Adapted from AIMA slides

Extended Bayesian networks

Peter Antal
antal@mit.bme.hu

Outline

- Reminder
- Bayesian network extensions
  - Canonical local models
  - Decision tree/graph local models
  - Dynamic Bayesian networks

Independence, Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z defined as follows

for all x,y and z with $P(z)>0$: $P(x;y|z)=P(x|z)\ P(y|z)$

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

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

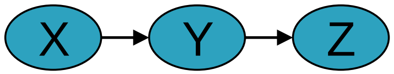
Other notations: $D_P(X;Y|Z) =\text{def}= \neg\ I_P(X;Y|Z)$

Direct dependence: $D_P(X;Y|V/\{X,Y\})$

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

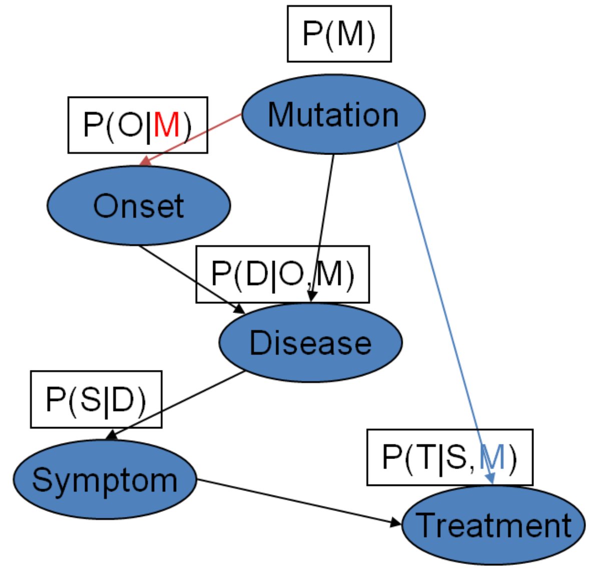
Normally/almost always: D(X;Z)

Exceptionally: I(X;Z)

Bayesian networks: three facets

3. Concise representation of joint distributions

$$P(M,O,D,S,T) =$$
$$P(M)P(O\,|\,M)P(D\,|\,O,M)P(S\,|\,D)P(T\,|\,S,M)$$



1. Causal model

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

2. Graphical representation of (in)dependencies

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:

