# Adapted from AIMA slides

## Bayesian networks

Peter Antal
antal@mit.bme.hu

# Outline

- Reminder: inference in the joint distribution
- Reminder: efficiency in Naïve Bayesian networks
- Properties of irrelevance
- Axiomatizatin of independencies
- Bayesian networks
- Special local models
  - Noisy-OR
  - Decision tree CPDs
  - Decision graph CPDs

# The joint probability distribution Classical vs probabilistic logic

| $P_1(X_1)$ | … | $P_k(X_k)$ | KB | $P(X_1 \ldots, X_1)$ |
|---|---|---|---|---|
| 0 | | 0 | 0 | .01 |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| 1 | | 1 | 1 | .1 |

# Inference by enumeration, contd.

Every question about a domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

# Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

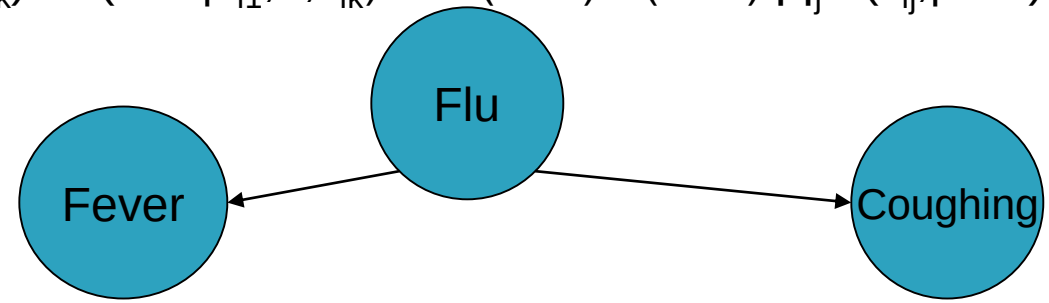
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$




$$p(Flu = present \mid Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

# Conditional independence

„Probability theory=measure theory+independence”

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z: $P(X;Y|z)=P(Y|z)\ P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z) =def= \neg\ I_P(X;Y|Z)$

Contextual independence: for not all z.

# Properties of irrelevance

(Properties of relevance: transitivity: If X is relevant for Y, and Y for Z, then X is relevant for Z.)

a Symmetry: The observational probabilistic conditional independence is symmetric.

b Decomposition: Any part of an irrelevant information is irrelevant.

c Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

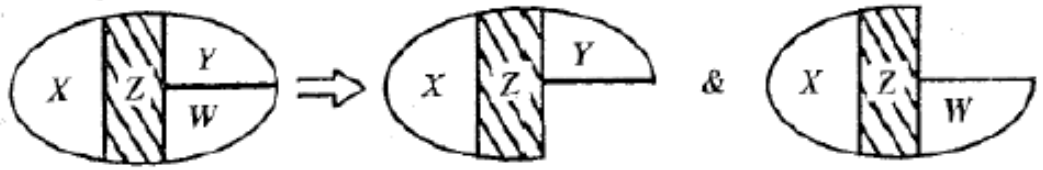
d Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

e Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies

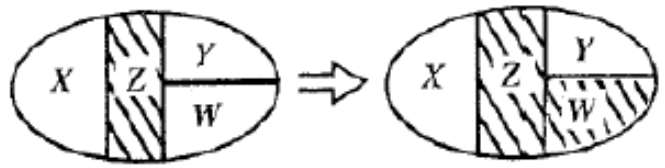
Pearl, Judea. *Probabilistic reasoning in intelligent systems: networks of plausible inference*.

# Properties of irrelevance

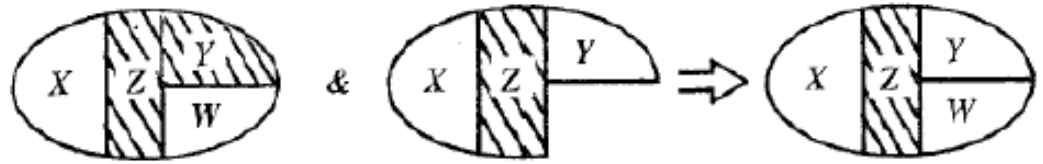
*Decomposition*



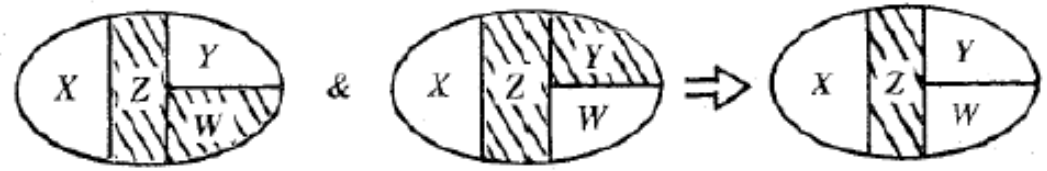

*Weak Union*




*Contraction*




*Intersection*




Pearl, Judea. *Probabilistic reasoning in intelligent systems: networks of plausible inference*.

1. Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution)
2. Graphoids: Semi-graphoids+Intersection (holds only in strictly positive distribution)

# Properties of independence

a Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ iff } I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

b Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

c Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

d Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

e Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

# Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  - 
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over $X_i$ for each combination of parent values

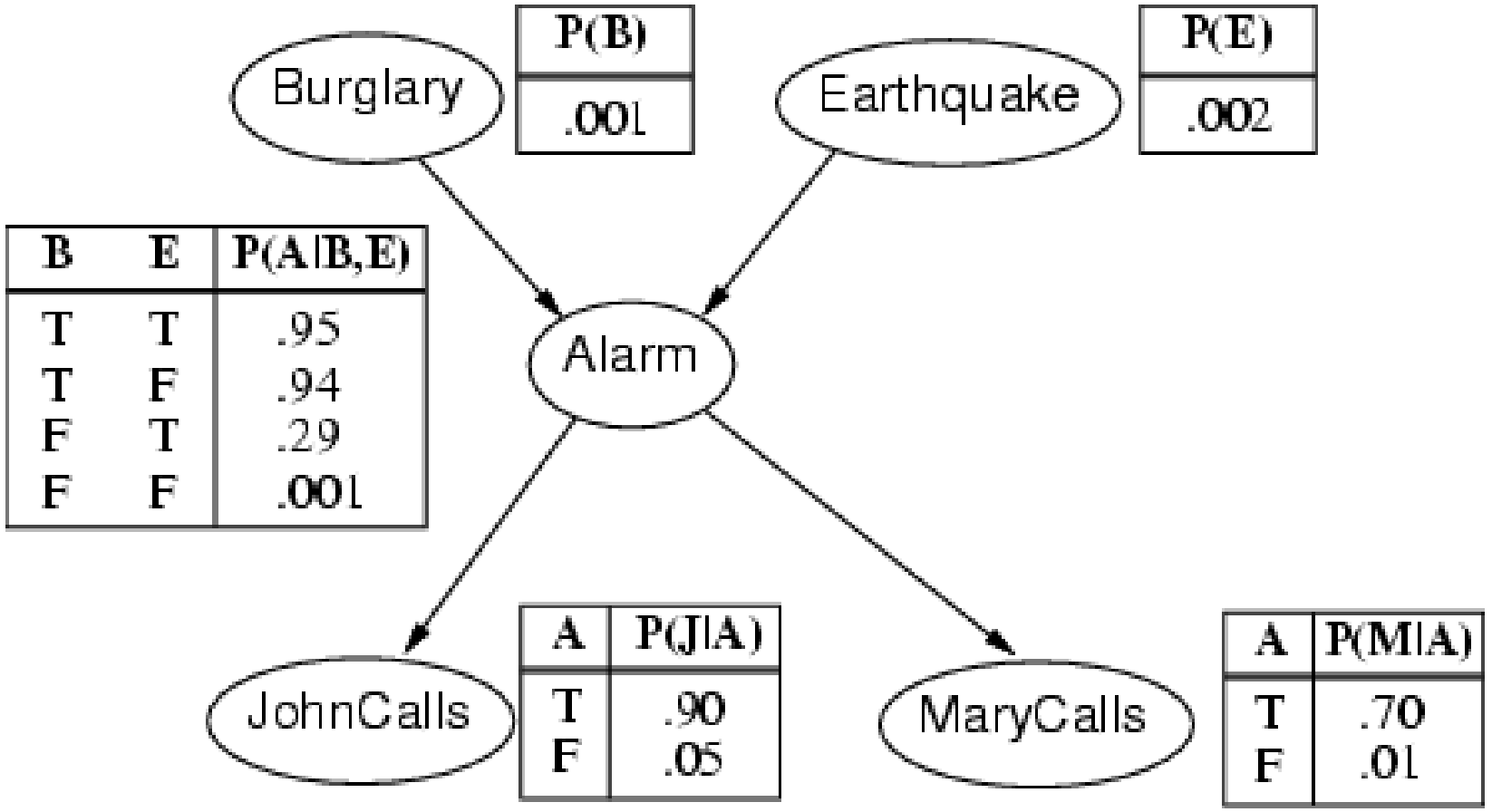
# Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

