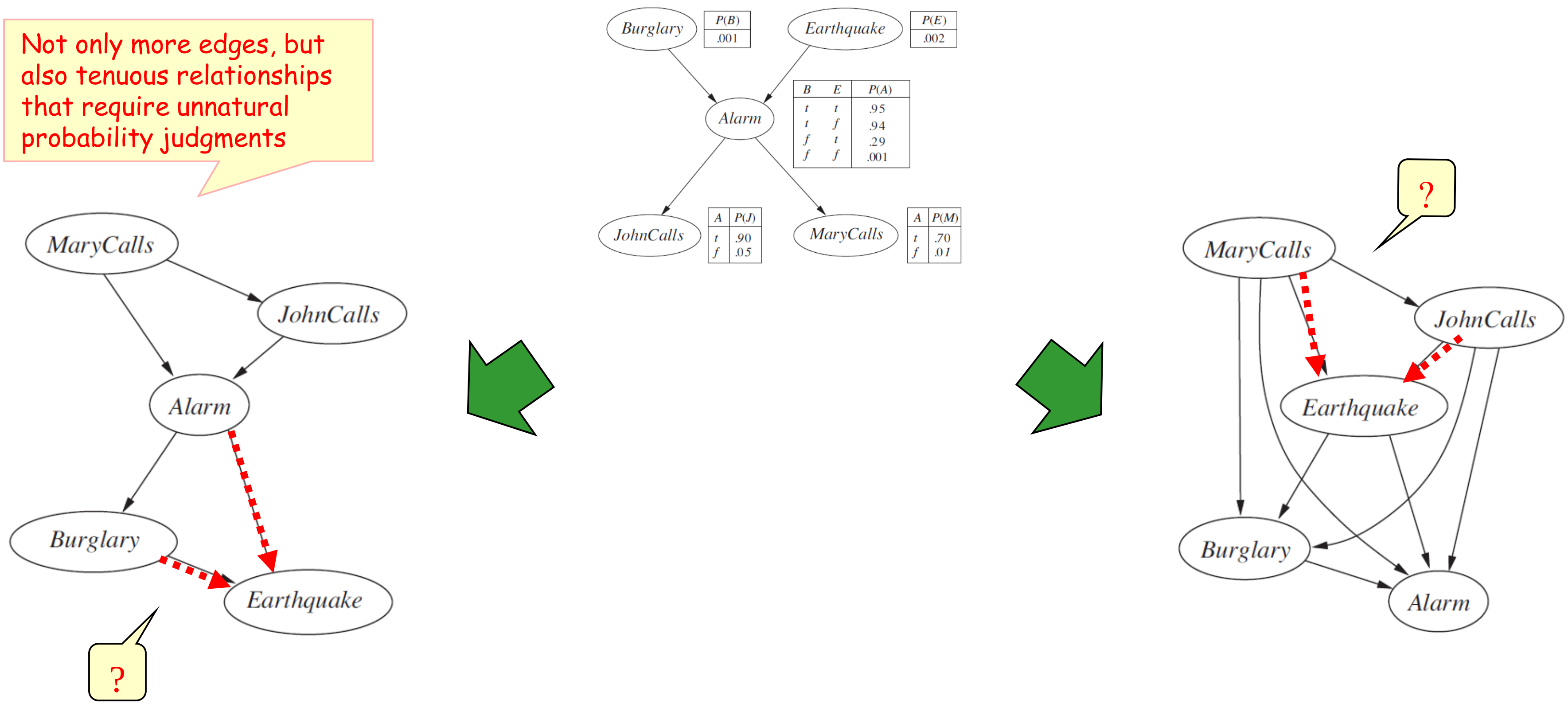
- There are multiple, **equivalent**, Bayesian networks, some of which are more **compact** than the others




