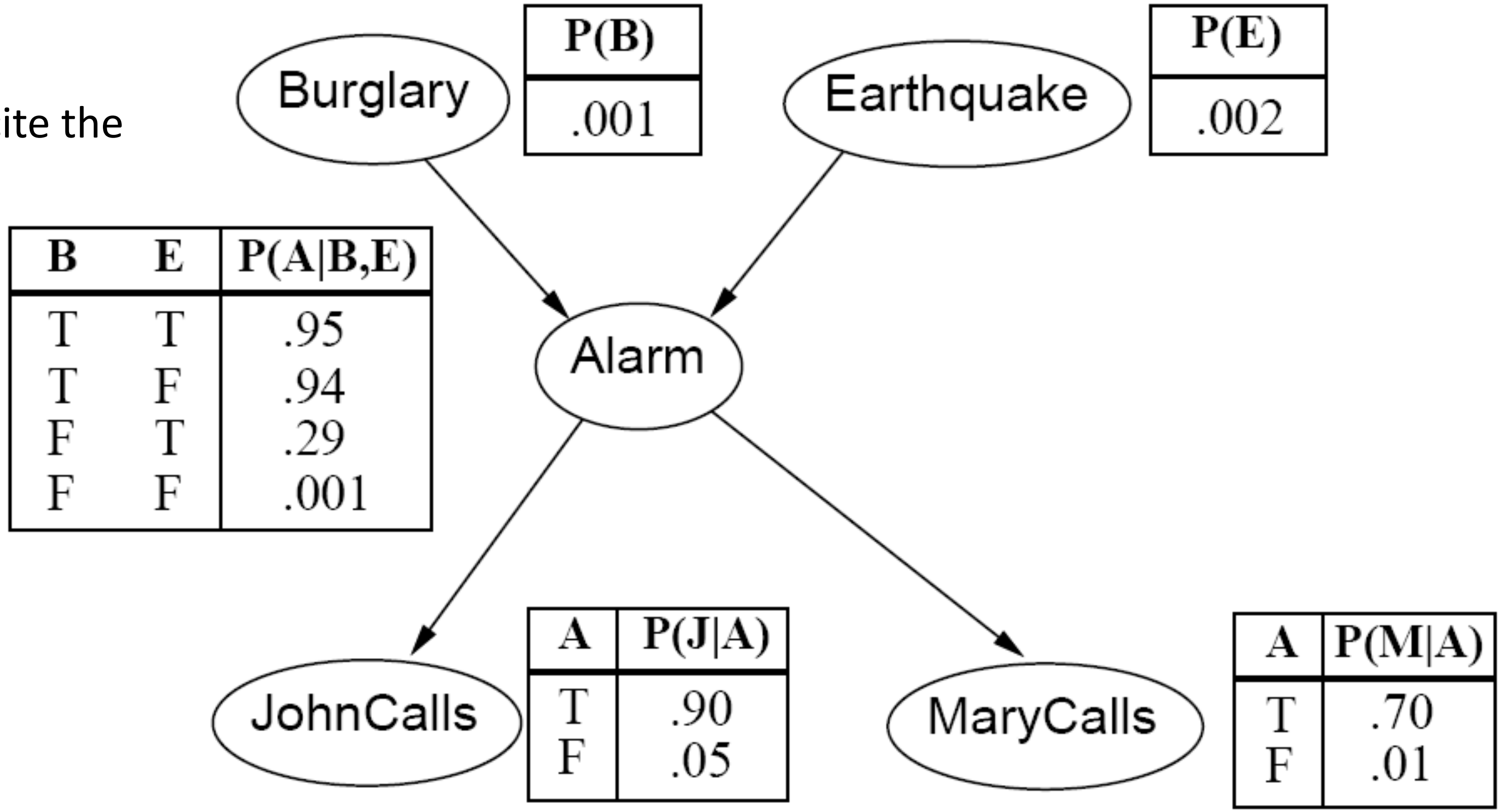
CS440/ECE448 Lecture 15: Bayesian Networks

By Mark Hasegawa-Johnson, 2/2020

With some slides by Svetlana Lazebnik, 9/2017



Review: Bayesian inference

- A general scenario:
  - Query *variables:* **X**
  - *Evidence* (*observed*) variables and their values: **E** = **e**
- **Inference problem**: answer questions about the query variables given the evidence variables
- This can be done using the posterior distribution P(**X** | **E** = **e**)
- Example of a useful question: **Which X is true?**
  - More formally: what value of **X** has the least probability of being wrong?
  - Answer: **MPE = MAP** (argmin P(error) = argmax P(X=x|E=e))

Today: What if P(X,E) is complicated?

- Very, very common problem: P(X,E) is complicated because both X and E depend on some hidden variable Y
- SOLUTION:
  - Draw a bunch of circles and arrows that represent the dependence
  - When your algorithm performs inference, make sure it does so in the order of the graph
- FORMALISM: Bayesian Network

Hidden Variables

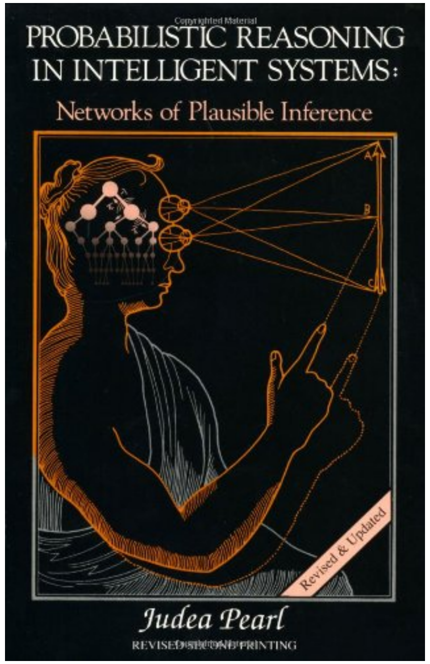
- A general scenario:
  - Query *variables:* **X**
  - *Evidence* (*observed*) variables and their values: **E** = **e**
  - *Unobserved* variables: **Y**
- **Inference problem**: answer questions about the query variables given the evidence variables
  - This can be done using the posterior distribution P(**X** | **E** = **e**)
  - In turn, the posterior needs to be derived from the full joint P(**X**, **E**, **Y**)