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

- 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

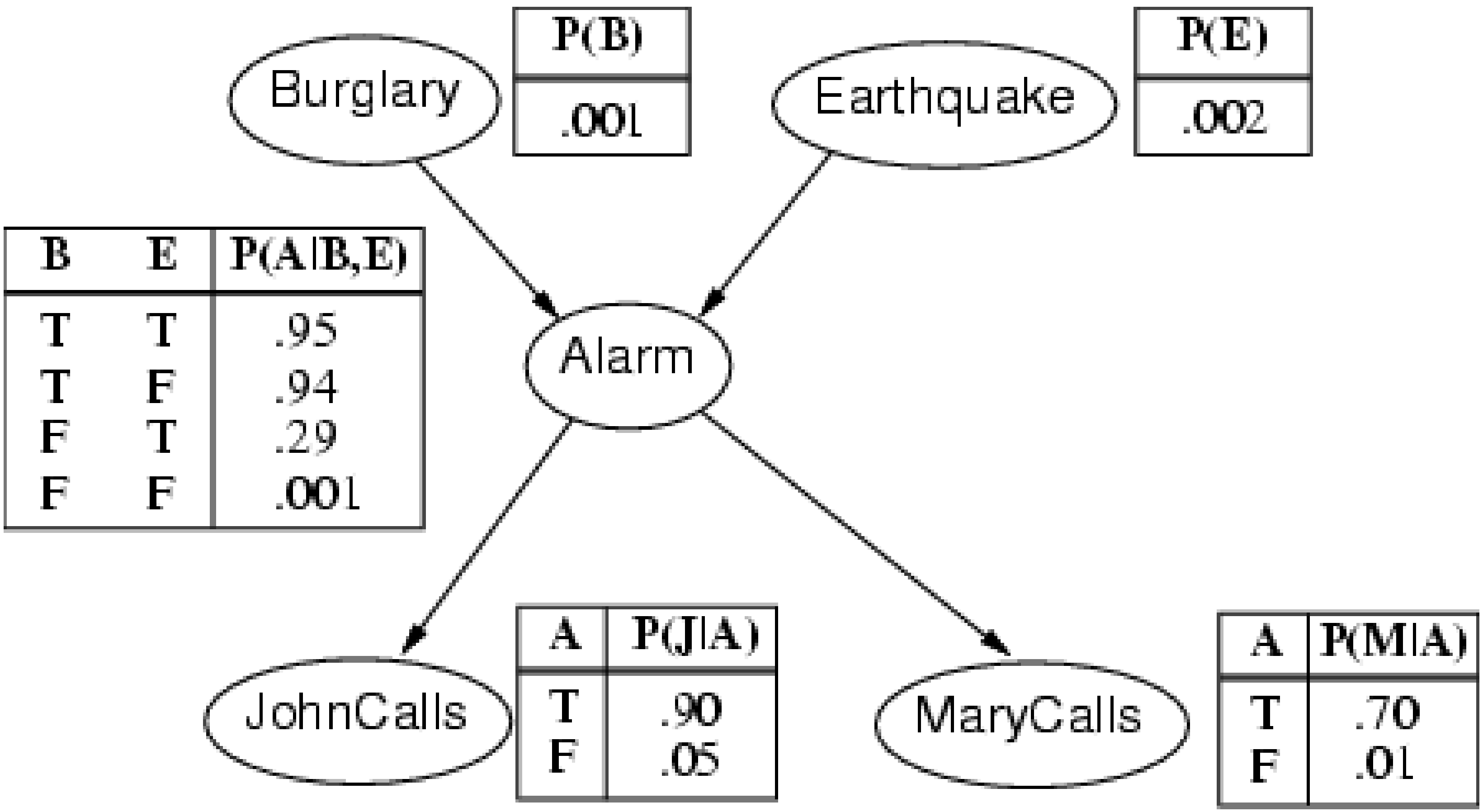
Example

- 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.



P(B)
.001

P(E)
.002

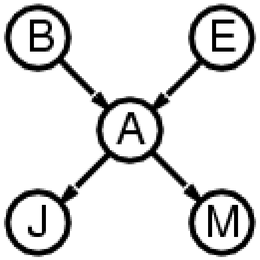
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

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

Compactness

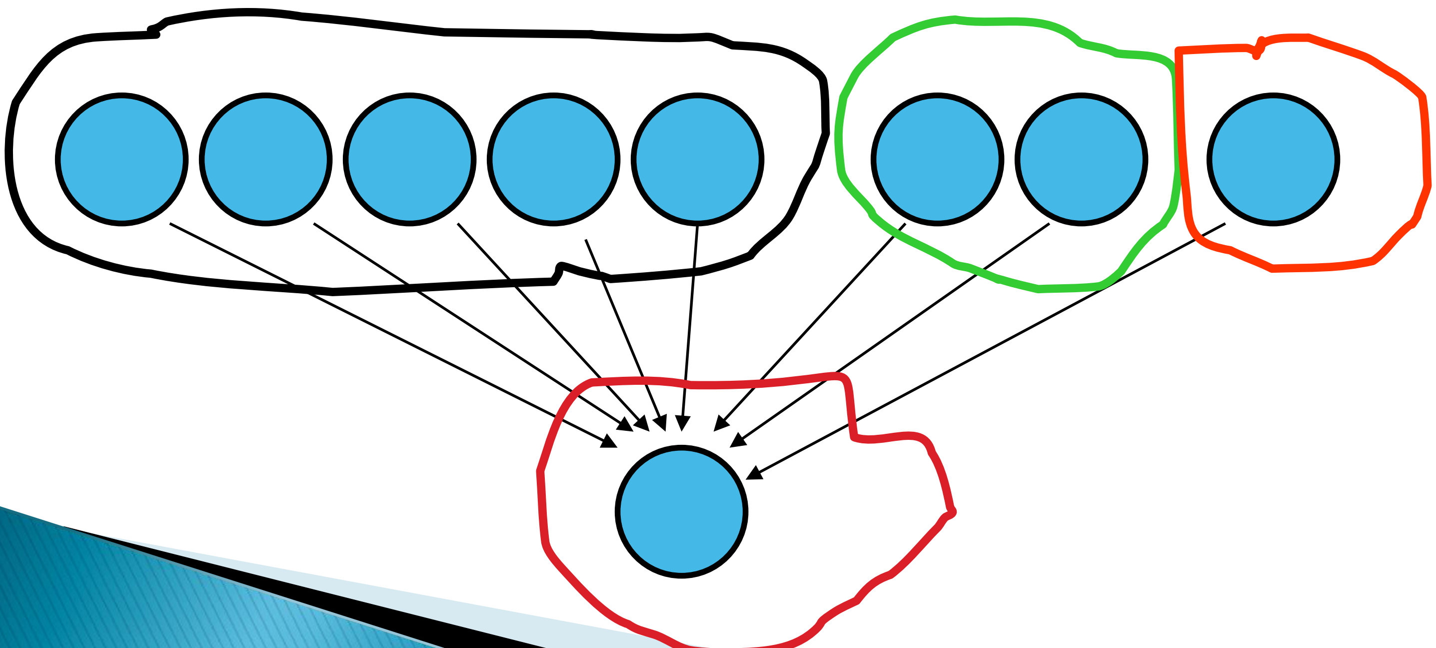
- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- 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. $2^5-1 = 31$)

A multinomiális általános eset I.

