Local Structure

COMPSCI 276, Fall 2014

Set 4b: Rina Dechter

(Reading: Darwiche chapter 5, Koller Chapter 5, Pearl chapter 4)

Outline

- Bayesian networks and queries
- Building Bayesian Networks
- Special representations of CPTs
  - Causal Independence
  - Context Specific Independence
  - Determinism
  - Mixed Networks

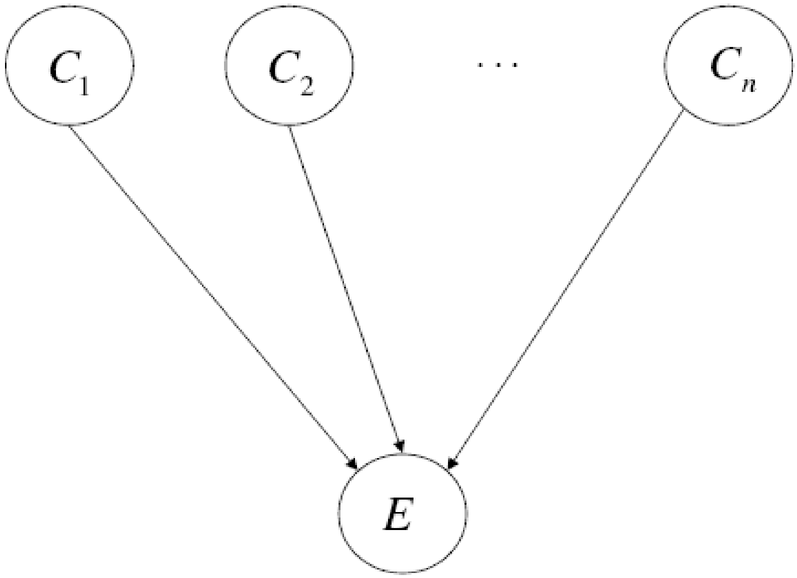
Dealing with Large CPTs

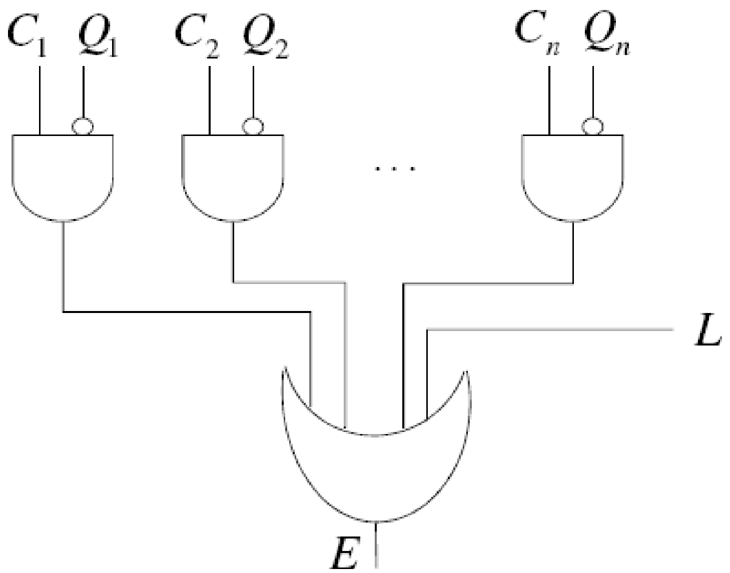
The size of a CPT

for binary variable E with binary parents $C_1, \ldots, C_n$

Number of Parents: n	Parameter Count: 2^n
2	4
3	8
6	64
10	1024
20	$1,048,576$
30	$1,073,741,824$

Micro Model





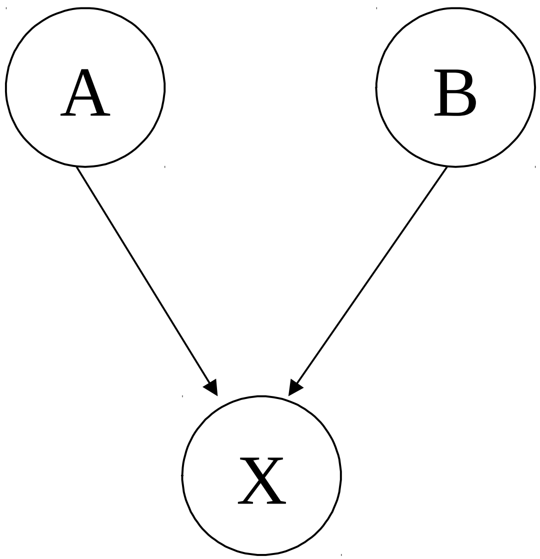
A noisy-or circuit

A micro model

details the relationship between a variable E and its parents $C_1, \ldots, C_n$.

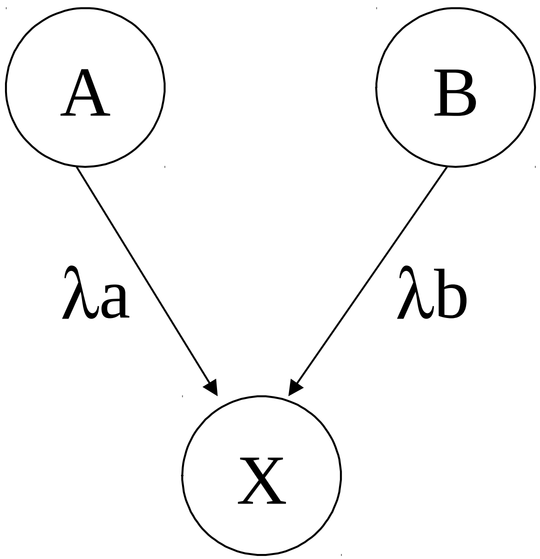
We wish to specify cpt with less parameters

Binary OR



A	B	P(X=0|A,B)	P(X=1|A,B)
0	0	1	0
0	1	0	1
1	0	0	1
1	1	0	1

Noisy-OR



"noise" is associated with each edge

described by noise parameter $\lambda \in [0,1]$:

Let q b=0.2, qa =0.1

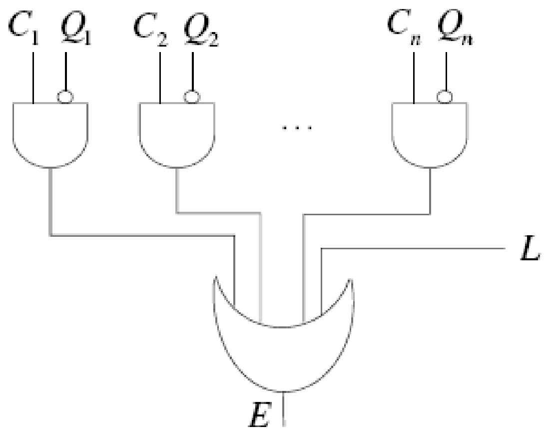
P(x=0|a,b)= (1-λa) (1-λb)

P(x=1|a,b)=1-(1-λa) (1-λb)