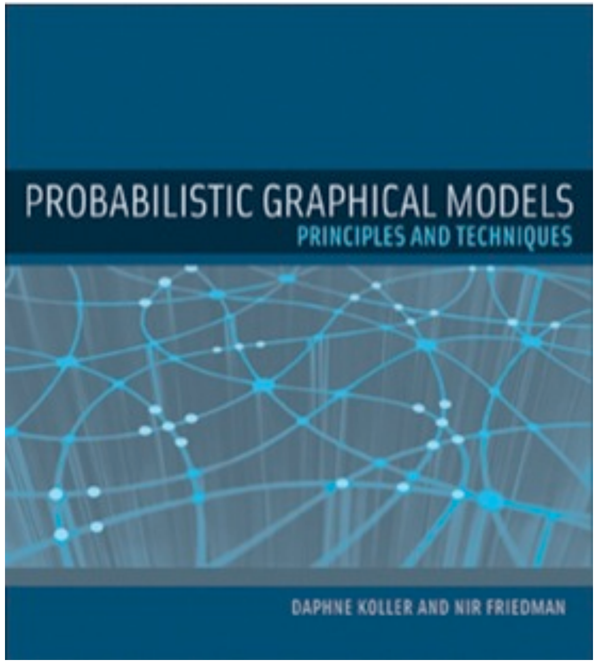
$$P(\boldsymbol{X} \mid \boldsymbol{E} = \boldsymbol{e}) = \frac{P(\boldsymbol{X}, \boldsymbol{e})}{P(\boldsymbol{e})} \propto \sum_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{e}, \boldsymbol{y})$$

- Bayesian networks are a tool for representing joint probability distributions efficiently

Bayesian networks

- More commonly called *graphical models*
- A way to depict conditional independence relationships between random variables
- A compact specification of full joint distributions

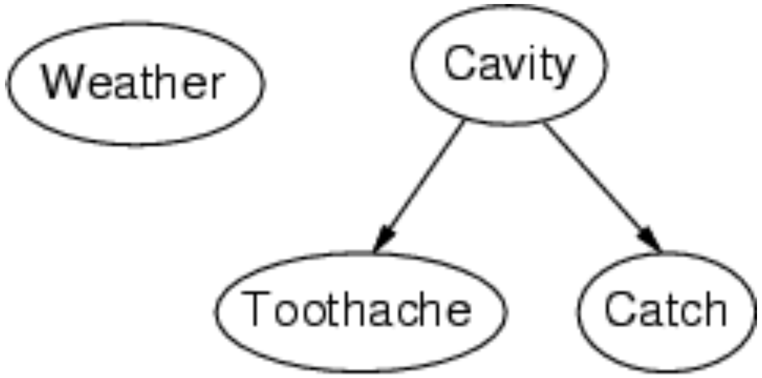




Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Inference in a Bayes network
- Conditional independence ≠ Independence

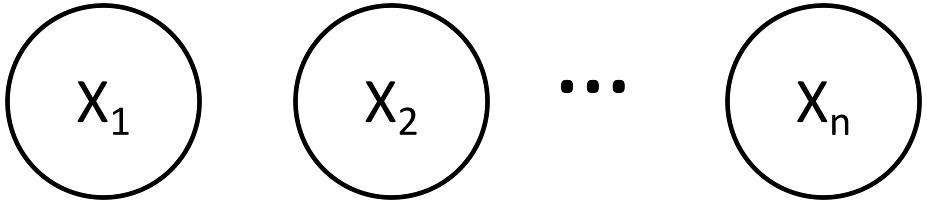
Bayesian networks: Structure



- **Nodes:** random variables
- **Arcs:** interactions
  - An arrow from one variable to another indicates direct influence
  - Must form a directed, *acyclic* graph

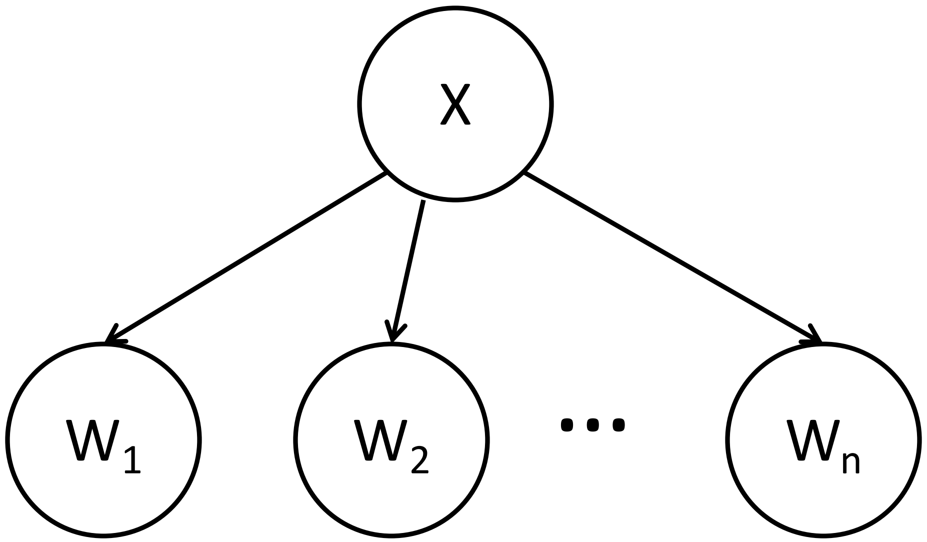
Example: N independent coin flips

- Complete independence: no interactions



Example: Naïve Bayes document model

- Random variables:
  - X: document class
  - $W_1, \ldots, W_n$: words in the document



