Bayesian networks

Syntax and Semantics

Bayesian Classifiers

- Generative models: model $P(x_1, \ldots, x_d, \omega)$
- E.g., Naive Bayes classifier
  - Naive Bayes assumption
- Use Bayesian networks to model $P(x_1, \ldots, x_d, \omega)$
  - Systematically use domain knowledge about conditional independence relationships

Uncertainty Reasoning

- Bayesian networks, a probabilistic reasoning system, have emerged as the method of choice for uncertainty reasoning.
- Probability theory provides a framework for representing and reasoning with uncertainty knowledge
- Joint probability distributions allow one to model uncertain beliefs and to answer any questions about the domain.

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Probabilistic Inference

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Joint probability distributions

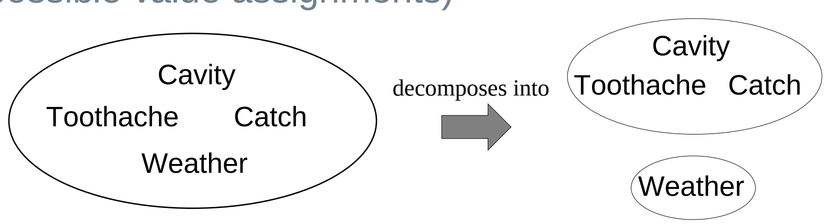
- In principle, joint distributions can be used to answer any probabilitic queries.
- A joint probability distribution has an exponential size in the number of variables of interest $O(r^n)$
  - Computational viewpoint: computing marginal and conditional probabilities poses a complexity challenge
  - Modelling viewpoint: requires a large number of probabilities that can be impossible to obtain directly in certain situations.
- Use domain knowledge to relieve this problem, one type of knowledge is independence relationships

Independence

Two sets of variables A and B are independent iff

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$

(for all possible value assignments)



$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$
$$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

32 entries reduced to 12

Conditional independence

X is independent of Y given Z, denoted $I(X,Z,Y)$, iff

$$P(x|y,z) = P(x|z), \forall x \in Dm(X), y \in Dm(Y), z \in Dm(Z)$$

Equivalent statements:

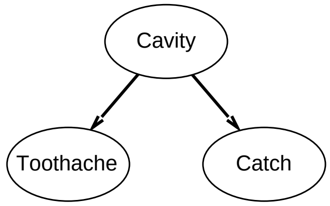
$$P(Y|X,Z) = P(Y|Z), \quad P(X,Y|Z) = P(X|Z)P(Y|Z)$$

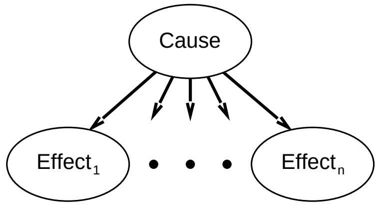
$Catch$ is conditionally independent of $Toothache$ given $Cavity$:
$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$
Conditional independence is one of the most basic and robust form of knowledge about uncertain environments.

Conditional independence

In some cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.
E.g., a naive Bayes model:

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





Bayesian networks

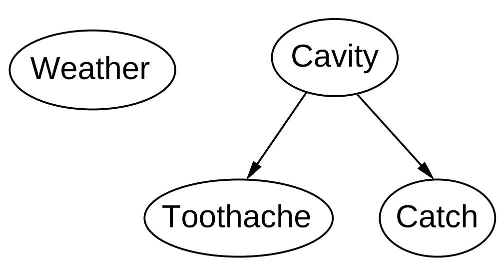
Bayesian networks is a graphical modelling tool for specifying probability distributions

- Encode conditional independence assertions explicitly
- Provides a compact representation of joint distribution
- Support efficient algorithms for answering probabilistic queries

Bayesian networks

Bayesian network is a directed acyclic graph (DAG)

- Nodes: random variables of interest
- Edges: *direct* (causal) influences
- Each node is annotated with a conditional distribution $\mathbf{P}(X_i|Parents(X_i))$
- Each variable is asserted to be conditionally independent of its non-descendants given its parents.