A	B	P(X=0|A,B)	P(X=1|A,B)
0	0	1	0
0	1	0.1	0.9
1	0	0.2	0.8
1	1	0.02	0.98

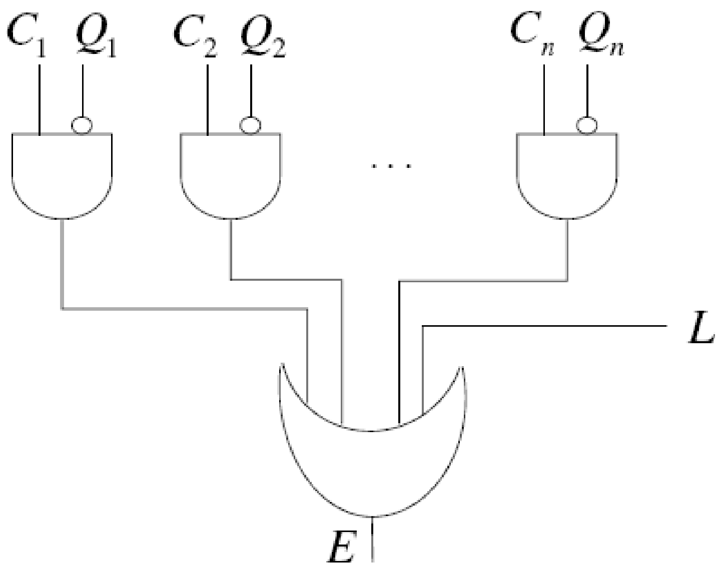
qi=P(X=0|A_i=1,…else =0)

Noisy-or Model



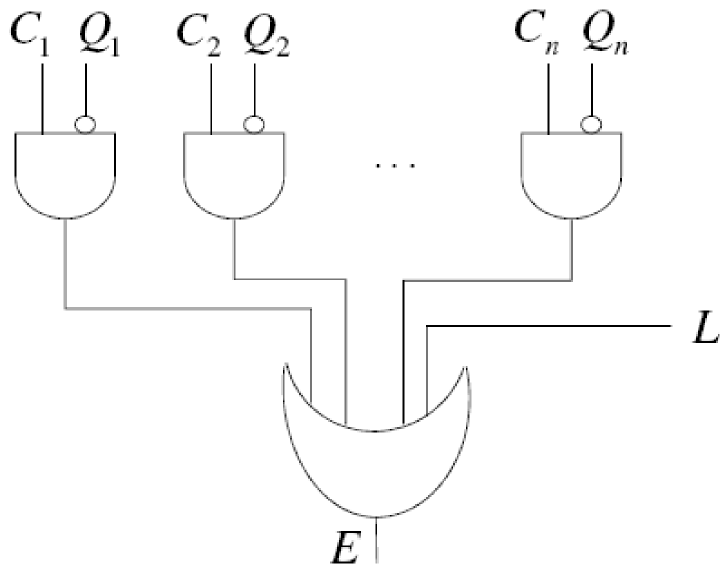
- Cause C_i is capable of establishing effect E, except under some unusual circumstances summarized by suppressor Q_i.
- When suppressor Q_i is active, C_i is no longer able to establish E.
- The leak variable L represents all other causes of E which were not modeled explicitly.
- When none of the causes C_i are active, the effect E may still be established by the leak variable L.

Noisy–or Model



The noisy-or model requires $n + 1$ parameters.

Noisy–or Model



The noisy-or model requires $n+1$ parameters.

To model the relationship between headache and ten different conditions

- $\theta_{q_i} = \Pr(Q_i = \text{active})$: probability that suppressor of C_i is active.
- $\theta_l = \Pr(L = \text{active})$: probability that leak is active.

Noisy-or Model

- Let I_α be the indices of causes that are active in α.

Noisy-or Model

- Let I_α be the indices of causes that are active in α.
- If

  α: $C_1 =$ active, $C_2 =$ active, $C_3 =$ passive, $C_4 =$ passive, $C_5 =$ active,

  then $I_\alpha = \{1, 2, 5\}$.

Noisy-or Model

- Let I_α be the indices of causes that are active in α.
- If

  α: $C_1 =$ active, $C_2 =$ active, $C_3 =$ passive, $C_4 =$ passive, $C_5 =$ active,

  then $I_\alpha = \{1, 2, 5\}$.
- We then have

$$
\begin{aligned}
\Pr(E = \text{passive} | \alpha) &= (1 - \theta_l) \prod_{i \in I_\alpha} \theta_{q_i} \\
\Pr(E = \text{active} | \alpha) &= 1 - \Pr(E = \text{passive} | \alpha).
\end{aligned}
$$

Noisy-or Model

- Let I_α be the indices of causes that are active in α.
- If

  α: C_1 = active, C_2 = active, C_3 = passive, C_4 = passive, C_5 = active,

  then $I_\alpha = \{1, 2, 5\}$.
- We then have

$$\begin{aligned}
\Pr(E = \text{passive}|\alpha) &= (1 - \theta_l) \prod_{i \in I_\alpha} \theta_{q_i} \\
\Pr(E = \text{active}|\alpha) &= 1 - \Pr(E = \text{passive}|\alpha).
\end{aligned}$$

The full CPT for variable E, with its 2^n parameters, can be induced from the $n + 1$ parameters of the noisy-or model.

Noisy-or Model

Example

Sore throat (S) has three causes: cold (C), flu (F), tonsillitis (T).

Noisy-or Model

Example

Sore throat (S) has three causes: cold (C), flu (F), tonsillitis (T).

If we assume that S is related to its causes by a noisy-or model

we can then specify the CPT for S by the following four probabilities:

- The suppressor probability for cold, say .15
- The suppressor probability for flu, say, .01
- The suppressor probability for tonsillitis, say .05
- The leak probability, say .02

Noisy-or Model

Example

Sore throat (S) has three causes: cold (C), flu (F), tonsillitis (T).

Noisy-or Model

Example

Sore throat (S) has three causes: cold (C), flu (F), tonsillitis (T).

The CPT for sore throat is then determined completely as follows:

C	F	T	S	$\theta_{s\|c,f,t}$	
true	true	true	true	0.9999265	$1-(1-.02)(.15)(.01)(.05)$
true	true	false	true	0.99853	$1-(1-.02)(.15)(.01)$
true	false	true	true	0.99265	$1-(1-.02)(.15)(.05)$
⋮	⋮	⋮	⋮	⋮	
false	false	false	true	.02	$1-(1-.02)$