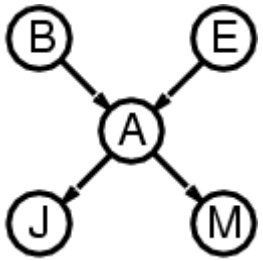
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

# Example contd.



# Compactness

- A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values

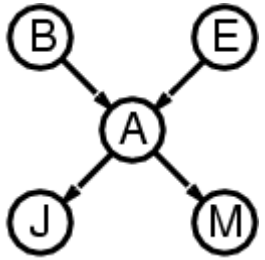



- Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

# Semantics

The full joint distribution is defined as the product of the local conditional distributions:

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




e.g., $\boldsymbol{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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

# Constructing Bayesian networks

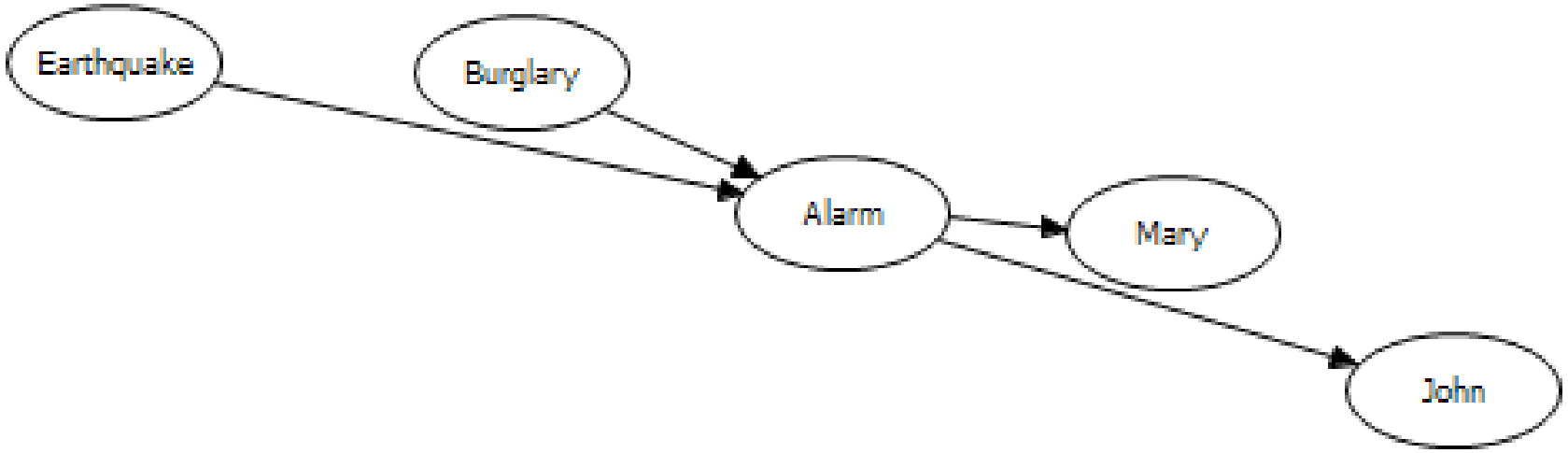
- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to $n$
  - add $X_i$ to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

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

This choice of parents guarantees:

$$\begin{aligned} \mathbf{P}(X_1, \ldots, X_n) &= \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) && //\text{(chain rule)} \\ &= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) && //\text{(by construction)} \end{aligned}$$

# (Re)constructing the example




1. Choose an ordering of variables $X_1$, … ,$X_n$
2. For $i$ = 1 to $n$
   add $X_i$ to the network
   select parents from $X_1$, … ,$X_{i-1}$ such that
   $$\boldsymbol{P}\,(X_i \mid Parents(X_i)) = \boldsymbol{P}\,(X_i \mid X_1, \ldots X_{i-1})$$

# Noisy–OR

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \dots U_k$ include all causes (can add leak node)
2) Independent failure probability $q_i$ for each cause alone

$$\Rightarrow P(X|U_1 \dots U_j, \neg U_{j+1} \dots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

| $Cold$ | $Flu$ | $Malaria$ | $P(Fever)$ | $P(\neg Fever)$ |
|---|---|---|---|---|
| F | F | F | **0.0** | 1.0 |
| F | F | T | 0.9 | **0.1** |
| F | T | F | 0.8 | **0.2** |
| F | T | T | 0.98 | $0.02 = 0.2 \times 0.1$ |
| T | F | F | 0.4 | **0.6** |
| T | F | T | 0.94 | $0.06 = 0.6 \times 0.1$ |
| T | T | F | 0.88 | $0.12 = 0.6 \times 0.2$ |
| T | T | T | 0.988 | $0.012 = 0.6 \times 0.2 \times 0.1$ |