Compactness depend on the node ordering

# Compactness and node ordering

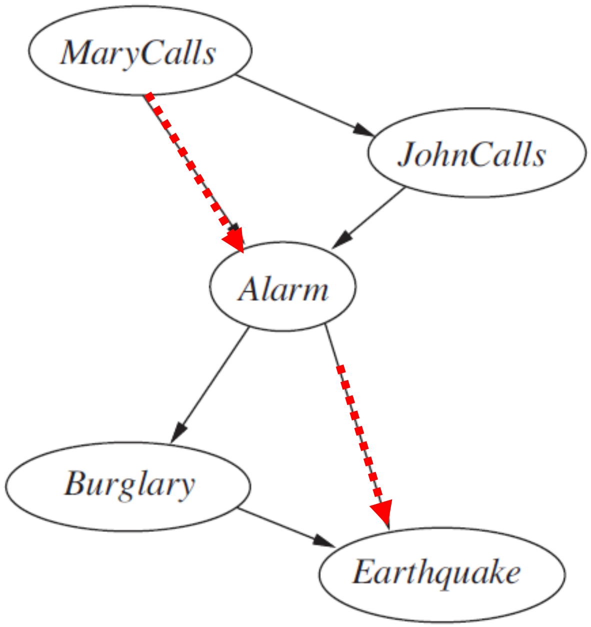
- There are multiple, **equivalent**, Bayesian networks, some of which are more **compact** than the others




Compactness depend on the node ordering

# Compactness and node ordering

- Distinction between **causal** model and **diagnostic** model
  - If we try to build a **diagnostic model**, with links **from symptoms to causes**, we have to specify additional dependencies between otherwise independent causes
  - Example: from ***MaryCalls*** to ***Alarm***, or from ***Alarm*** to ***Burglary***




**Solution**: stick to a *causal model*

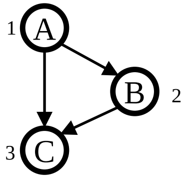
*(1) Fewer dependences*
*(2) Easier to come up with probability*

# Let’s go through an example

- Note: each way of **factoring joint distribution** corresponds to a **different Bayesian network**

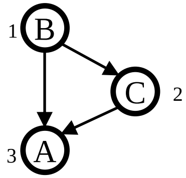
- Example: 6 ways of factoring P(A, B, C), including

  – P(A, B, C) = P(C | B, A) P(B, A) = P(C | B, A) P(B | A) P(A)




(First picking A, then picking B, and finally picking C, each time conditioning the picked random variable on all random variables picked earlier)

  – P(A, B, C) = P(A | B, C) P(B, C) = P(A | B, C) P(C | B) P(B)



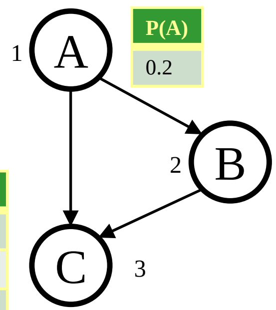

(First picking B, then picking C and finally picking A, each time conditioning the picked random variable on all random variables picked earlier)

# Bayesian network is not unique

- The network topology determines how many probabilities need to be specified for the conditional probability tables.
  - Let's choose P(A, B, C) = P(C | B, A) P(B | A) P(A).

| A | B | C | P(A, B, C) |
|---|---|---|---|
| true | true | true | 0.054 |
| true | true | false | 0.126 |
| true | false | true | 0.002 |
| true | false | false | 0.018 |
| false | true | true | 0.432 |
| false | true | false | 0.288 |
| false | false | true | 0.032 |
| false | false | false | 0.048 |

| A | B | P(C \| A, B) |
|---|---|---|
| true | true | 0.3 |
| true | false | 0.1 |
| false | true | 0.6 |
| false | false | 0.4 |




| P(A) |
|---|
| 0.2 |

| A | P(B \| A) |
|---|---|
| true | **0.9** |
| false | **0.9** |

# Independence detected

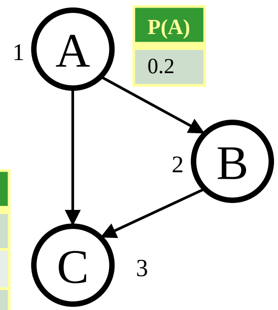
- Here: $P(B \mid A) = P(B \mid \neg A)$.
- Thus, A and B are **independent**

- Detailed explanation
  - $P(B) = P(B \wedge A) + P(B \wedge \neg A)$

$$= P(B \mid A)\, P(A) + P(B \mid \neg A)\, P(\neg A)$$
$$= P(B \mid A)\, P(A) + P(B \mid A)\, P(\neg A)$$
$$= P(B \mid A)\, (P(A) + P(\neg A))$$
$$= P(B \mid A)$$