Noisy/OR CPDs

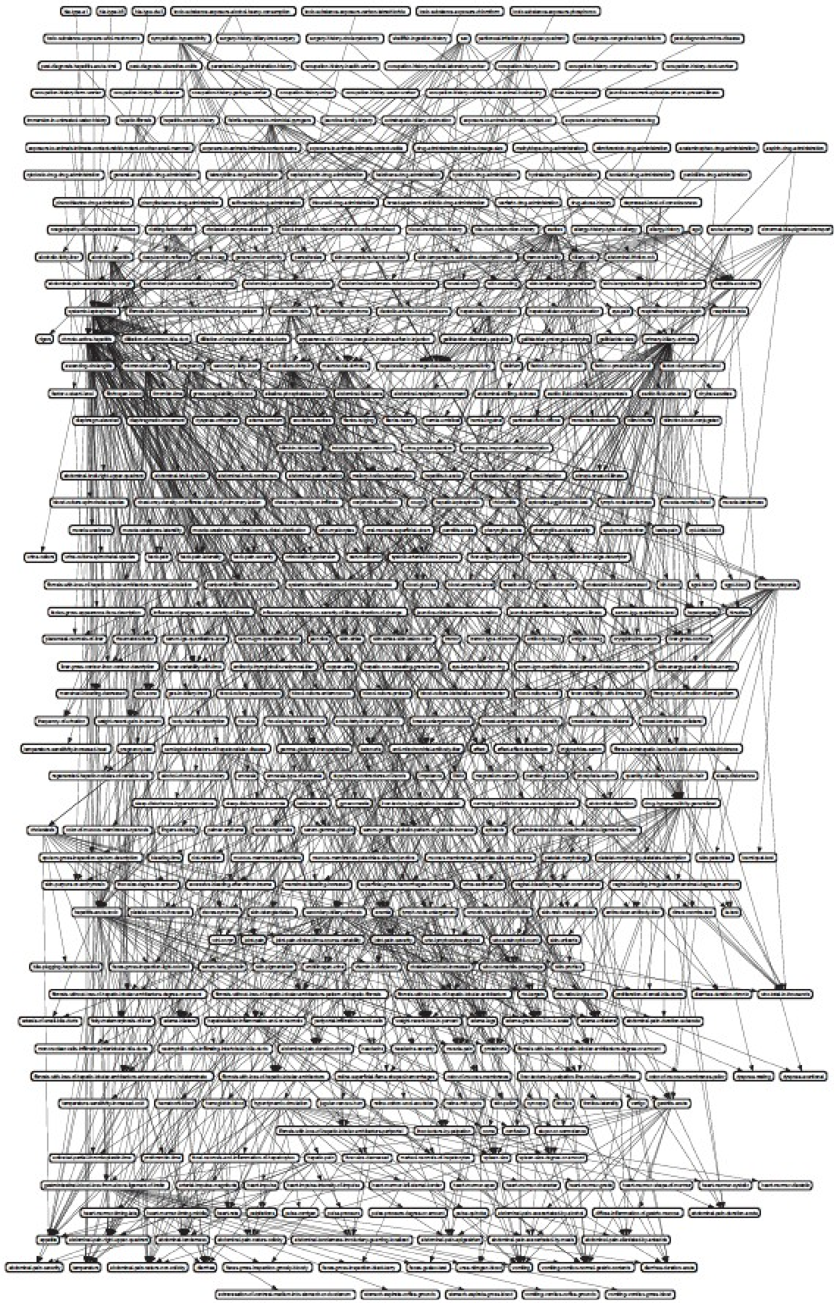


Figure 11: the CPCS network for diagnosis of internal diseases. The network contains 448 nodes, 906 links.

Independence of Causal Influence

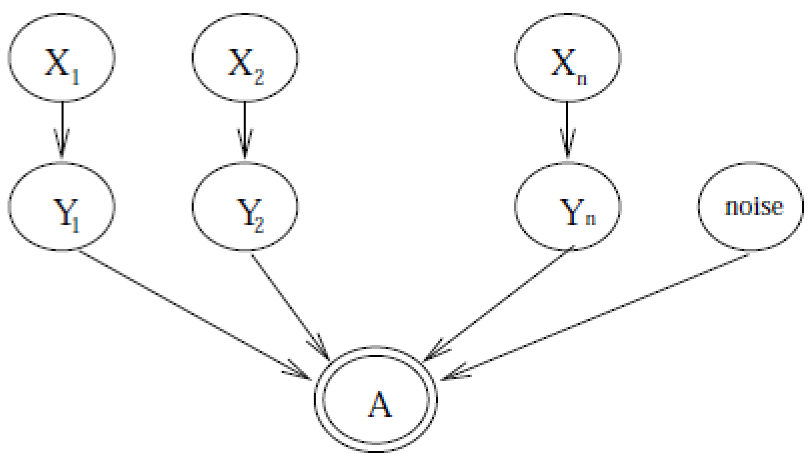


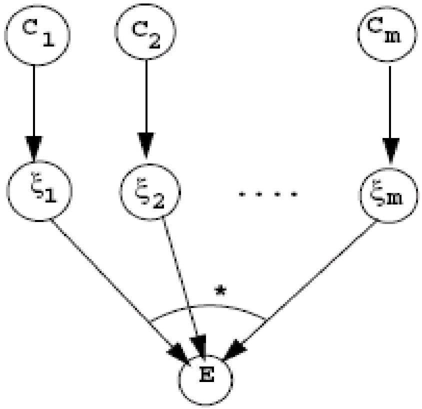
Figure 10: Independence of causal influence

Definition 2

Let A be a random variable with k parents X1,…,Xk.

The CPT P(Y|X1,…Xk) exhibits ***independence of causal influence*** (ICI) if it is described via a network fragment of the structure shown in on the left where CPT of Z is a deterministic functions f.

Causal Independence



- Formally, C_1, C_2 ..., and C_m are said to be **causally independent** w.r.t effect E if
  - there exist random variables ξ_1, ξ_2 ..., and ξ_m such that
    1. For each i, ξ_i probabilistically depends on C_i and is conditionally independent of all other C_j's and all other ξ_j's given C_i, and
    2. There exists a commutative and associative binary operator $*$ over the domain of e such that

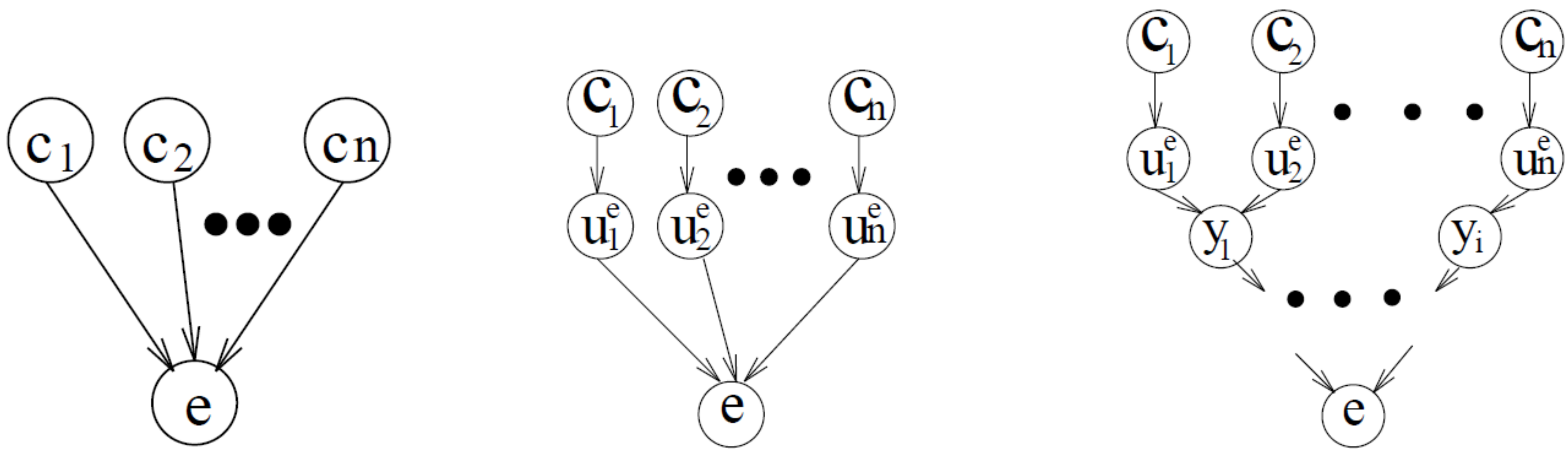
$$E = \xi_1 * \xi_2 * \ldots * \xi_m$$

.