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: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

Network topology reflects "causal" knowledge:

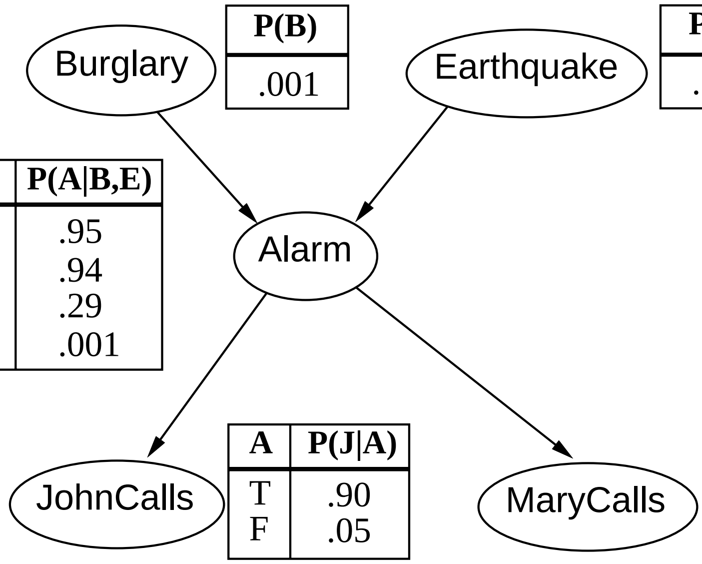
– A burglar can set the alarm off

– 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

BNs - Qualitative part

- We interpret each DAG G as a compact representation of the following independence statements: $\{I(V, Parents(V), Non-Descendants(V)):$ for all variables V in $G\}$
  - Every variable is conditionally independent of its non-descendants given its parents.
- This set of independence statements are often referred to as the *local Markovian assumptions* of DAG G

BN as a Knowledge Base

- Since the joint distribution must satisfy the independence assumptions, the chain rule of BN

$$Pr(x_1,\ldots,x_n) = \prod_i Pr(x_i|pa_i)$$

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

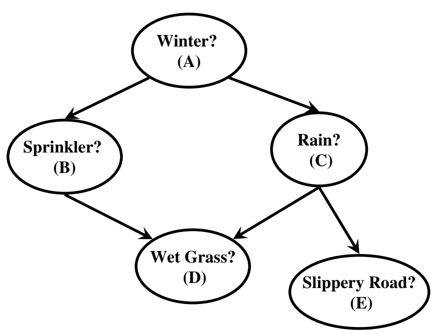
$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

The joint distribution can be constructed by specifying the local conditional distributions $Pr(x_i|pa_i)$'s

Parameterizing BNs

- The joint distribution can be constructed by specifying the local CPDs $P(x_i|pa_i)$'s
- A *parameterization* Θ of the DAG G:
  - Θ consists of a set of parameters $\Theta_{X_i|Pa_i}$ for each CPD
  - For discrete random variables: conditional probability tables (CPTs)

A Bayesian network



A	Θ_A
true	.6
false	.4

A	B	$\Theta_{B\|A}$
true	true	.2
true	false	.8
false	true	.75
false	false	.25

A	C	$\Theta_{C\|A}$
true	true	.8
true	false	.2
false	true	.1
false	false	.9

B	C	D	$\Theta_{D\|B,C}$
true	true	true	.95
true	true	false	.05
true	false	true	.9
true	false	false	.1
false	true	true	.8
false	true	false	.2
false	false	true	0
false	false	false	1

C	E	$\Theta_{E\|C}$
true	true	.7
true	false	.3
false	true	0
false	false	1

Bayesian network

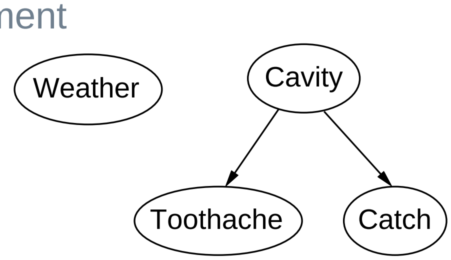
- A *Bayesian network* over a set of variables $X_1, \ldots, X_n$ is a pair (G, Θ) such that

$$Pr(x_1, \ldots, x_n) = \prod_i \theta_{x_i|pa_i}$$

- A BN provides a compact representation of joint distribution: $O(n * d^{k+1})$ vs. $O(d^n)$ (every variable takes up to d values and has at most k parents)
  For burglary net, $1+1+4+2+2=10$ numbers (vs. $2^5-1=31$)
- BNs support efficient algorithms for answering probabilistic queries