# Simplifying Bayesian network

- **Independence** allows us to **simplify** Bayesian network

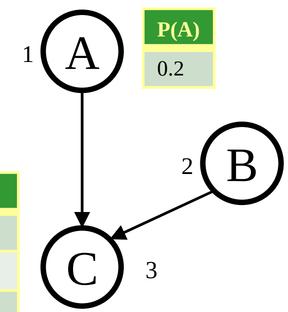



| P(A) |
|---|
| 0.2 |

| A | P(B \| A) |
|---|---|
| true | 0.9 |
| false | 0.9 |

| A | B | P(C \| A, B) |
|---|---|---|
| true | true | 0.3 |
| true | false | 0.1 |
| false | true | 0.6 |
| false | false | 0.4 |

Need to specify 7 probabilities for all conditional probability tables




| P(A) |
|---|
| 0.2 |

| P(B) |
|---|
| 0.9 |

| A | B | P(C \| A, B) |
|---|---|---|
| true | true | 0.3 |
| true | false | 0.1 |
| false | true | 0.6 |
| false | false | 0.4 |

Need to specify only 6 probabilities for all conditional probability tables

# Semantics of Bayesian networks

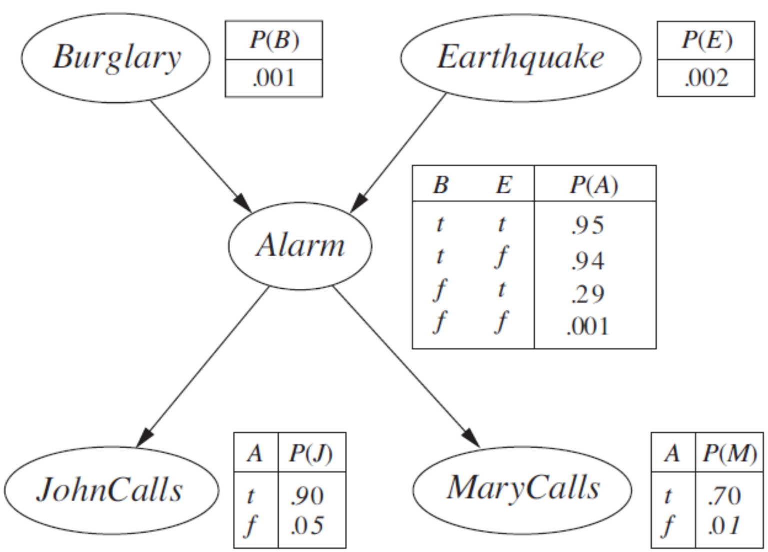
- Two different, but equivalent **views**
  - Representation of the **joint probability distribution**
  - Encoding of a collection of **conditional independence** statements

- How to construct networks
- How to design inference procedures

# Bayesian network: *Representing the full joint distribution*

- Each entry $P(x_1,\ldots,x_n)$ in the full joint distribution, which is the abbreviation of $P(X_1=x_1 \wedge \ldots \wedge X_n=x_n)$, is the product of the elements of the CPTs defined as follows:

$$\boxed{P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))}$$

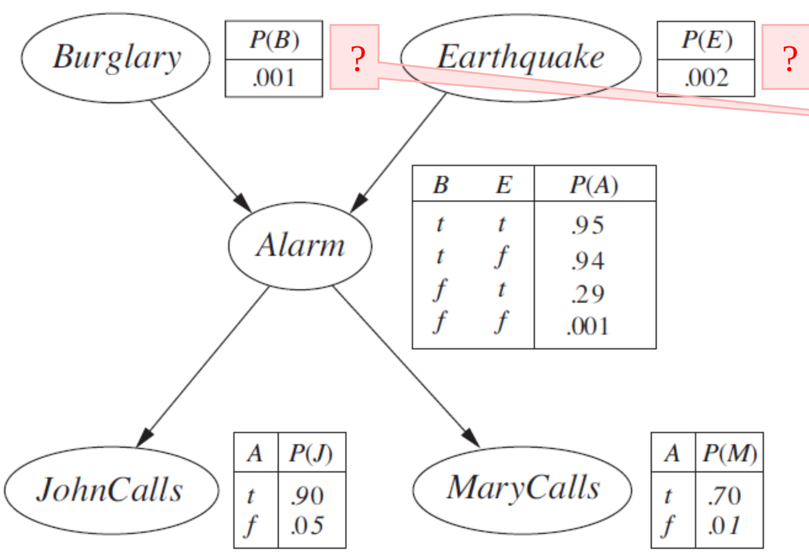



$$\boxed{P(j, m, a, \neg b, \neg e)}$$

# Bayesian network: *Representing the full joint distribution*

- Each entry $P(x_1,\ldots,x_n)$ in the full joint distribution, which is the abbreviation of $P(X_1{=}x_1 \wedge \ldots \wedge X_n{=}x_n)$, is the product of the elements of the CPTs defined as follows:

$$\boxed{P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \,|\, parents(X_i))}$$