Causal Independence

- Example: Lottery
  - C_i: money spent on buying lottery of type i.
  - E: change of wealth.
  - ξ_i: change in wealth due to buying the ith type lottery.
  - Base combination operator: "+". (**Noisy-adder**)

- Other causal independence models:
  1. Noisy MAX-gate — *max*
  2. Noisy AND-gate — $\wedge$



(a) (b) (c)

Figure 2: (a) causally-independent belief network; (b) its dependency graph and (c) its decomposition network

We can use a linear number of hidden variables to decompose

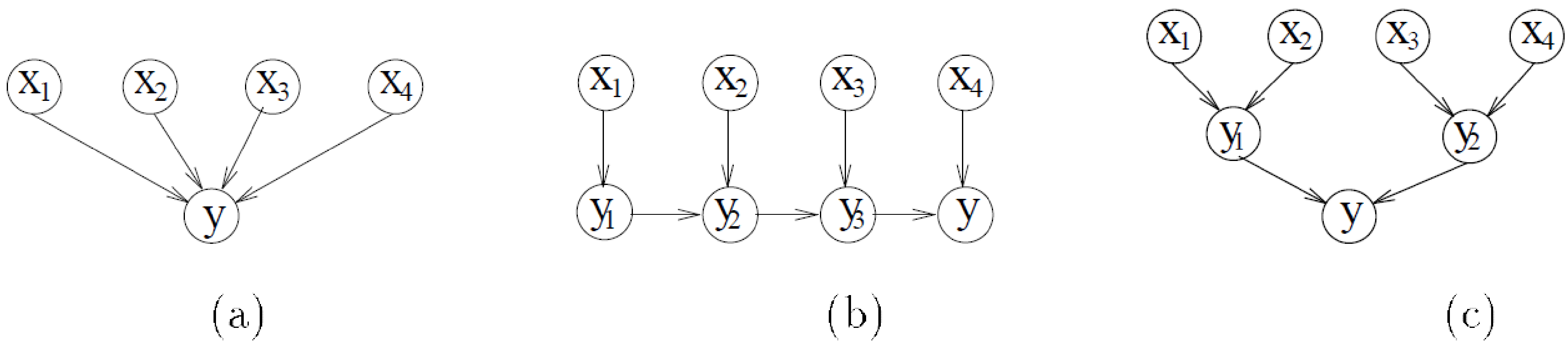


Figure 3: (a) A Bayesian network; (b) temporal transformation (c) parent divorcing.

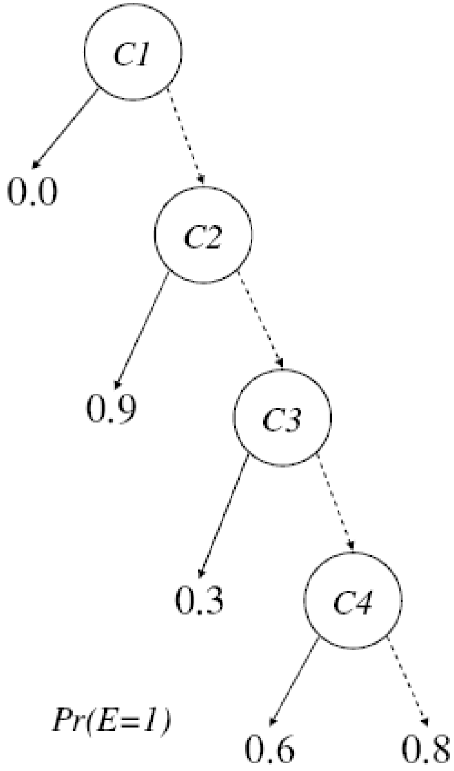
Context Specific Independence

- When there is conditional independence in some specific variable assignment
- Reading:
  - Darwiche chapter 5
  - *Koller & Freidman Chapter 5*
  - Pearl Chapter 4

Decision Trees

Can we use hidden variables?

C1	C2	C3	C4	Pr(E=1)
1	1	1	1	0.0
1	1	1	0	0.0
1	1	0	1	0.0
1	1	0	0	0.0
1	0	1	1	0.0
1	0	1	0	0.0
1	0	0	1	0.0
1	0	0	0	0.0
0	1	1	1	0.9
0	1	1	0	0.9
0	1	0	1	0.9
0	1	0	0	0.9
0	0	1	1	0.3
0	0	1	0	0.3
0	0	0	1	0.6
0	0	0	0	0.8



If-Then Rules

A CPT for variable E can be represented using a set of if-then rules of the form

If α_i then $\Pr(e) = p_i$, for each value e of variable E, where α_i is a propositional sentence constructed using the parents of variable E.

If-Then Rules

A CPT for variable E can be represented using a set of if-then rules of the form

If α_i then $\Pr(e) = p_i$, for each value e of variable E, where α_i is a propositional sentence constructed using the parents of variable E.

If $C_1 = 1$	then	$\Pr(E = 1) = 0.0$
If $C_1 = 0 \wedge C_2 = 1$	then	$\Pr(E = 1) = 0.9$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 1$	then	$\Pr(E = 1) = 0.3$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 0 \wedge C_4 = 1$	then	$\Pr(E = 1) = 0.6$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 0 \wedge C_4 = 0$	then	$\Pr(E = 1) = 0.8$

If-Then Rules

A CPT for variable E can be represented using a set of if-then rules of the form

If α_i then $\Pr(e) = p_i$, for each value e of variable E, where α_i is a propositional sentence constructed using the parents of variable E.

For the rule-based representation to be complete and consistent

- The premises α_i must be mutually exclusive. That is, $\alpha_i \wedge \alpha_j$ is inconsistent for $i \neq j$. This ensures that the rules will not conflict with each other.
- The premises α_i must be exhaustive. That is, $\bigvee_i \alpha_i$ must be valid. This ensures that every CPT parameter $\theta_{e|\ldots}$ is implied by the rules.

Context specific independence (CSI)

- Let **C** be a set of variables. A **context** on **C** is an assignment of one value to each variable in **C**.
- We denote a context by **C**=**c**, where **c** is a set of values of variables in **C**.
- Two contexts are **incompatible** if there exists a variable that is assigned different values in the contexts.
- They are **compatible** otherwise.

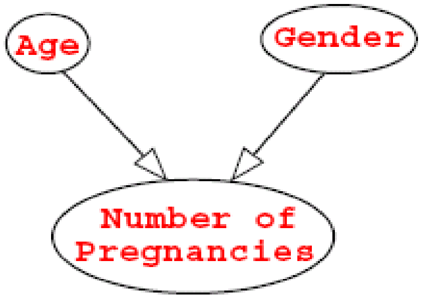
Context-specific independence

- Let **X**, **Y**, **Z**, and **C** be four disjoint sets of variables.
- **X** and **Y** are **independent given Z in context C=c** if

$$P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \mathbf{C}{=}\mathbf{c}) = P(\mathbf{X}|\mathbf{Z}, \mathbf{C}{=}\mathbf{c})$$

  whenever $P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}{=}\mathbf{c}) > 0$.
- When **Z** is empty, one simply says that **X** and **Y** are **independent in context C=c**.