BN as modeling tool

- Human good at low order mariginal and conditional probabilities, much difficult to judge joint probability
- The parents of X are those variables judged to be *direct causes* of X or have *direct influence* on X
- The parameters requested from model builders are conditional probabilities that quantify conceptual relationships in one's mind, e.g., cause-effect relations, which are psychologically meaningful, and may be obtained by direct measurement



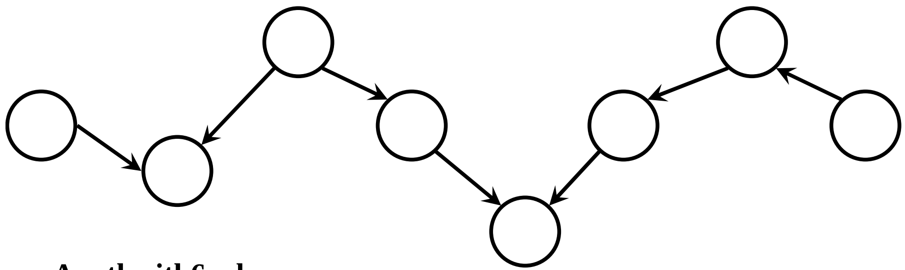
BNs as a Logic of Dependences

- A BN can be viewed as an inference instrument for deducing new independence relationships from those used in constructing the network.
- Input independence statements, the local Markovian assumptions, $\{I(X_i, PA_i, \{X_1, X_2, \ldots, X_{i-1}\} - PA_i)\}$
- Are there other independencies that hold in *every* distribution that factorizes over G?
- Additional independencies can be deduced by logical inference rules, captured using a graphical test known as *d-separation*, without reference to numerical quantities

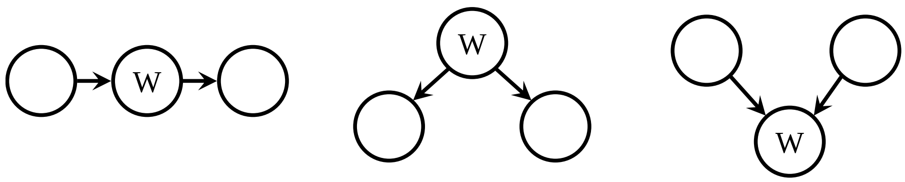
Capturing Indep. Graphically

- How to represent dependence relations using a DAG G?
- To decide $I(X, Z, Y)$, we need to consider every path between a node in X and a node in Y, and then ensure that the path is blocked by Z
- The best way to understand the notion of blocking is to view the path as a *pipe*, and to view each variable W on the path as a *valve*
- A valve W is either *open* or *closed*

Capturing Indep. Graphically



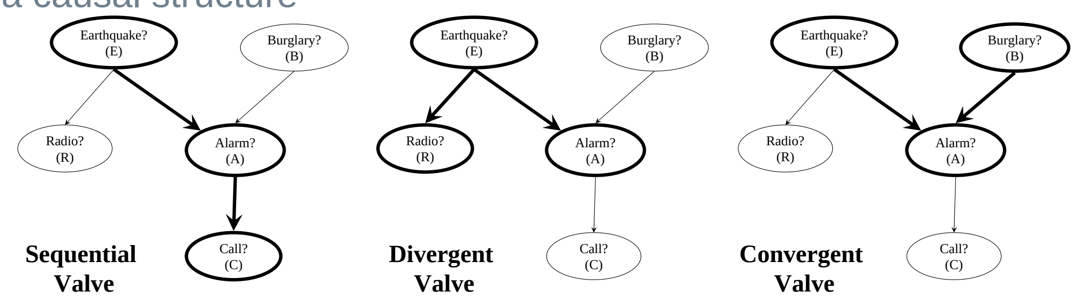
Apathwith6valves



Sequentialvalve **Divergentvalve** **Convergentvalve**

Capturing Indep. Graphically

To obtain more intuition on how these types of valves correspond to independence relations, it is best to interpret the given DAG as a causal structure