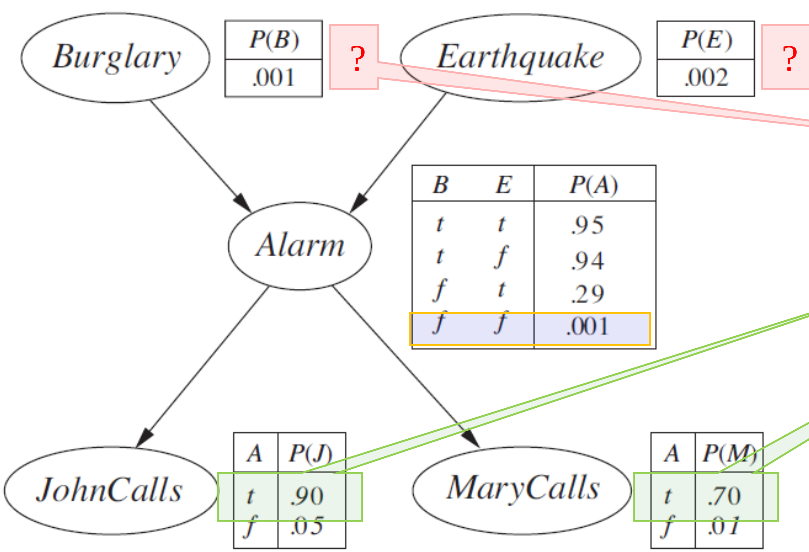

$$\boxed{P(j,m,a,\neg b,\neg e)}$$

$$= P(j \,|\, a)P(m \,|\, a)P(a \,|\, \neg b \wedge \neg e)P(\neg b)P(\neg e)$$

# Bayesian network: *Representing the full joint distribution*

- Each entry $P(x_1,\ldots,x_n)$ in the full joint distribution, which is the abbreviation of $P(X_1=x_1 \wedge \ldots \wedge X_n=x_n)$, is the product of the elements of the CPTs defined as follows:

$$\boxed{P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))}$$



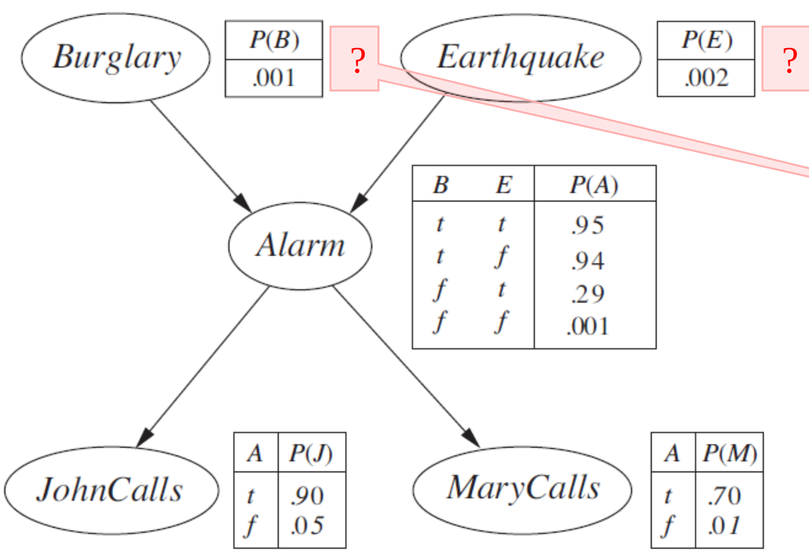

$$\boxed{P(j, m, a, \neg b, \neg e)}$$

$$= P(j \mid a)P(m \mid a)P(\boxed{a \mid \neg b \wedge \neg e})P(\neg b)P(\neg e)$$