Context-specific independence

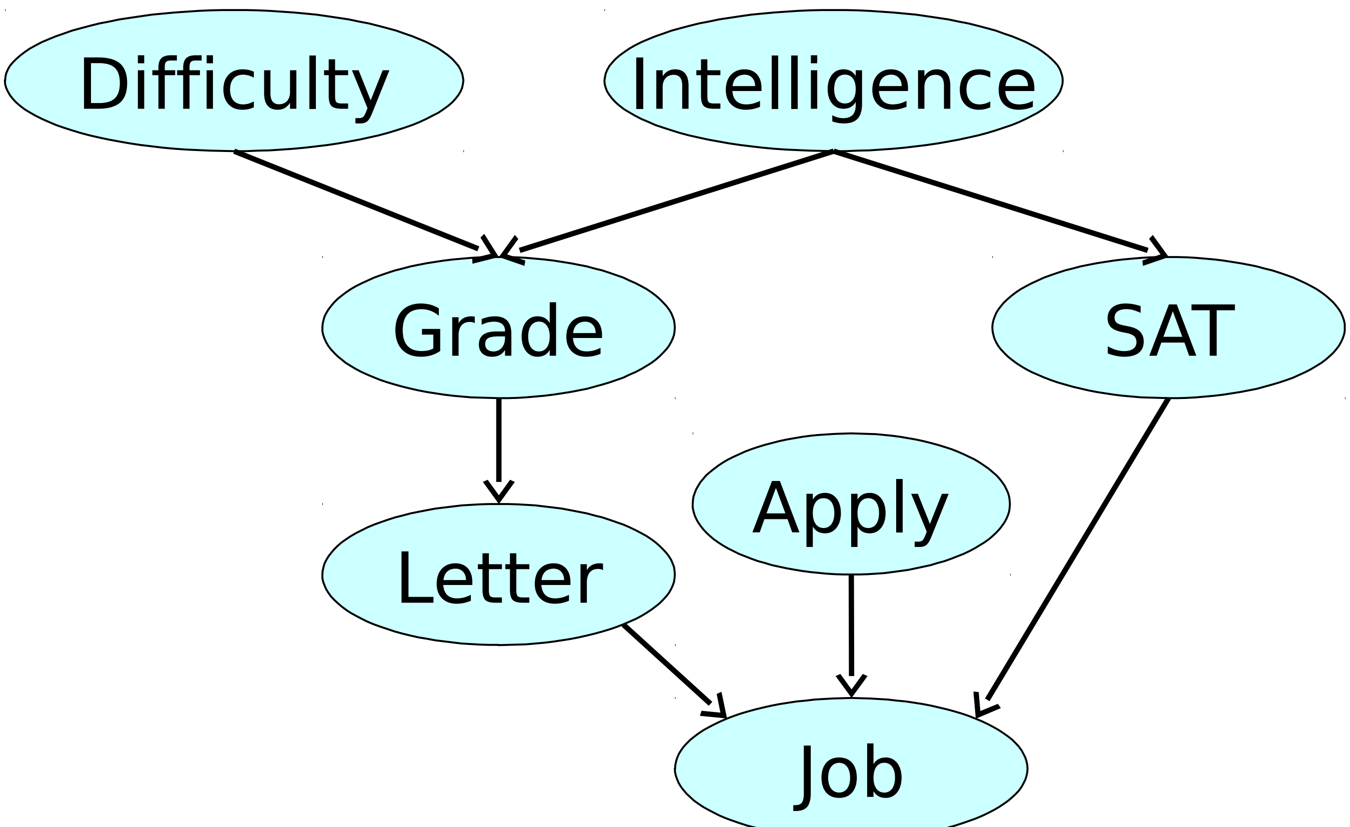


- Shafer's Example:
  - *Number of pregnancies (N)* is independent of *Age (A)* in the context *Gender=Male (G=m)*.

$$P(N|A, G{=}m) = P(N|G{=}m)$$

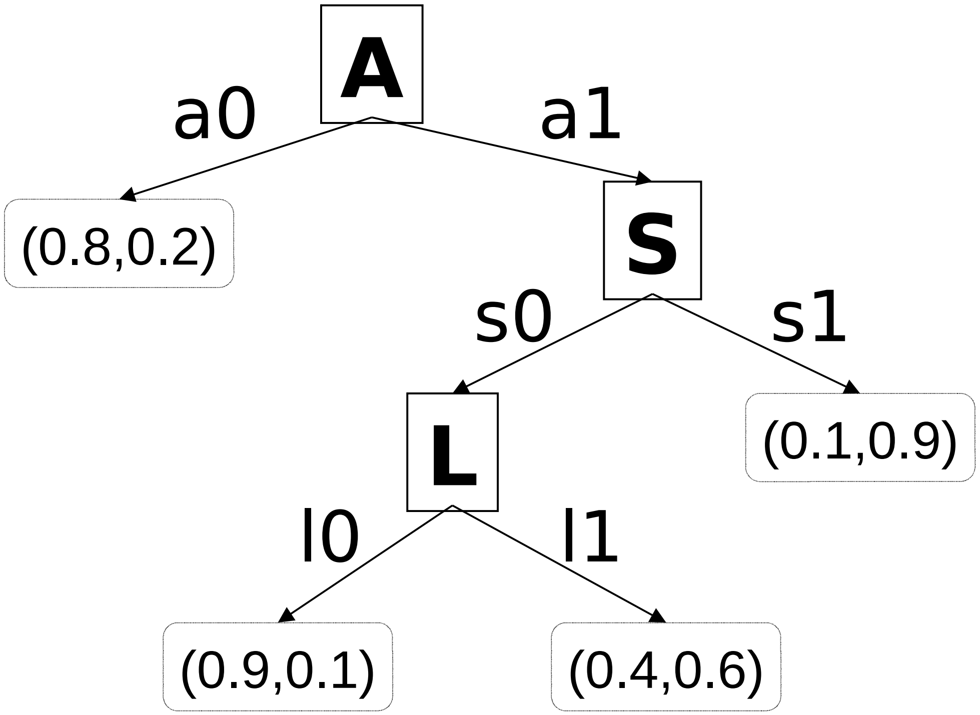
- Number of parameters reduced by $(|A|-1)(|N|-1)$.