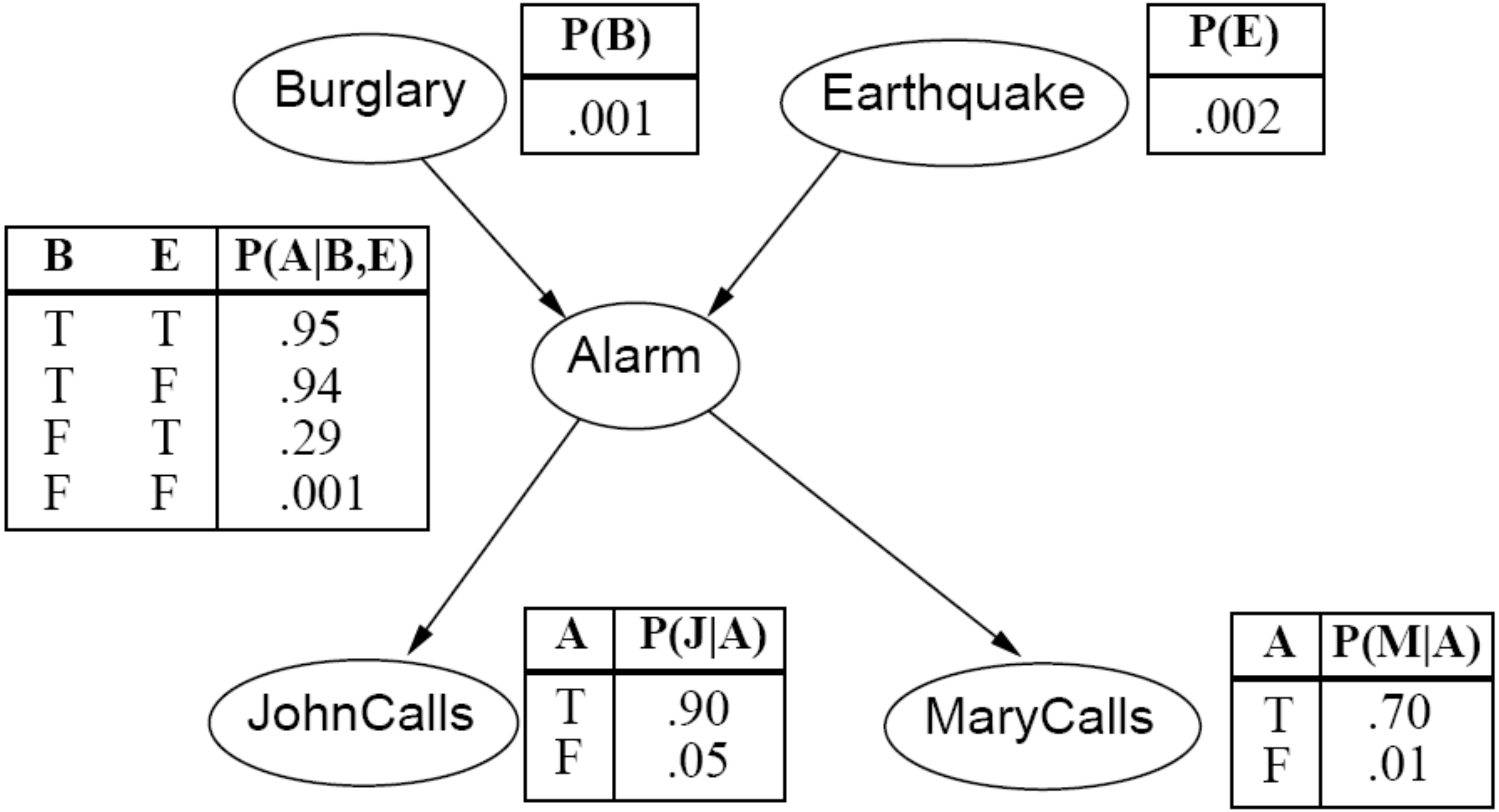
Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Inference in a Bayes network
- Conditional independence ≠ Independence

Example: Los Angeles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn’t call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Burglar Alarm



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Conditional independence and the joint distribution

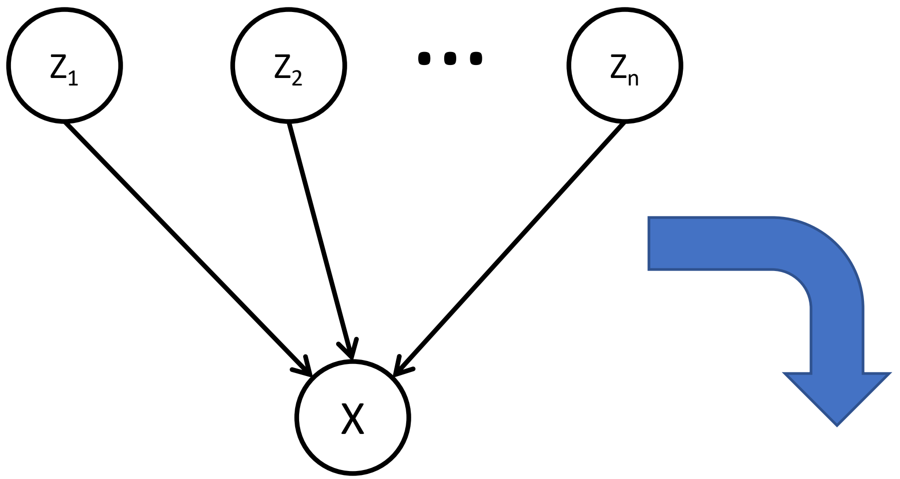
- Key property: each node is conditionally independent of its *non-descendants* given its *parents*
- Suppose the nodes $X_1, \ldots, X_n$ are sorted in topological order
- To get the joint distribution $P(X_1, \ldots, X_n)$, use chain rule:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid X_1,\ldots,X_{i-1}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i \mid Parents(X_i)\right)$$