Tfh: 5 szülő csomópont bináris értékű
2 szülő csomópont 3-as értékű
1 szülő csomópont 4-es értékű és
az eredmény csomópont 5-ös értékű ?????



A multinomiális általános eset II.

Sz1	Sz2	Sz3	Sz4	Sz5	Sz6	Sz7	Sz8	Kimeneti változó				
								e1	e2	e3	e4	e5
.	.	.	.	.	.	.	.	P	P	P	P	P
.	.	.	.	.	.	.	.	P	P	P	P	P
1	1	1	1	1	.	.	.	P	P	P	P	P
0	0	0	0	0	e1	e1	.	P	P	P	P	P
.	.	.	.	.	e2	e2	.	P	P	P	P	P
.	.	.	.	.	e3	e3	.	P	P	P	P	P
.	.	.	.	.	.	.	e1	P	P	P	P	P
.	.	.	.	.	.	.	e2	P	P	P	P	P
.	.	.	.	.	.	.	e3	P	P	P	P	P
.	.	.	.	.	.	.	e4	P	P	P	P	P
.	.	.	.	.	.	.	.	P	P	P	P	P
.	.	.	.	.	.	.	.	P	P	P	P	P

Minden kombináció

2^5 x 3^2 x 4 szülői feltétel van (FVT sor) és 4 (független érték) (FVT oszlop) = összesen: (32 x 9 x 4) x 4 = 4608

együttes eloszláshoz kell: 2^5 x 3^2 x 4 x 5 – 1 = 5759

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:

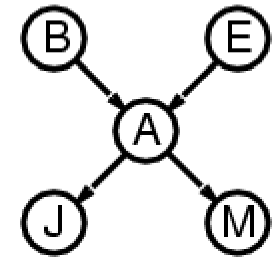
$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \qquad \text{//(chain rule)}$$

$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \qquad \text{//(by construction)}$$

Semantics

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

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



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

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

Noisy–OR

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)

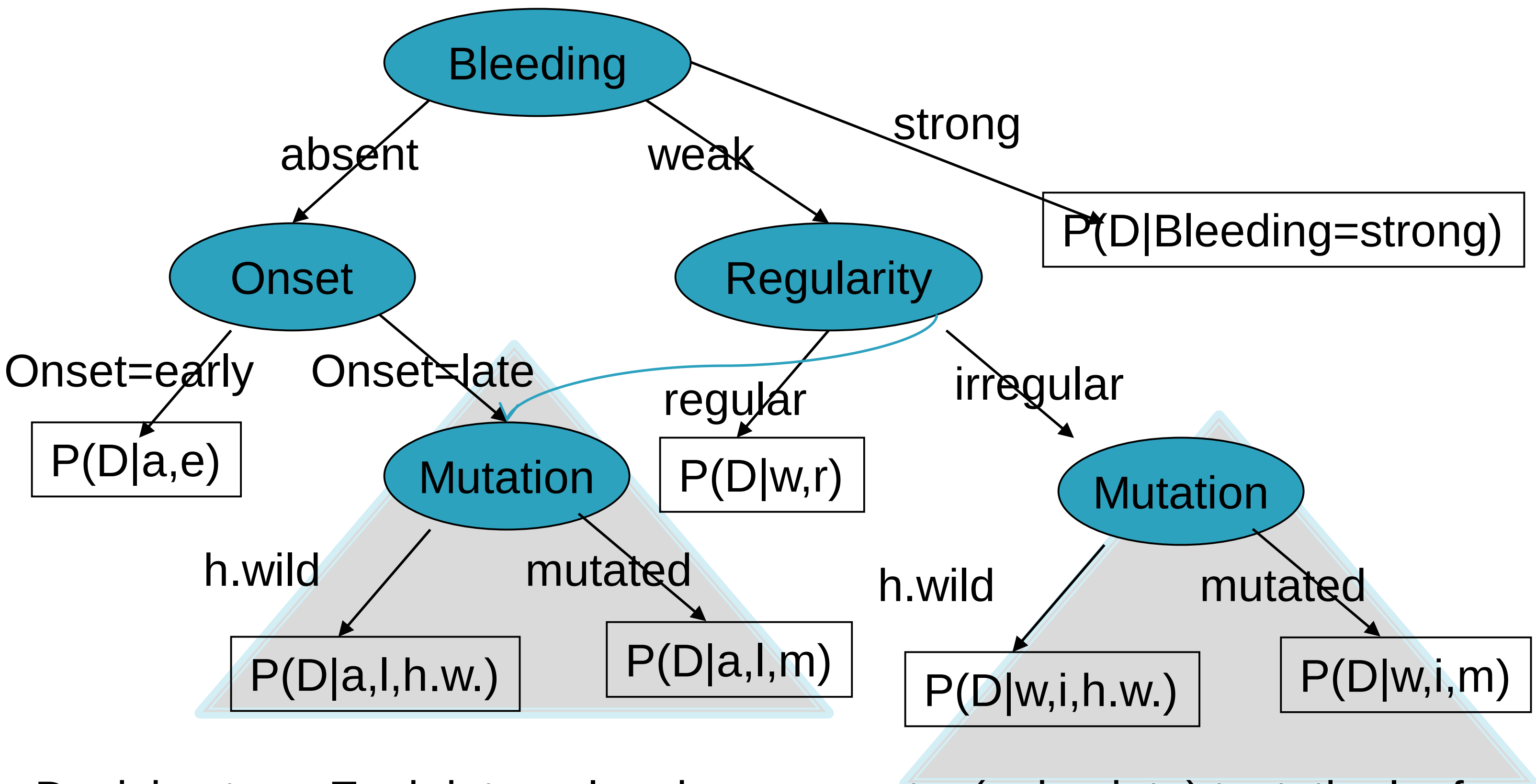
2) Independent failure probability q_i for each cause alone