A student's example



Tree CPD

If the student does not *A*pply, *S*AT and *L* are irrelevant

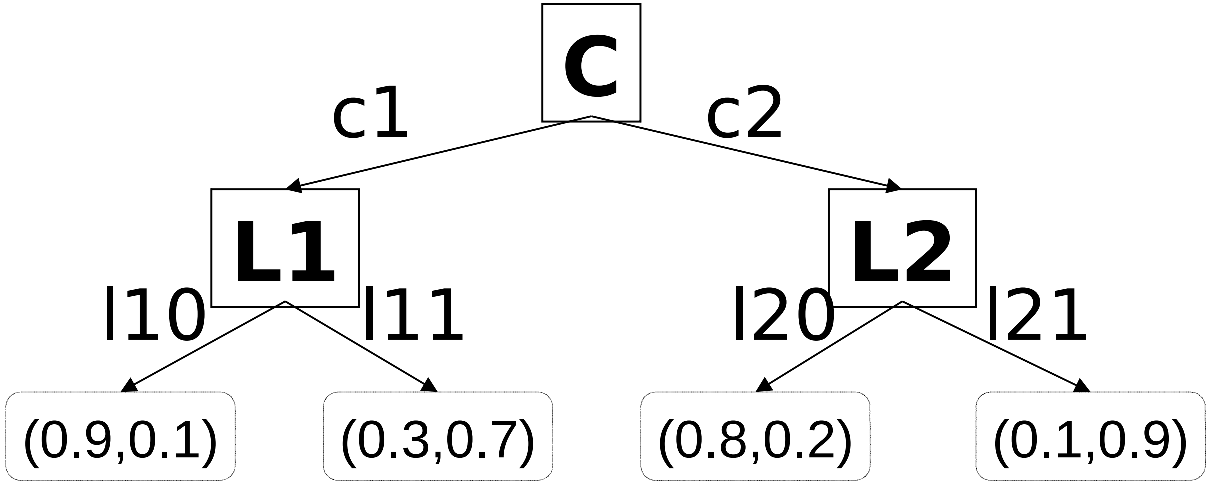


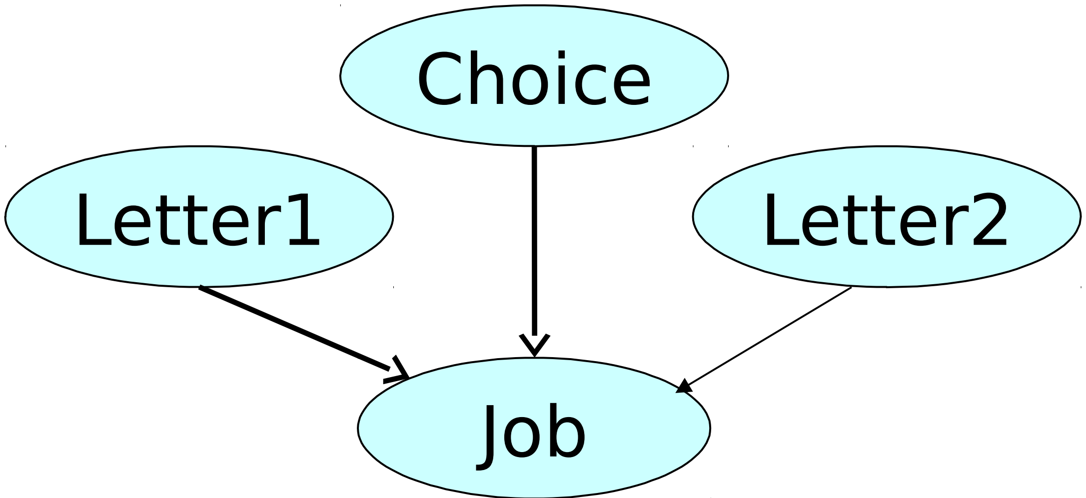
Tree-CPD for job

Definition of CPD-tree

- A CPD-tree of P(z|pa_zpa) is a tree whose leaves are labeled by P(z) and internal nodes correspond to parents branching over their values.