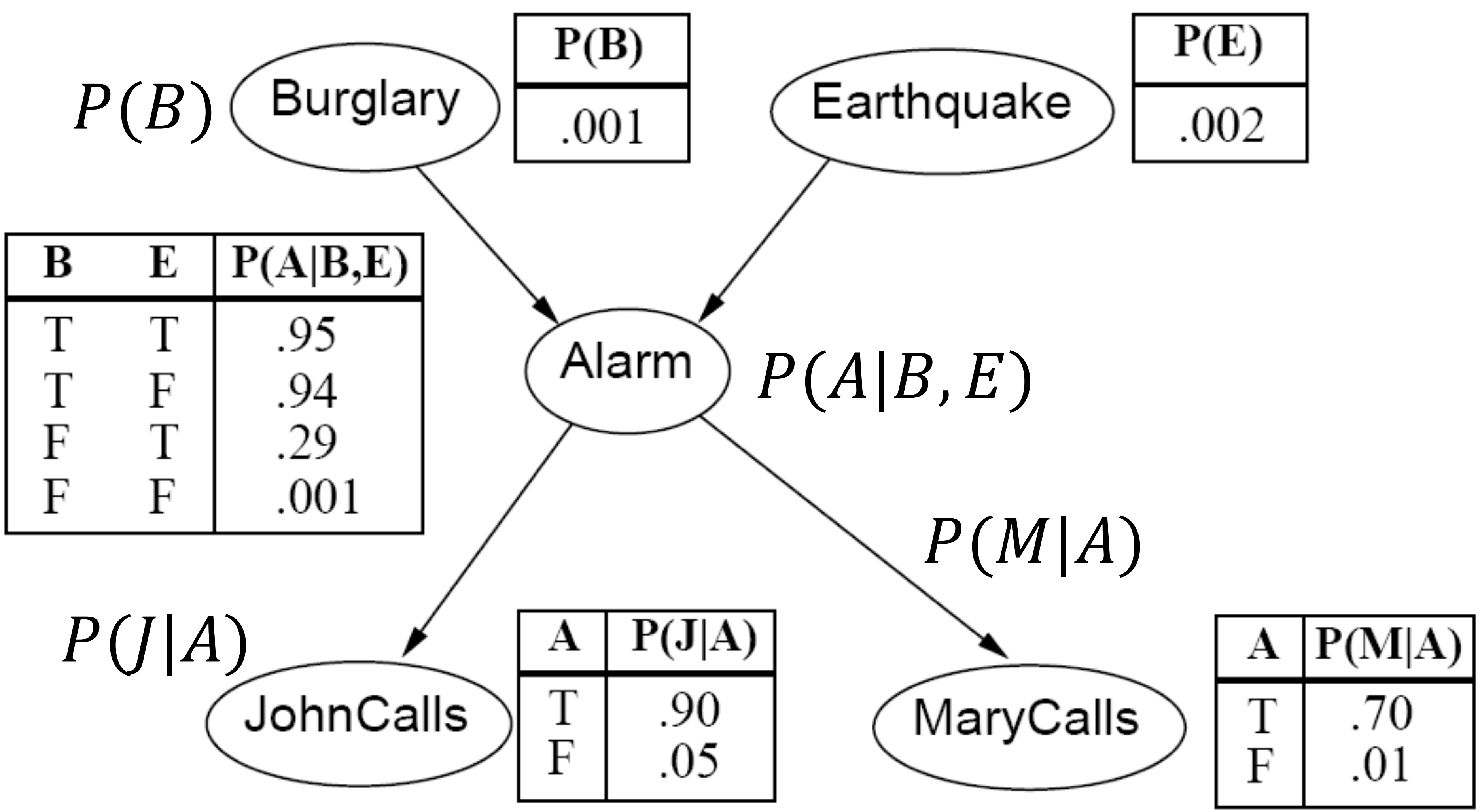
Conditional probability distributions

- To specify the full joint distribution, we need to specify a *conditional* distribution for each node given its parents: $P\ (X \mid \text{Parents}(X))$