# Bayesian network: *Representing the full joint distribution*

- Each entry $P(x_1,\ldots,x_n)$ in the full joint distribution, which is the abbreviation of $P(X_1=x_1 \wedge \ldots \wedge X_n=x_n)$ is the product of the elements of the CPTs defined as follows:

$$\boxed{P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))}$$




$$\boxed{P(j, m, a, \neg b, \neg e)}$$
$$= P(j \mid a)P(m \mid a)P(a \mid \neg b \wedge \neg e)P(\neg b)P(\neg e)$$
$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628\ .$$

# How to construct Bayesian networks

- Starting from the full joint distribution, first, we rewrite the entries in terms of conditional probability

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1}, \ldots, x_1)$$

- Then, we repeat the process

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1) \,. \end{aligned}$$

Each node must be **conditionally independent** of its other predecessors in the node ordering, given its parents

- Now, compare to Bayesian network

$$\boxed{P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))}$$

$$\mathbf{P}(X_i \mid \boxed{X_{i-1}, \ldots, X_1}) = \mathbf{P}(X_i \mid \boxed{Parents(X_i)})$$

# How to construct Bayesian networks *(cont'd)*

- Starting from the full joint distribution, first, we rewrite the entries in terms of conditional probability

$$
\begin{aligned}
P(x_1,\ldots,x_n) &= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1} \mid x_{n-2},\ldots,x_1) \cdots P(x_2 \mid x_1)P(x_1) \\
&= \prod_{i=1}^{n} P(x_i \mid x_{i-1},\ldots,x_1) .
\end{aligned}
$$

- Now, compare to Bayesian network

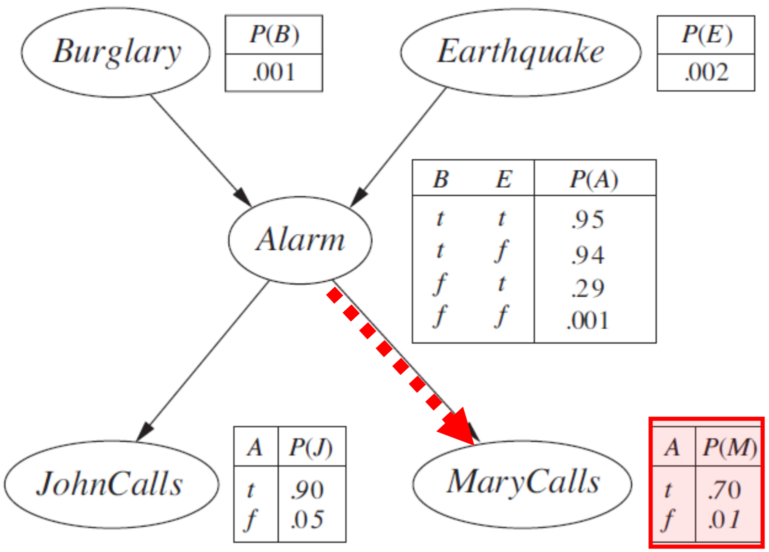
$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

Each node must be **conditionally independent** of its other predecessors in the node ordering, given its parents

$$\mathbf{P}(X_i \mid X_{i-1},\ldots,X_1) = \mathbf{P}(X_i \mid Parents(X_i))$$

- For ***i* = *1*** to ***n*** do:
  - Choose, from *$X_1$, …, $X_{i-1}$*, a minimum set of parents for *$X_i$*, to satisfy the equation
  - Insert edges from the parents to *$X_i$*
  - CTPs: write down the conditional probability table, *$P(X_i \mid Parents(X_i))$*

# How to construct Bayesian networks *(cont'd)*



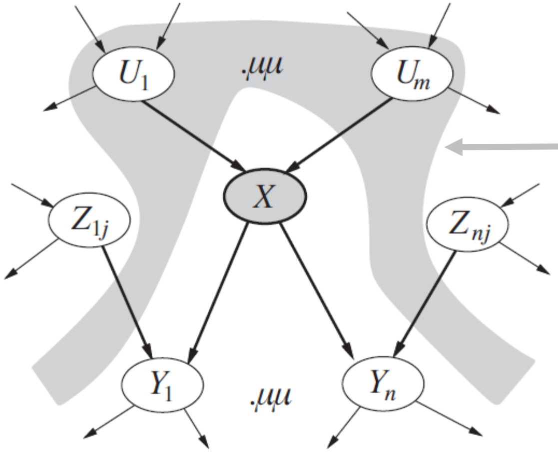

*P(MaryCalls | JohnCalls, Alarm, Earthquake, Burglary)*
*= P(MaryCalls | Alarm)*

$$\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid Parents(X_i))$$

- For ***i* = 1** to ***n*** do:
  - Choose, from $X_1, \ldots, X_{i-1}$, a minimum set of parents for $X_i$, to satisfy the equation
  - Insert edges from the parents to $X_i$
  - CTPs: write down the conditional probability table, $P(X_i \mid Parents(X_i))$