Captures irrelevant variables

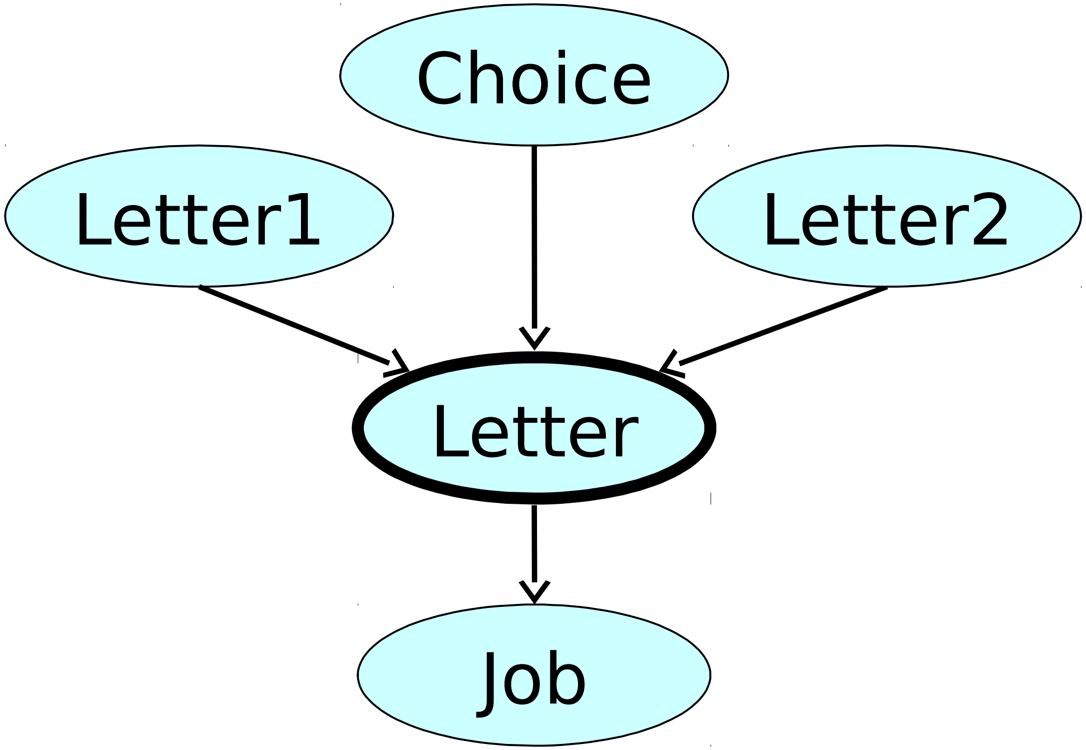




Multiplexer CPD

A CPD P(Y|A,Z1,Z2,…,Zk) is a multiplexer iff Val(A)=1,2,…k, and

P(Y|A,Z1,…Zk)=Z_a



Mixture of trees

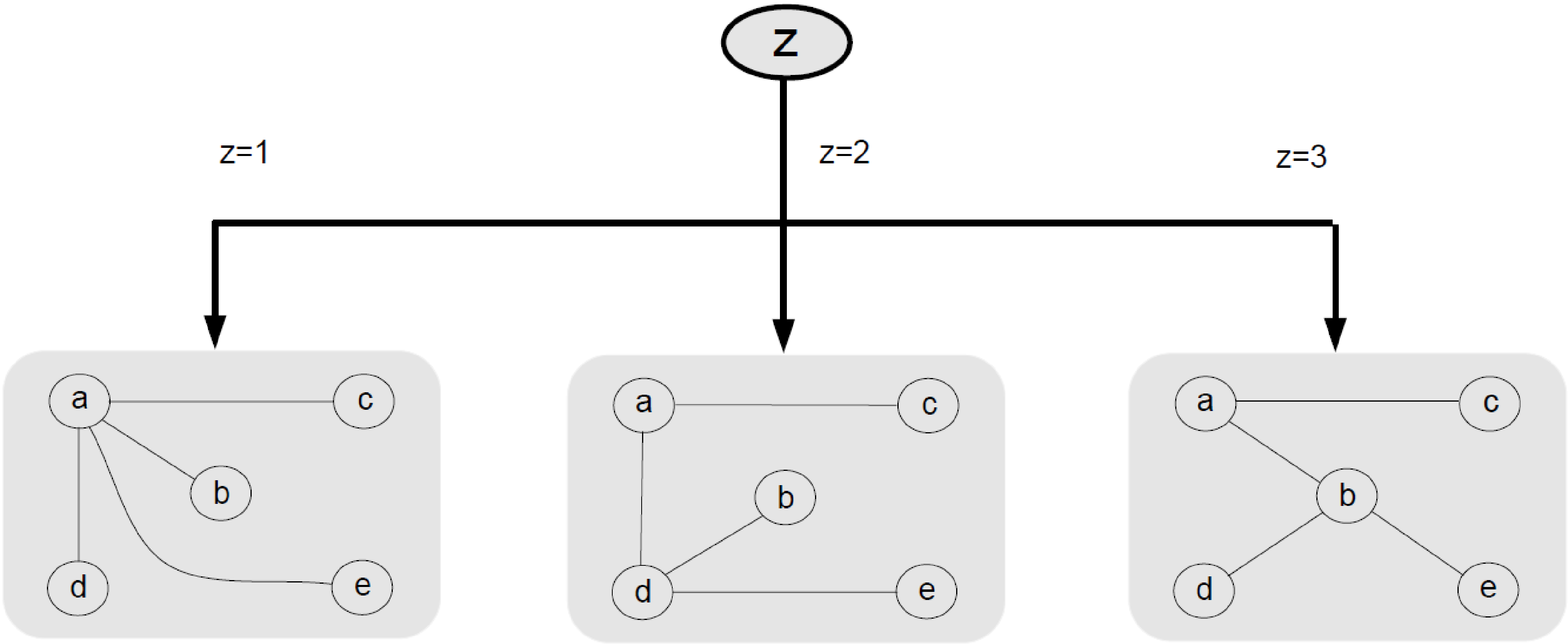


Figure 1: A mixture of trees over a domain consisting of random variables $V = \{a, b, c, d, e\}$, where z is a hidden *choice variable*. Conditional on the value of z, the dependency structure is a tree. A detailed presentation of the mixture-of-trees model is provided in Section 3.

Meila and Jordan, 2000

Mixture model with shared structure

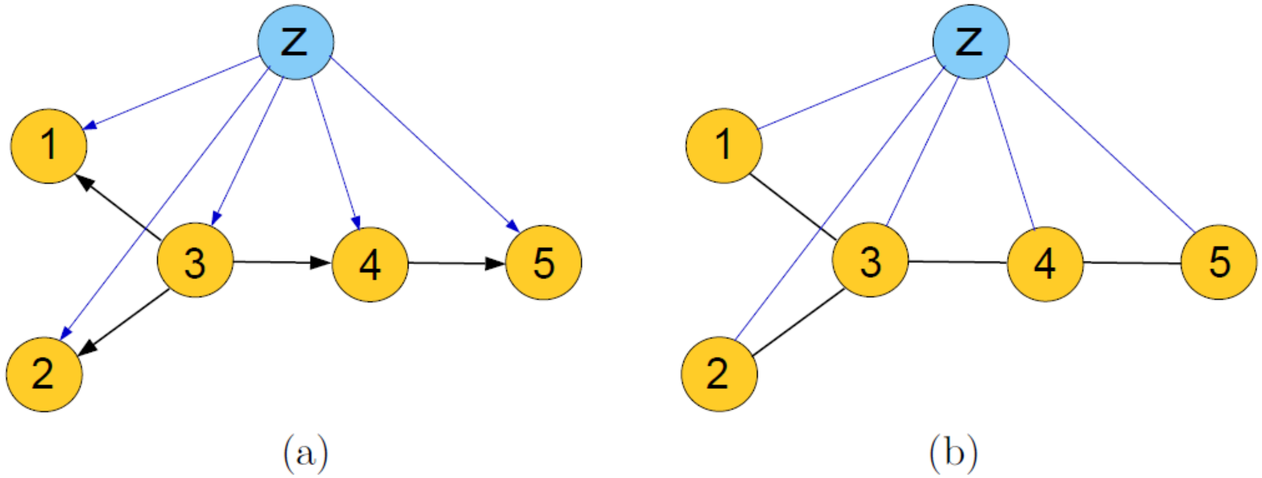


Figure 4: A mixture of trees with shared structure (MTSS) represented as a Bayes net (a) and as a Markov random field (b).

Meila and Jordan, 2000

Deterministic CPTs

A deterministic, or functional CPT

is one in which every probability is either 0 or 1

A deterministic CPT for variable E with values $e_1, \ldots, e_m$

can be represented by a set of propositional sentences of the form:

$$\Gamma_i \iff E = e_i,$$

where we have one rule for each value e_i of E, and the premises Γ_i are mutually exclusive and exhaustive.

The CPT for variable E is then given by

$$\theta_{e_i|\alpha} = \begin{cases} 1, & \text{if parent instantiation } \alpha \text{ is consistent with } \Gamma_i; \\ 0, & \text{otherwise.} \end{cases}$$

Deterministic CPTs

Can we use hidden variables?

A	X	C	$\theta_{c\|a,x}$
high	ok	high	0
low	ok	high	1
high	stuckat0	high	0
low	stuckat0	high	0
high	stuckat1	high	1
low	stuckat1	high	1

We can represent this CPT as follows

$$(X = \text{ok} \wedge A = \text{high}) \vee X = \text{stuckat0} \iff C = \text{low}$$
$$(X = \text{ok} \wedge A = \text{low}) \vee X = \text{stuckat1} \iff C = \text{high}$$

Generalized linear models

(see Koller 5.4.2)

Let Y be a binary-valued variable with parents the X_i's that can take a numerical value (discrete).
The CPT P(Y|X_1,…X_n) is a logistic CDT if there are w's such that

$$P(y|x_1, \ldots, x_n) = \text{sigmoid}(w_0 + \sum_{i=1}^{k} w_i x_i)$$

$$sigmoid(z) = \frac{e^z}{1 + e^z}$$

Mixed Networks

(Dechter 2013)

- Augmenting Probabilistic networks with constraints because:
  - Some information in the world is deterministic and undirected (X not-eq Y)
  - Some queries are complex or evidence are complex (cnfs)
- Queries are probabilistic queries