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

$Cold$	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

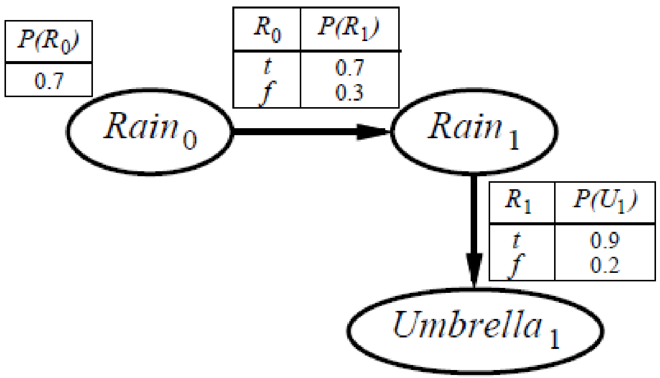
Decision trees, decision graphs

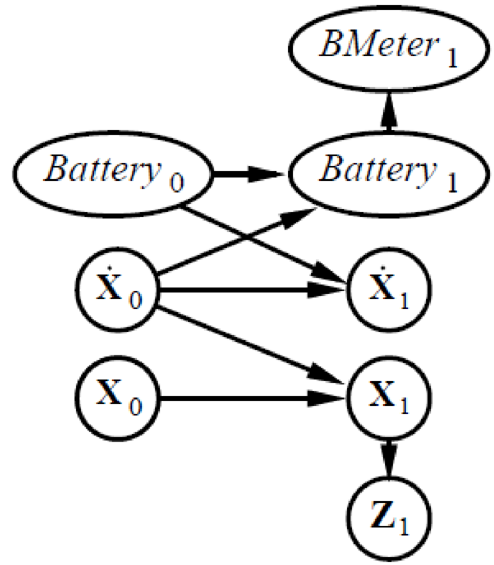


Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.
Decision graph: If conditions are equivalent, then subtrees can be merged.
E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

Dynamic Bayesian networks

$\mathbf{X}_t$, $\mathbf{E}_t$ contain arbitrarily many variables in a replicated Bayes net