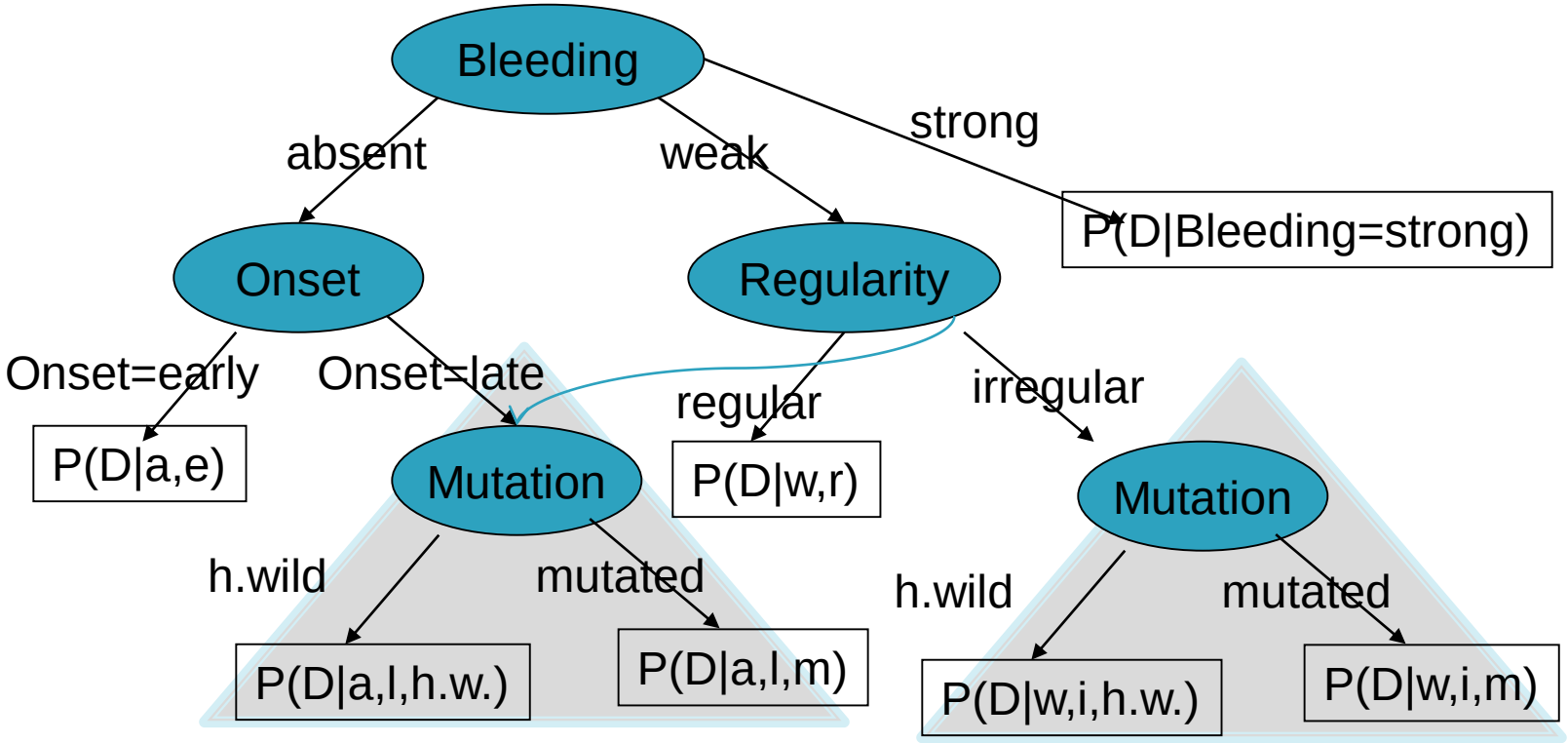
Number of parameters **linear** in number of parents

# Summary

- Conditional independencies allows:
  - efficient representation of the joint probabilitity distribution,
  - efficient inference to compute conditional probabilites.
- Bayesian networks use directed acyclic graphs to represent
  - conditional independencies,
  - conditional parameters,
  - optionally, causal mechanisms (see Knowledge engineering lecture later!).
- Design of variables and order of the variables can drastically influence structure
  - (see Knowledge engineering lecture later!)
- Canonical conditional models can further increase efficiency.

- **Suggested reading:**
  - Charniak: Bayesian networks without tears, 1991

# Decision trees, decision graphs




Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.
Decision graph: If conditions are equivalent, then subtrees can be merged.
E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)