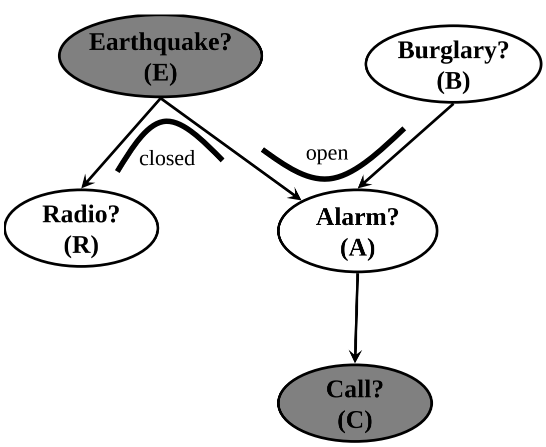
A general pattern of causal relationships: observation on a common consequence of two independence causes tend to render those causes dependent – "Explaining away effect"

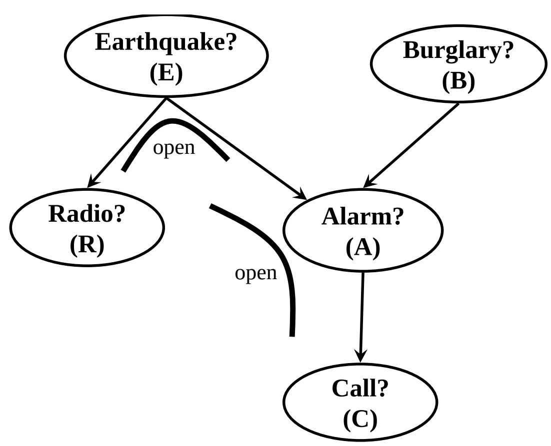
d-separation

- A sequential valve $\rightarrow W \rightarrow$ is closed iff W appears in Z
- A divergent valve $\leftarrow W \rightarrow$ is closed iff W appears in Z
- A convergent valve $\rightarrow W \leftarrow$ is closed iff neither variable W nor any of its descendants appears in Z

Definition [d-separation] A path is said to be *blocked* by a set of nodes Z iff at least one valve on the path is closed given Z. (Otherwise, the path is said to be *unblocked* or *active*.)
A set of nodes X and Y are *d-separated* by a set Z in a DAG G, denoted by $dsep_G(X, Z, Y)$, iff every path between a node in X and a node in Y is blocked by Z.

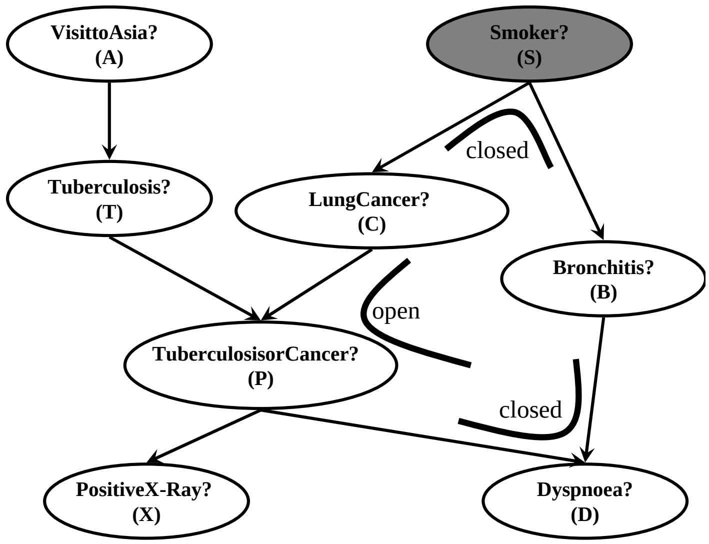
d-separation





$dsep_G(R, EC, B)$?
$dsep_G(R, \emptyset, C)$?