$P\ (X \mid Z_1, \ldots, Z_n)$

Example: Burglar Alarm



P(B)
.001

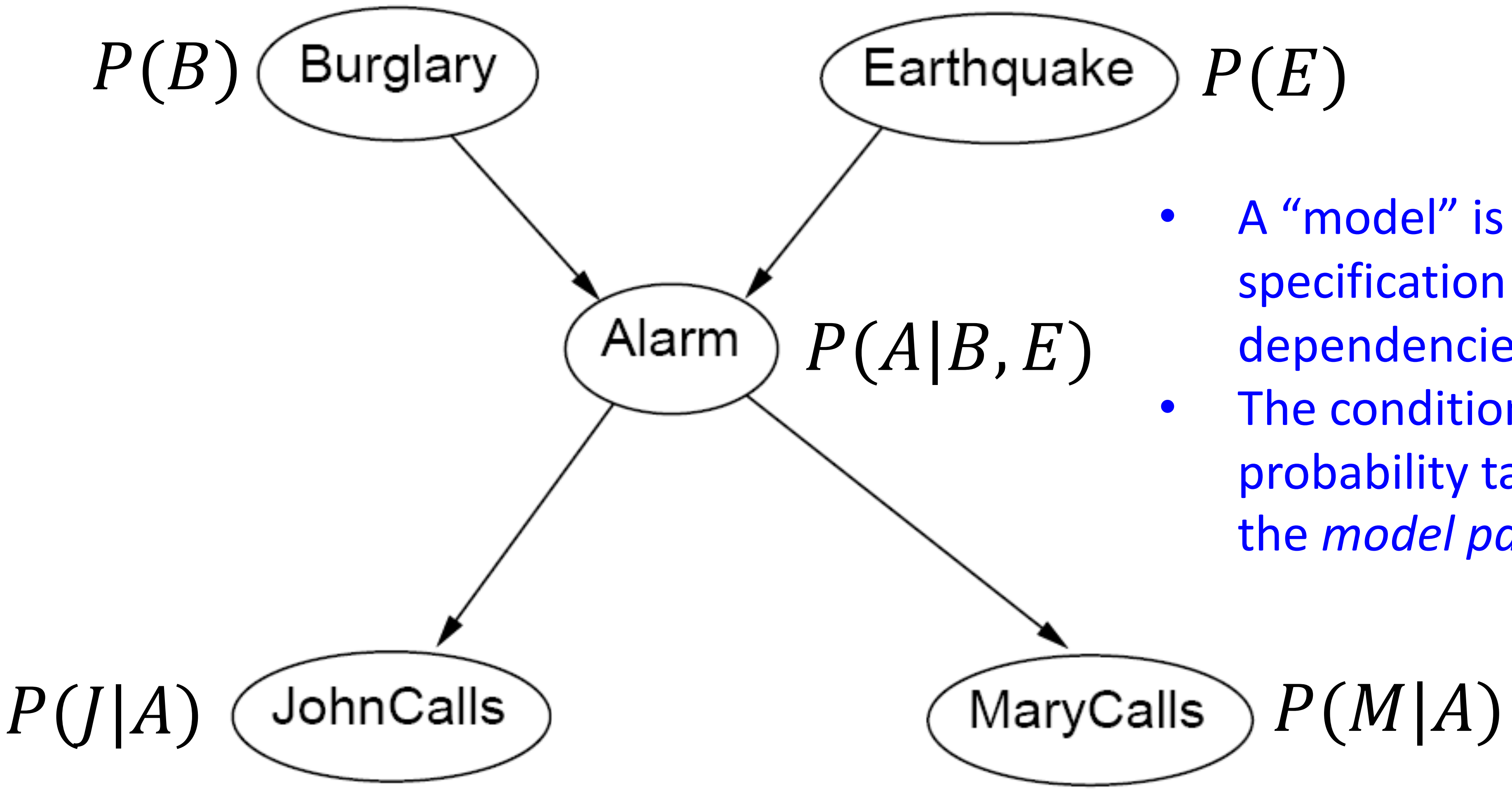
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Example: Burglar Alarm



- A "model" is a complete specification of the dependencies.
- The conditional probability tables are the *model parameters.*

Outline

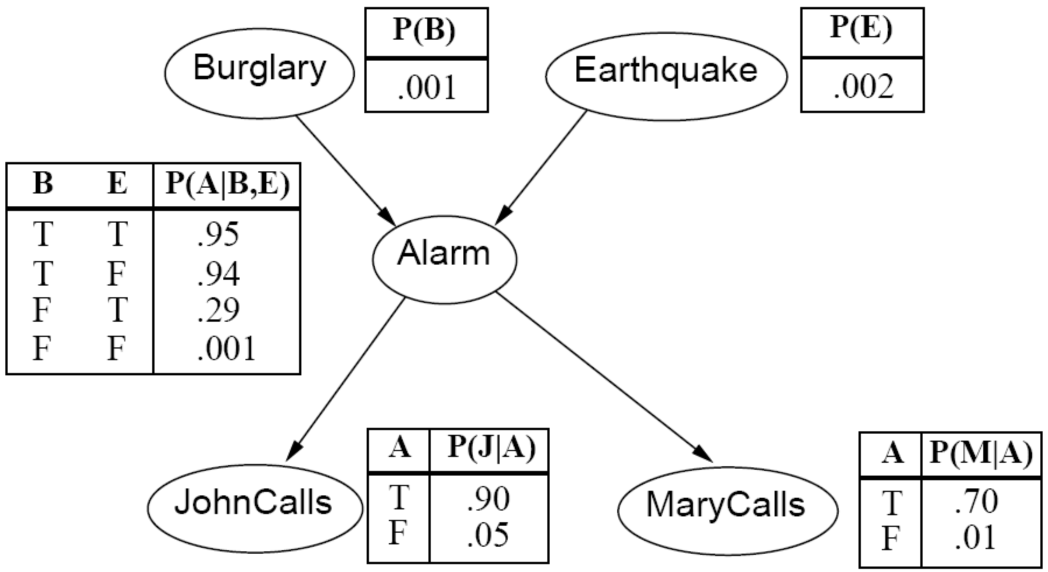
- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Inference in a Bayes network
- Conditional independence ≠ Independence

Classification using probabilities

- Suppose Mary has called to tell you that you had a burglar alarm. Should you call the police?
  - Make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision. You decide that you have a burglar in your house if and only if

$$P(Burglary|Mary) > P(\neg Burglary|Mary)$$

Using a Bayes network to estimate a posteriori probabilities

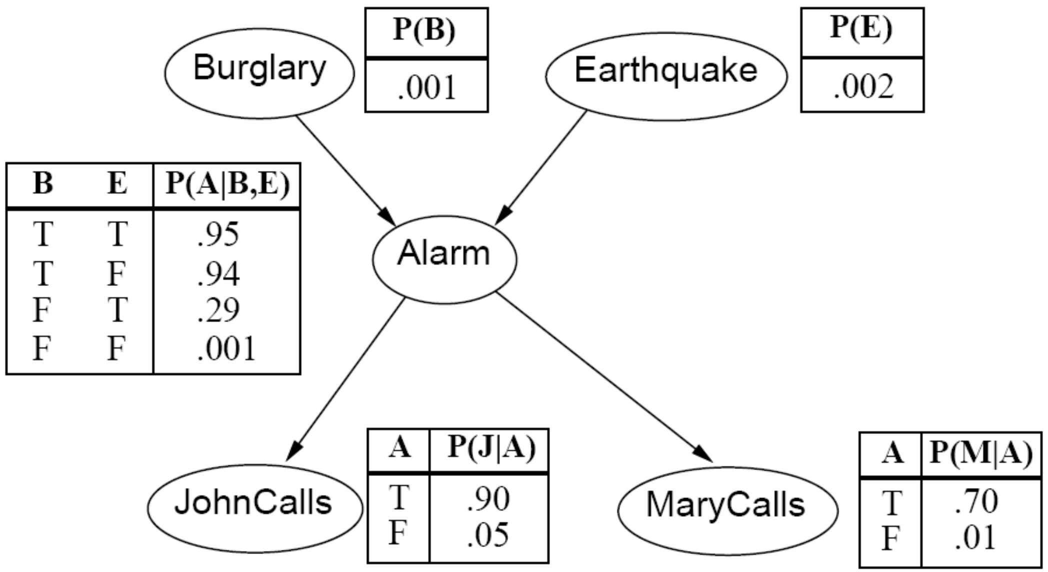


- Notice: we don't know $P(Burglary|Mary)$! We have to figure out what it is.
- This is called "inference".
- First step: find the joint probability of B (and $\neg B$), M (and $\neg M$), and any other variables that are necessary in order to link these two together.