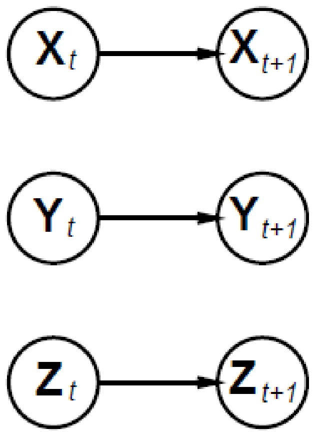


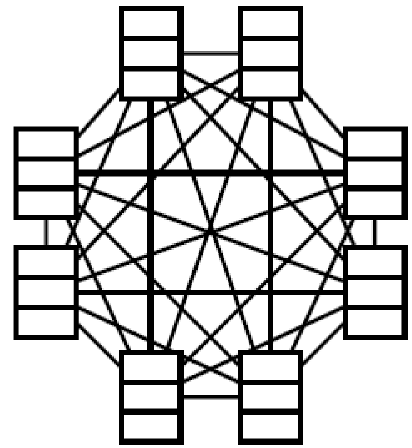


http://phoenix.mit.bme.hu:49080/kgt/

DBNs vs. HMMs

Every HMM is a single-variable DBN; every discrete DBN is an HMM





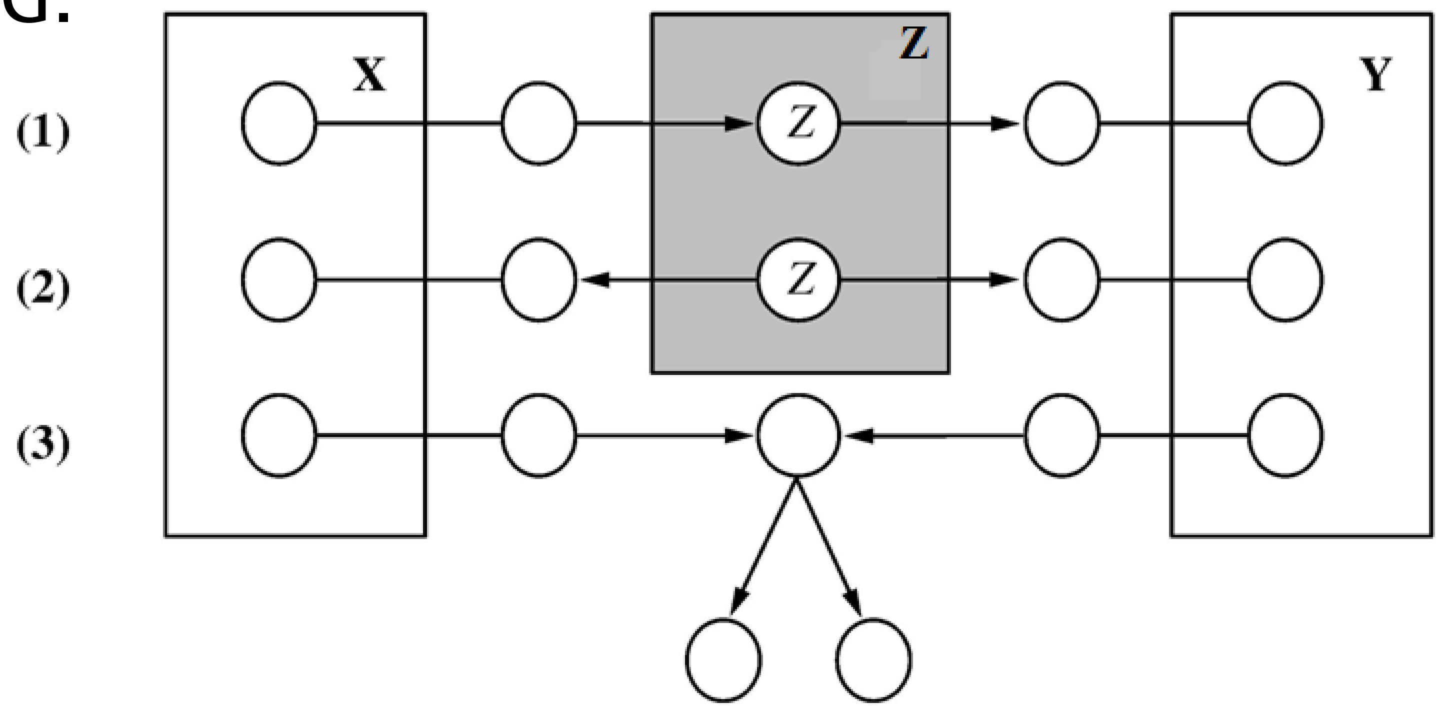
Sparse dependencies ⇒ exponentially fewer parameters;

e.g., 20 state variables, three parents each

DBN has $20 \times 2^3 = 160$ parameters, HMM has $2^{20} \times 2^{20} \approx 10^{12}$

Inferring independencies from structure: d-separation

- $I_G(X;Y|Z)$ denotes that X is d-separated (directed separated) from Y by Z in directed graph G.



d-separation and the global Markov condition

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_P, \tag{9}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

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 probability distributions,
  - causal mechanisms.
- Design of variables and order of the variables can drastically influence structure

- **Suggested reading:**
  - Charniak: Bayesian networks without tears, 1991
  - Koller, Daphne, et al. "Graphical models in a nutshell." *Introduction to statistical relational learning* (2007): 13-55.