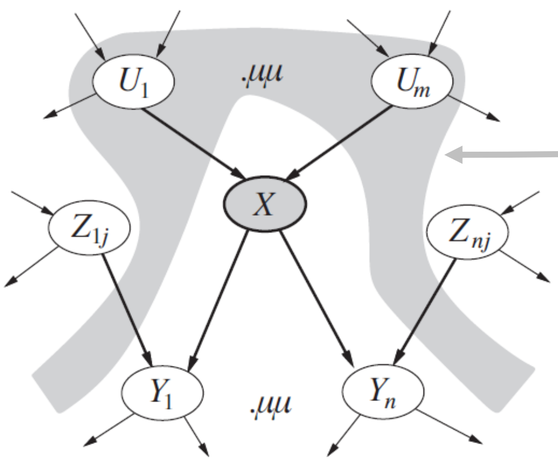
# Semantics of Bayesian networks

- Two different, but equivalent views
  - Representation of the joint probability distribution
  - Encoding of a collection of **conditional independence** statements
- How to construct Bayesian networks
  - a node is conditionally independent of its other predecessors, given its parents
- X is conditionally independent of its non-descendants, given its parents



# Semantics of Bayesian networks

- Two different, but equivalent views
  - Representation of the joint probability distribution
  - Encoding of a collection of **conditional independence** statements
- How to construct Bayesian networks
  - a node is conditionally independent of its other predecessors, given its parents
- X is conditionally independent of its **non-descendants**, given **its parents**



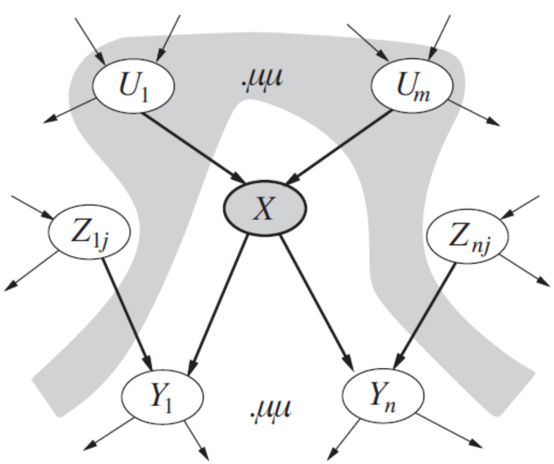

X may even be conditionally independent of some of these descendants

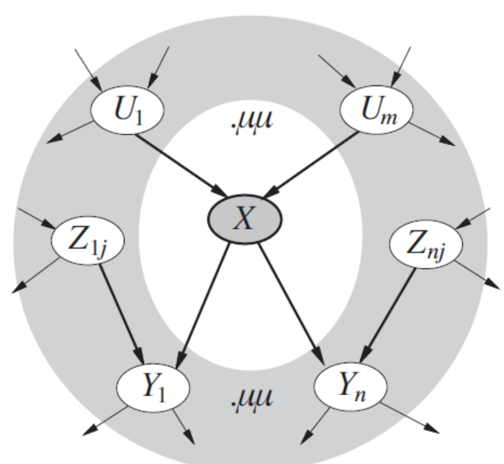
# Semantics of Bayesian networks

- Two different, but equivalent views
  - Representation of the joint probability distribution
  - Encoding of a collection of conditional independence statements

Markov blanket of X includes
(1) its parents
(2) its children,
(3) its children's parents

- How to construct Bayesian networks
  - a node is conditionally independent of its other predeces
- X is conditionally independent of all other nodes, given Markov blanket



# Summary

- Bayesian network is a well-developed representation for uncertain knowledge
  - It is often **exponentially smaller** than full joint distribution
  - It plays similar role as propositional logic for definite knowledge

- Bayesian network is a DAG where
  - a node denotes a random variable, together with local conditional distribution of that variable, given its parents
  - It's a **concise representation** of conditional independence
  - It represents full joint distribution, as product of corresponding entries in the local conditional distributions