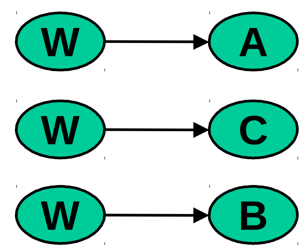
Probabilistic Reasoning

Party example: the weather effect

Alex is-likely-to-go in bad weather

Chris rarely-goes in bad weather

Becky is indifferent but unpredictable



P(A|W=bad)=.9

P(C|W=bad)=.1

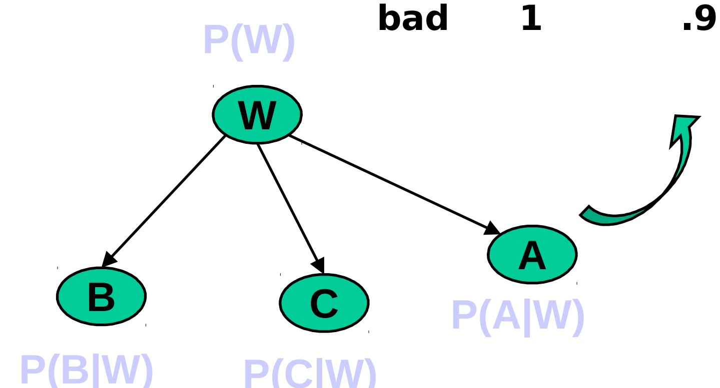
P(B|W=bad)=.5

Questions:

Given bad weather, which group of individuals is most likely to show up at the party?

What is the probability that Chris goes to the party but Becky does not?

W	A	P(A|W)
good	**0**	**.01**
good	**1**	**.99**
bad	**0**	**.1**
bad	**1**	**.9**

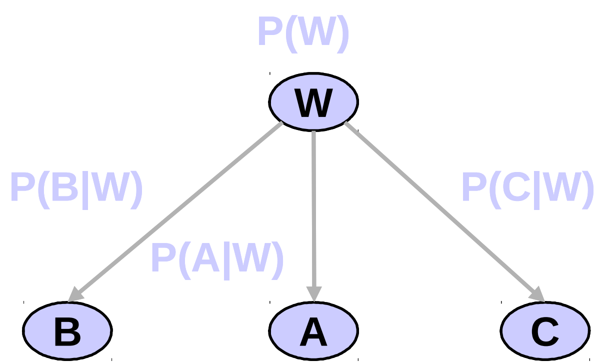


P(W,A,C,B) = P(B|W) · P(C|W) · P(A|W) · P(W)

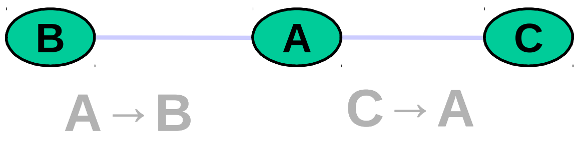
P(A,C,B|W=bad) = 0.9 · 0.1 · 0.5

Party example again

PN



CN



Query:
Is it likely that Chris goes to the party if Becky does not but the weather is bad?

$$P(C, \neg B \mid w = bad, A \rightarrow B, C \rightarrow A)$$

Semantics?

Algorithms?

Markov Networks

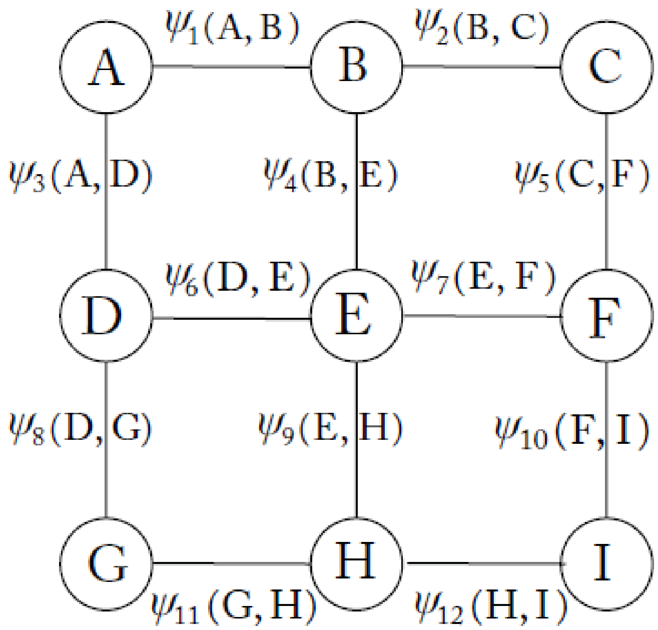
Definition 2.23 Markov networks. A Markov network is a graphical model $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{H}, \prod \rangle$ where $\mathbf{H} = \{\psi_1, \ldots, \psi_m\}$ is a set of potential functions where each potential ψ_i is a non-negative real-valued function defined over a scope of variables $\mathcal{S} = \{\mathbf{S}_1, ..., \mathbf{S}_m\}$. $\mathbf{S}_i$. The Markov network represents a global joint distribution over the variables $\mathbf{X}$ given by:

$$P_{\mathcal{M}} = \frac{1}{Z} \prod_{i=1}^{m} \psi_i \quad , \quad Z = \sum_{\mathbf{X}} \prod_{i=1}^{m} \psi_i$$

where the normalizing constant Z is called the partition function.

Dechter, chapter 2

Markov network example



(a)

D	E	$\psi_6(D, E)$
0	0	20.2
0	1	12
1	0	23.4
1	1	11.7

(b)

Figure 2.6: (a) An example 3×3 square grid Markov network (ising model) and (b) an example potential $H_6(D, E)$.

$$F(a, b, c, d, e, f, g, h, i) \propto$$

$$\psi_1(a, b) \cdot \psi_2(b, c) \cdot \psi_3(a, d) \cdot \psi_4(b, e) \cdot \psi_5(c, f) \cdot \psi_6(d, e) \cdot \psi_7(e, f) \cdot \psi_8(d, g)$$

$$\cdot \psi_9(e, h) \cdot \psi_{10}(f, i) \cdot \psi_{11}(g, h) \cdot \psi_{12}(h, I)$$

where $Z = \sum_{a,b,c,d,e,f,g,h,i} F(a, b, c, d, e, f, g, h, i)$ is the partition function.

Mixed Networks

Definition 2.25 Mixed networks. Given a belief network $\mathcal{B} = \langle \mathbf{X}, \mathbf{D}, \mathbf{P}_G, \prod \rangle$ that expresses the joint probability $P_{\mathcal{B}}$ and given a constraint network $\mathcal{R} = \langle \mathbf{X}, \mathbf{D}, \mathbf{C}, \bowtie \rangle$ that expresses a set of solutions denoted ρ, a mixed network based on $\mathcal{B}$ and $\mathcal{R}$ denoted $\mathcal{M}_{(\mathcal{B},\mathcal{R})} = \langle \mathbf{X}, \mathbf{D}, \mathbf{P}, \mathbf{C} \rangle$ is created from the respective components of the constraint network and a Bayesian network as follows: the variables $\mathbf{X}$ and their domains are shared (we could allow non-common variables and take the union), and the functions include the CPTs in $\mathbf{P}_G$ and the constraints in $\mathbf{C}$. The mixed network expresses the conditional probability $P_{\mathcal{M}}(\mathbf{X})$:

$$P_{\mathcal{M}}(\mathbf{x}) = \begin{cases} P_{\mathcal{B}}(\mathbf{x} \mid \mathbf{x} \in \rho), & if \;\; \mathbf{x} \in \rho \\ 0, & otherwise. \end{cases}$$