$$P(B,E,A,M) = P(B)P(E)P(A|B,E)P(M|A)$$

$P(BEAM)$	$\neg M, \neg A$	$\neg M, A$	$M, \neg A$	M, A
$\neg B, \neg E$	0.986045	2.99×10^{-4}	9.96×10^{-3}	6.98×10^{-4}
$\neg B, E$	1.4×10^{-3}	1.7×10^{-4}	1.4×10^{-5}	4.06×10^{-4}
$B, \neg E$	5.93×10^{-5}	2.81×10^{-4}	5.99×10^{-7}	6.57×10^{-4}
B, E	9.9×10^{-8}	5.7×10^{-7}	10^{-9}	1.33×10^{-6}

Using a Bayes network to estimate a posteriori probabilities

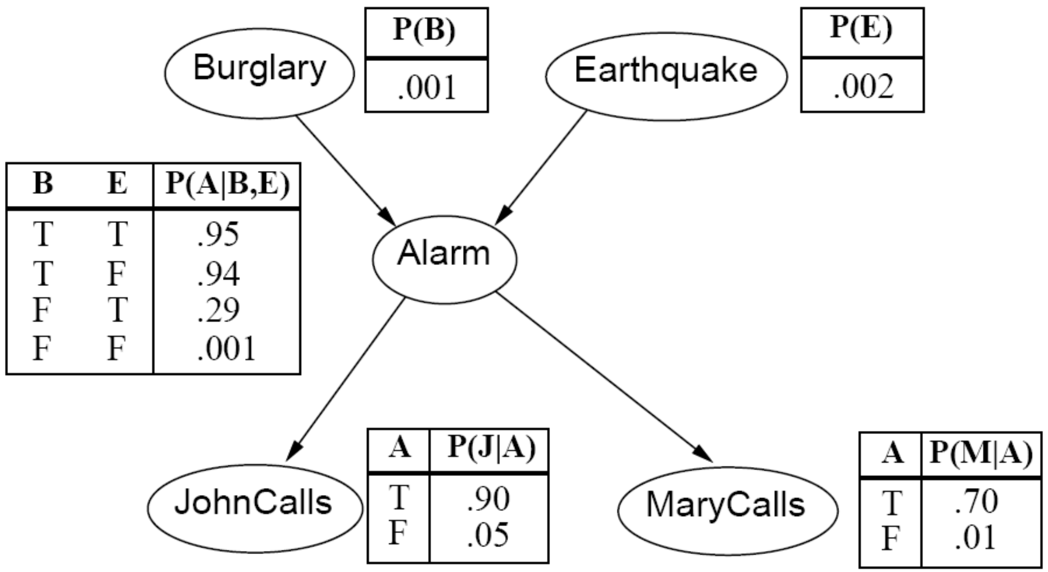


- Second step: marginalize (add) to get rid of the variables you don't care about.

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

$P(B,M)$	$\neg M$	$\boldsymbol{M}$
$\neg B$	0.987922	0.011078
B	0.000341	0.000659

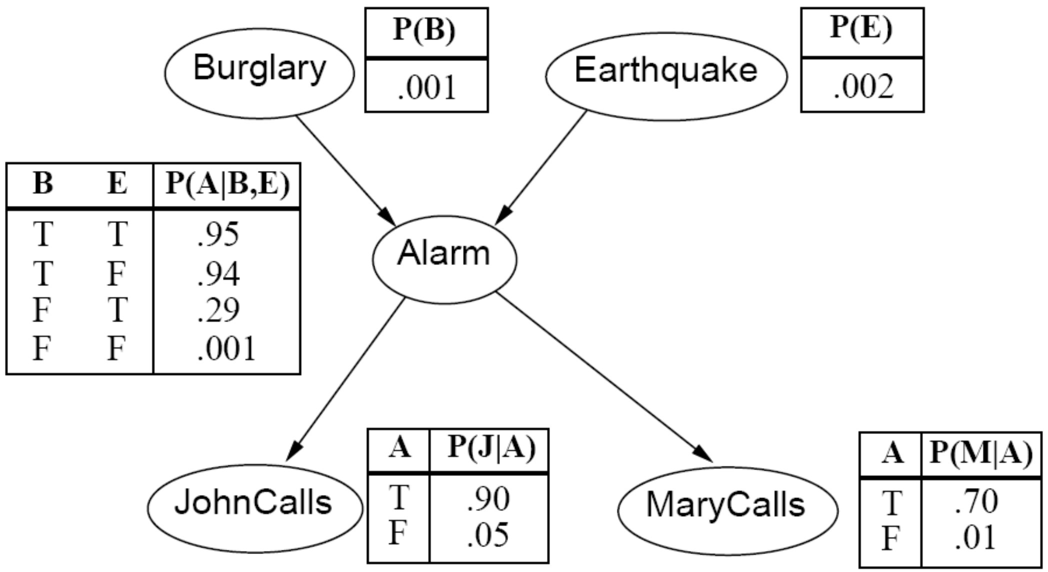
Using a Bayes network to estimate a posteriori probabilities



- Third step: ignore (delete) the column that didn't happen.

$P(B,M)$	$\boldsymbol{M}$
$\neg B$	0.011078
B	0.000659

Using a Bayes network to estimate a posteriori probabilities



- Fourth step: use the definition of conditional probability.