d-separation



$dsep_G(C, S, B)$?

d-separation

Theorem (soundness) If $Pr(.)$ is induced by a BN G, then

$$dsep_G(X, Z, Y) \Longrightarrow I(X, Z, Y)$$

i.e., every d-separation condition displayed in G corresponds to a valid independence relationship

- Often called *global Markovian property*
- *Some* distributions will induce independences not revealed by d-separation
- Complexity: d-separation can be decided in linear time in the size of G

I-map

- A DAG G is said to be an **I-map** (Independence MAP) of a probability distribution Pr if for every three disjoint sets of vertices X, Y, and Z

  $$dsep_G(X,Z,Y) \Longrightarrow I(X,Z,Y)$$

  i.e., every d-separation condition displayed in G corresponds to a valid independence relationship
- A DAG is a **minimal I-map** of Pr if none of its edges can be deleted without destroying its I-mapness.
- Alternative definition of BN: a DAG G is called a Bayesian network of Pr iff G is a minimal I-map of Pr.

Bayesian networks

"Given a distribution Pr, can we construct a BN G?"

Given a probability distribution $Pr(X_1, X_2, \dots, X_n)$ and an ordering $d = (X_1, X_2, \dots, X_n)$ of the variables, the DAG created by designating as parents of X_i any minimal set PA_i of predecessors satisfying

$$Pr(x_i|pa_i) = Pr(x_i|x_1, \dots, x_{i-1}),\ PA_i \subseteq \{X_1, X_2, \dots, X_{i-1}\}$$

is a Bayesian network of Pr. If Pr is strictly positive, then all of the parent sets are unique and the Bayesian network is unique given d.

Constructing Bayesian networks

Given a distribution Pr, can we construct a BN?

1. Choose an ordering of variables $X_1, \ldots, X_n$

2. For i = 1 to n

add X_i to the network

identify a minimal subset $Parents(X_i)$ from $X_1, \ldots, X_{i-1}$ such that

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

Need a series of locally testable assertions of conditional independence

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: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$

Causal knowledge:

– A burglar can set the alarm off

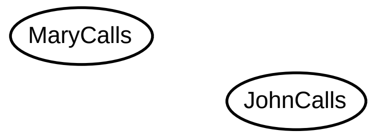
– An earthquake can set the alarm off

– The alarm can cause Mary to call

– The alarm can cause John to call

Example

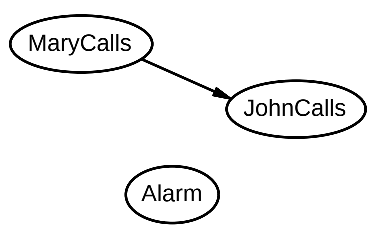
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?

Example

Suppose we choose the ordering M, J, A, B, E

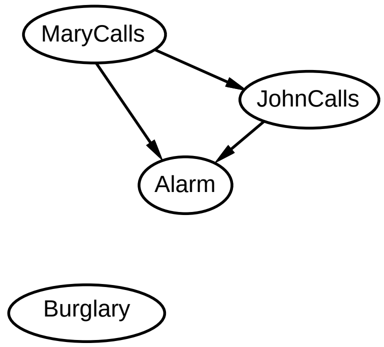


$P(J|M) = P(J)$? No

$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$?

Example

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

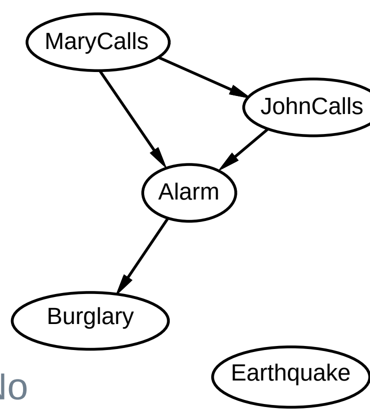
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No

$P(B|A,J,M) = P(B|A)$?

$P(B|A,J,M) = P(B)$?