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

$P(B\|M)$	$\boldsymbol{M}$
$\neg B$	0.943883
B	0.056117

Some unexpected conclusions

- Burglary is so unlikely that, if only Mary calls or only John calls, the probability of a burglary is still only about 5%.
- If both Mary and John call, the probability is ~50%.

unless …

Some unexpected conclusions

- Burglary is so unlikely that, if only Mary calls or only John calls, the probability of a burglary is still only about 5%.
- If both Mary and John call, the probability is ~50%.

unless …

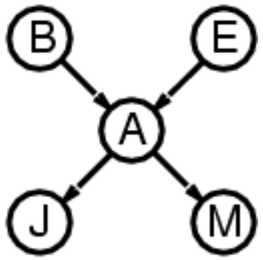
- If you know that there was an earthquake, then the probability is, the alarm was caused by the earthquake. In that case, the probability you had a burglary is vanishingly small, even if twenty of your neighbors call you.
- This is called the “explaining away” effect. The earthquake “explains away” the burglar alarm.

Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Inference in a Bayes network
- Conditional independence ≠ Independence

The joint probability distribution

$$\boxed{P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\big(X_i \mid Parents(X_i)\big)}$$

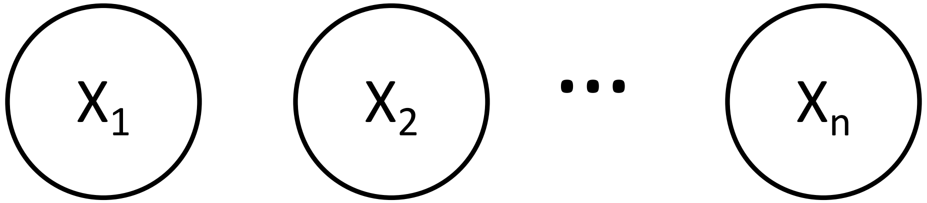


For example,

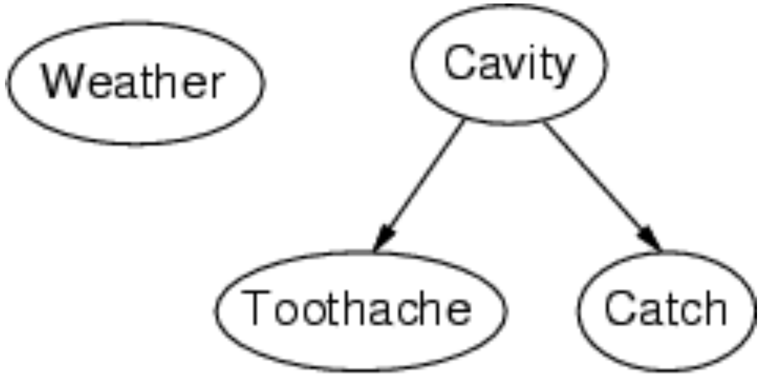
$P(j, m, a, \neg b, \neg e) = P(\neg b)\, P(\neg e)\, P(a \mid \neg b, \neg e)\, P(j \mid a)\, P(m \mid a)$

Independence

- By saying that X_i and X_j are independent, we mean that
$$\mathrm{P}(X_j, X_i) = \mathrm{P}(X_i)\mathrm{P}(X_j)$$
- X_i and X_j are independent if and only if they have no common ancestors
- Example: *independent coin flips*

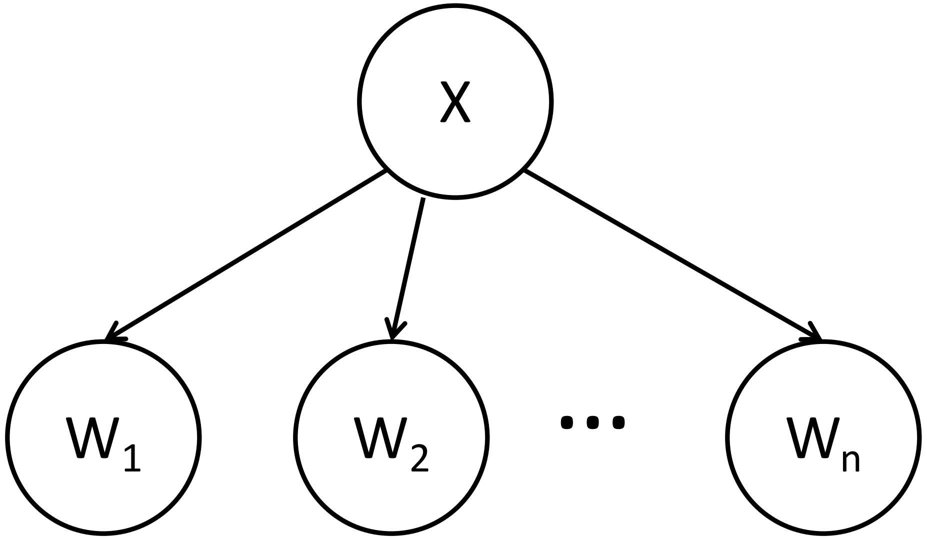


- Another example: Weather is independent of all other variables in this model.



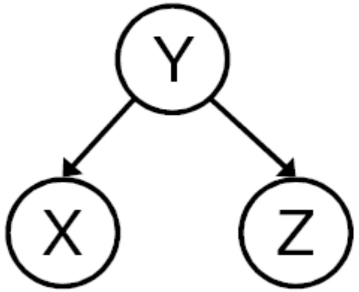
Conditional independence

- By saying that W_i and W_j are conditionally independent given X, we mean that
$$\mathrm{P}\left(W_i, W_j \middle| X\right) = \mathrm{P}(W_i|X)\mathrm{P}(W_j|X)$$
- W_i and W_j are conditionally independent given X if and only if they have no common ancestors other than the ancestors of X.
- Example: *naïve Bayes model:*



Conditional independence ≠ Independence

Common cause: Conditionally Independent



Y: Project due

X: Newsgroup busy

Z: Lab full

Are X and Z independent? **No**

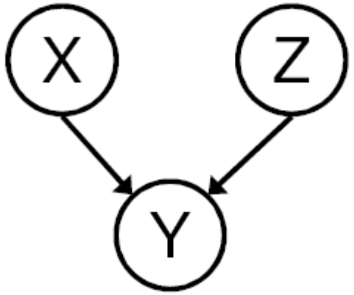
$$P(Z,X) = \sum_Y P(Z|Y)P(X|Y)P(Y)$$

$$P(Z)P(X) = \left(\sum_Y P(Z|Y)P(Y)\right)\left(\sum_Y P(X|Y)P(Y)\right)$$

Are they conditionally independent given Y? **Yes**

$$P(Z,X|Y) = P(Z|Y)P(X|Y)$$

Common effect: Independent



X: Raining

Z: Ballgame

Y: Traffic

Are X and Z independent? **Yes**

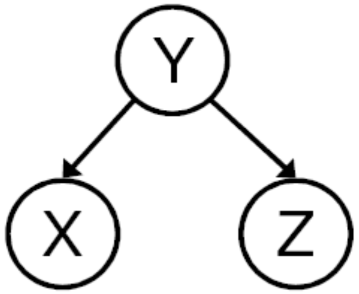
$$P(X,Z) = P(X)P(Z)$$

Are they conditionally independent given Y? **No**

$$P(Z,X|Y) = \frac{P(Y|X,Z)P(X)P(Z)}{P(Y)} \neq P(Z|Y)P(X|Y)$$

Conditional independence ≠ Independence

Common cause: Conditionally Independent



Y: Project due

X: Newsgroup busy

Z: Lab full

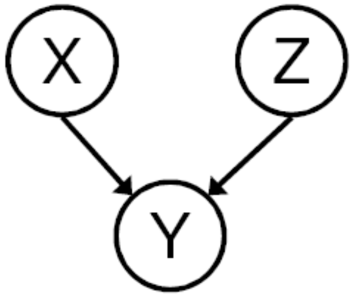
Are X and Z independent? **No**

Knowing X tells you about Y, which tells you about Z.

Are they conditionally independent given Y? **Yes**

If you already know Y, then X gives you no useful information about Z.

Common effect: Independent



X: Raining

Z: Ballgame

Y: Traffic

Are X and Z independent? **Yes**

Knowing X tells you nothing about Z.

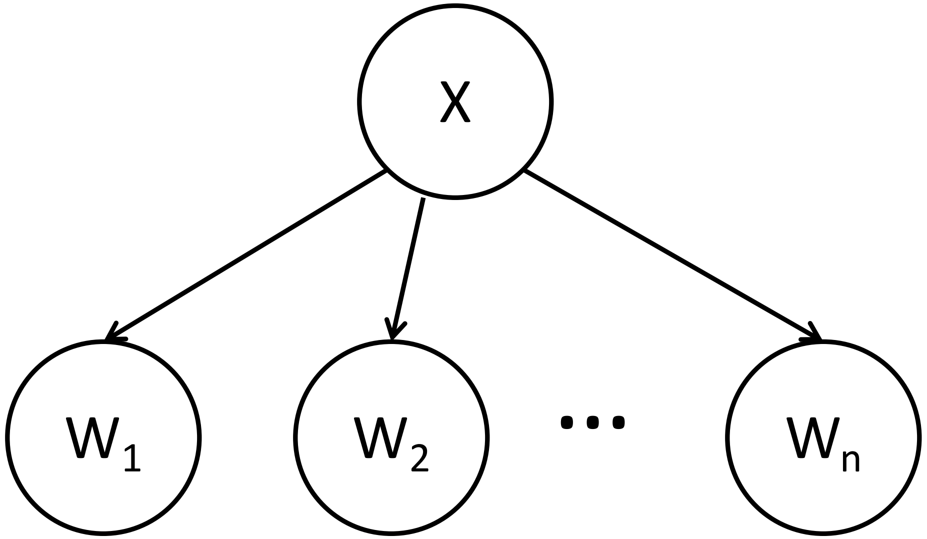
Are they conditionally independent given Y? **No**

If Y is true, then either X or Z must be true.

Knowing that X is false means Z must be true.

We say that X “explains away” Z.

Conditional independence ≠ Independence

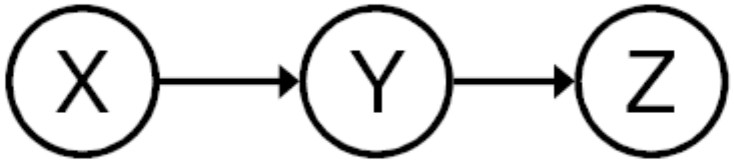


Being conditionally independent given X does NOT mean that W_i and W_j are independent. Quite the opposite. For example:

- The document topic, X, can be either "sports" or "pets", equally probable.
- W_1=1 if the document contains the word "food," otherwise W_1=0.
- W_2=1 if the document contains the word "dog," otherwise W_2=0.
- Suppose you don't know X, but you know that W_2=1 (the document has the word "dog"). Does that change your estimate of p(W_1=1)?

Conditional independence

Another example: *causal chain*



X: Low pressure

Y: Rain

Z: Traffic

- X and Z are conditionally independent given Y, because they have no common ancestors other than the ancestors of Y.

- Being conditionally independent given Y does NOT mean that X and Z are independent. Quite the opposite. For example, suppose $\mathrm{P}(X) = 0.5$, $\mathrm{P}(Y|X) = 0.8$, $\mathrm{P}(Y|\neg X) = 0.1$, $\mathrm{P}(Z|Y) = 0.7$, and $\mathrm{P}(Z|\neg Y) = 0.4$. Then we can calculate that $\mathrm{P}(Z|X) = 0.64$, but $\mathrm{P}(Z) = 0.535$

Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Inference in a Bayes network
- Conditional independence ≠ Independence