Example

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No

$P(B|A,J,M) = P(B|A)$? Yes

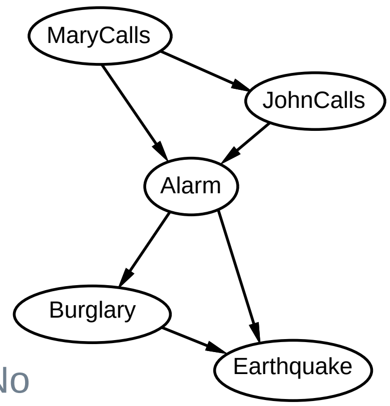
$P(B|A,J,M) = P(B)$? No

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

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

Example

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No

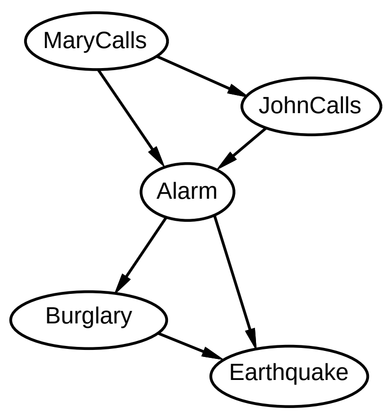
$P(B|A,J,M) = P(B|A)$? Yes

$P(B|A,J,M) = P(B)$? No

$P(E|B,A,J,M) = P(E|A)$? No

$P(E|B,A,J,M) = P(E|A,B)$? Yes

Example contd.



Deciding conditional independence is hard in noncausal directions (Causal models and conditional independence seem hardwired for humans!)

Assessing conditional probabilities is hard in noncausal directions

Network is less compact: $1+2+4+2+4=13$ numbers needed (vs. 10)

Role of Causality

- The interpretation of directed acyclic graphs as carriers of independence assumptions does not necessarily imply causation
- The ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation
- In practice, DAG models are rarely used in any variable ordering other than those which respect the direction of time and causation
- There are many advantages of building DAG models around causal rather than associational information

Role of Causality

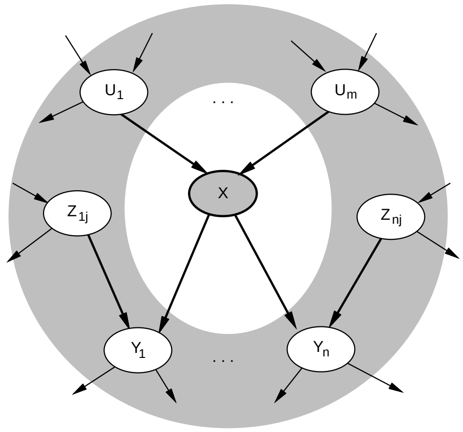
- The judgments required in the construction of the model are more meaningful, more accessible, and hence more reliable.
- Conditional independence judgments are accessible (hence reliable) only when they are anchored onto more fundamental building blocks of our knowledge, such as causal relationships.
- If conditional independence judgments are byproducts of stored causal relationships, then representing those relationships directly would be a more natural way of expressing what we know or believe about the world
  - The philosophy behind *causal Bayesian networks*.

Markov Blanket

- A *Markov Blanket* for X is a set of variables B which, when known, will render every other variable irrelevant to X, i.e., $I(X, B, R)$, where R is the set of all variables other than X and B
- A minimal Markov Blanket is known as a *Markov Boundary*, i.e., none of its proper subsets is a Markov blanket.
- The Markov Boundary for a variable is not unique, unless the distribution is strictly positive
- Feature selection in classification

Markov Blanket

- If Pr is induced by DAG G, then a Markov blanket for variable X can be constructed using its parents, children, and spouses (parents of its children) in G
- Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Independence equivalence

Independence equivalence: two DAGs may encode the same set of indpendence relations

$$dsep_{G_1}(X,Z,Y) \iff dsep_{G_2}(X,Z,Y)$$

- Aka observational equivalence: G_1 is a BN of P iff G_2 is a BN of P
- Consequences to learning BNs from data: place a limit on our ability to infer directionality from probabilities alone

Independence equivalence

Theorem Two DAGs are independence equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow.