Spezielle Themen der Künstlichen Intelligenz

8. Termin:

Bayesian Networks

CITEC

Sociable Agents

Degrees of belief as probabilities

Degree of belief or probability of a *world* $Pr(\omega)$

Degree of belief or probability of a *sentence* $Pr(\alpha) := \sum_{\omega \models \alpha} Pr(\omega)$

State of belief or joint probability distribution $\sum_{\omega_i} Pr(\omega_i) = 1$

World	**Earthquake**	**Burglary**	**Alarm**	*Pr(.)*
w1	*true*	*true*	*true*	*.0190*
w2	*true*	*true*	*false*	*.0010*
w3	*true*	*false*	*true*	*.0560*
w4	*true*	*false*	*false*	*.0240*
w5	*false*	*true*	*true*	*.1620*
w6	*false*	*true*	*false*	*.0180*
w7	*false*	*false*	*true*	*.0072*
w8	*false*	*false*	*false*	*.7128*

$$Pr(Earthquake) = .1$$
$$Pr(Burglary) = .2$$
$$Pr(Alarm) = .2442$$

Independence

(Absolute) Independence

a given state of beliefs finds an event independent of another event iff

$$Pr(\alpha|\beta) = Pr(\alpha) \text{ or } Pr(\beta) = 0$$
$$Pr(\alpha \wedge \beta) = Pr(\alpha) \cdot Pr(\beta)$$

Conditional Independence

state of belief Pr finds α conditionally independent of β given event γ iff

$$Pr(\alpha|\beta \wedge \gamma) = Pr(\alpha|\gamma) \text{ or } Pr(\beta \wedge \gamma) = 0$$
$$Pr(\alpha \wedge \beta|\gamma) = Pr(\alpha|\gamma)Pr(\beta|\gamma) \text{ or } Pr(\gamma) = 0$$

Independence is a dynamic notion!

- new evidence can make *(in-)dependent* facts *conditionally (in-)dependent*
- determined by the initial state of belief (joint full distribution) one has

Conditional Independence

Example:

Given two noisy, unreliable sensors

Initial beliefs

- *Pr(Temp=normal)=.80*
- *Pr(Sensor1=normal)=.76*
- *Pr(Sensor2=normal)=.68*

Temp	*sensor1*	*sensor2*	*Pr(.)*
normal	*normal*	*normal*	*.576*
normal	*normal*	*extreme*	*.144*
normal	*extreme*	*normal*	*.064*
normal	*extreme*	*extreme*	*.016*
extreme	*normal*	*normal*	*.008*
extreme	*normal*	*extreme*	*.032*
extreme	*extreme*	*normal*	*.032*
extreme	*extreme*	*extreme*	*.128*

After checking sensor1 and finding its reading is *normal*

- *Pr(Sensor2=normal | Sensor1=normal) ~ .768* ➔ initially dependent

But after observing that temperatur is *normal* ....

- *Pr(Sensor2=normal | Temp=normal) = .80*
- *Pr(Sensor2=normel | Temp=normal, Sensor1=normal) = .80* ➔ cond. independent

Variable set independence

<u>*Notation*</u>:
independence between sets of variables **X,Y, Z** in a belief state *Pr* denoted as $I_{Pr}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$

<u>*Example*</u>:

- **X**={A,B}, **Y**={C}, **Z**={D,E} (all Boolean variables)
- $I_{Pr}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ denotes 4x2x4=32 different independent statements:
  - A∧B indep. of C given D∧E
  - A∧¬B indep. of C given D∧E
  - ..
  - ..
  - ¬A∧¬B indep. of ¬C given ¬D∧¬E

Conditional Independence

Is a special case of mutual information:

$$MI(X;Y) := \sum_{x,y} Pr(x,y) log_2 \frac{Pr(x,y)}{Pr(x)Pr(y)}$$

- quantifies impact of observing one variable on uncertainty in another
- non-negative
- zero **iff** *X* and *Y* are independent

Relation to entropy as defined earlier:

$$MI(X;Y) := ENT(X) - ENT(X|Y) = ENT(Y) - ENT(Y|X)$$

- with conditional entropy: $ENT(X|Y) := \sum_y Pr(y) log_2 ENT(X|y)$
  $ENT(X|y) := -\sum_x Pr(x|y) log_s Pr(x|y)$
- <u>*Note*</u>: $ENT(X|Y) \leq ENT(X)$

Properties of beliefs

Repeated application of Bayes Conditioning gives chain rule

$$Pr(\alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_n) = Pr(\alpha_1|\alpha_2 \wedge ... \wedge \alpha_n)Pr(\alpha_2|\alpha_3 \wedge ... \wedge \alpha_n)...Pr(\alpha_n)$$

If events β_i are mutually exclusive and exhaustive, we can apply case analysis (or law of total probability) to compute a belief in α :

$$Pr(\alpha) = \sum_i Pr(\alpha \wedge \beta_i) = \sum_i Pr(\alpha|\beta_i)Pr(\beta_i)$$

- compute belief in α by adding up beliefs in exclusive cases $\alpha \wedge \beta_i$ that cover the conditions under which α holds

Bayes rule or Bayes theorem:

- follows directly from product rule

$$Pr(\alpha|\beta) = \frac{Pr(\beta|\alpha)P(\alpha)}{Pr(\beta)}$$

Proposed a solution to problem of "inverse probability"

- published posthumously by R. Price in Phil. Trans. of Royal Soc. Lond. (1763)

Bayes' theorem

- expresses the posterior (i.e. after evidence *E* is observed) of a hypothesis *H* in terms of the priors of *H* and *E*, and the prob of *E* given *H*
- implies that evidence has a stronger confirming effect if it was more unlikely before being observed

Thomas Bayes (1702–1761)

$$Pr(\alpha|\beta) = \frac{Pr(\beta|\alpha)P(\alpha)}{Pr(\beta)}$$

Bayes rule

Example:

A patient has been tested positive for a disease D, which one in every 1000 people has. The test T is not reliable (2% false positive rate and 5% false negative rate). What is our belief *Pr(D|T)*?

$$Pr(D) = 1/1000$$
$$Pr(T|\neg D) = 2/100 \quad \Rightarrow \quad Pr(\neg T|\neg D) = 98/100$$
$$Pr(\neg T|D) = 5/100 \quad \Rightarrow \quad Pr(T|D) = 95/100$$

$$P(D|T) = \frac{95/100 \cdot 1/1000}{Pr(T)}$$
$$P(T) = Pr(T|D)Pr(D) + Pr(T|\neg D)Pr(\neg D)$$
$$Pr(D|T) = \frac{95}{2093} = 4.5\%$$

Soft & hard evidence

Useful to disntighuish two types of evidence

- hard evidence: information that some event has occurred
- soft evidence: unreliable hint that an event have may occurred
  - neighbour with hearing problem tells us he had heard the alarm trigger
  - can be interpreted in terms of noisy sensors

So far, conditioning on hard evidence. How to update in light of soft evidence? Two methods:

1. new state of beliefs *Pr'* = old beliefs + new evidence („all things considered") ➔ bayesian conditioning leads to Jeffrey's rule:

$$Pr'(\alpha) = qPr(\alpha|\beta) + (1-q)Pr(\alpha|\neg\beta) \text{ with } Pr'(\beta) = q$$
$$Pr'(\alpha) = \sum_i q_i Pr(\alpha|\beta_i) \text{ with } q_i \text{ exclusive and exhaustive}$$

Soft & hard evidence

2. use strenght of evidence, independent of current beliefs („nothing else considered“)

- *Definition*: Odds of event: $O(\beta) := \frac{Pr(\beta)}{Pr(\neg\beta)}$
  - states how many times we believe more in β than in $\neg\beta$
- *Definition*: Bayes factor of the „strength“ of evidence: $k := \frac{O'(\beta)}{O(\beta)}$
  - relative change induced on odds of β
  - k=1: indicates neutral evidence
    k=2: indicates evidence strong enough to double the odds of β
    k➔inf.: hard evidence conforming β, k➔0: hard evidence against β
- update according to evidence β with known Bayes factor *k*:

$$Pr'(\beta) = \frac{kPr(\beta)}{kPr(\beta)+Pr(\neg\beta)}$$

(from def. of *O*)

$$Pr'(\alpha) = \frac{kPr(\alpha\wedge\beta)+Pr(\alpha\wedge\neg\beta)}{kPr(\beta)+Pr(\neg\beta)}$$

(together with Jeffrey‘s rule)

Soft evidence

Example: Murder with three suspects, investigator Rich has the following state of belief:

- *O*(killer=david) = *Pr*(david)/*Pr*(not david) =2

	Killer	*Pr(.)*
ω_1	david	2/3
ω_2	dick	1/6
ω_3	jane	1/6

new soft evidence is obtained that triples the odds of killer=david

- *k*=*O*‘(killer=david)/*O*(killer=david) = 3

➔ new belief in David being the killer:

- *Pr*‘(killer=david) = (3*2/3) / (3*2/3+1/3) = 6/7

only the first statement (*k*; nothing else considered) can be used by other agents to update their belief according to β

So, what's this all good for?

Key observation:

Full joint distribution or state of belief can be used to model uncertain beliefs and update them in face of soft or hard evidence

- determines prob for every event given any combination of evidence

But, the joint distribution is exponential

- $O(d^n)$ with n random variables and domain size d

Independence would help: $O(d^n) \rightarrow O(n)$

- absolute independence unfortunately rare
- conditional independence not so rare
  „our most basic, robust, and commonly available knowledge about uncertain environments"

So, what's this all good for?

Independence allows to decompose the joint distribution

- *Pr(Cavity,Catch,Toothache)* ➔ 2^3=8 worlds needed
  = *Pr(Tootha.,Catch|Cavity) Pr(Cavity)* (Bayes rule)
  = *Pr(Tootha.|Cavity) Pr(Catch|Cavity) Pr(Cavity)* (cond. ind. of *Tootha.* & *Catch* given *Cavity*) ➔ 2+2+1=5 worlds needed

Common pattern:

a cause *directly implies* multiple effects, all of which are conditionally independent given the cause

$$Pr(Cause, E_1, ..., E_n) = Pr(Cause) \prod_i Pr(E_i|Cause)$$

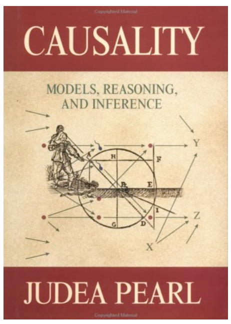
- the cause sufficiently „explains" each effect, knowing about other effects doesn't change the belief in it anymore
- Naive Bayes model (also called Bayesian classifier):
  Bayes rule + presumed independence where there is no

Bayesian networks

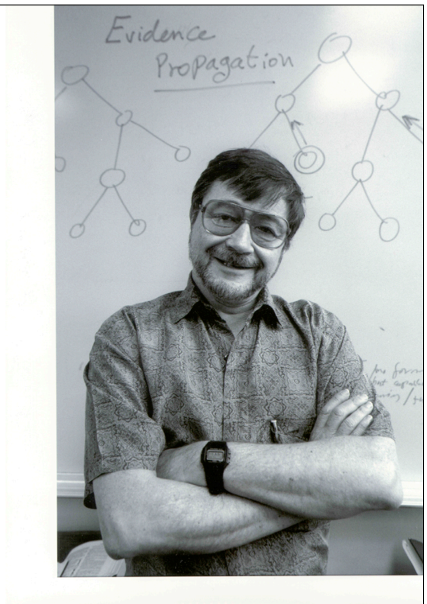
<u>*Definition:*</u> *A* Bayesian network for variables Z is a pair (G,Θ) with

- Structure G: a directed acyclic graph with
  - a set of nodes, one per random variable
  - a set of edges representing *direct causal influence* between variables
- Parametrization Θ: a conditional probability table (CPT) for each variable
  - probability distribution for each node given its parents: *$Pr(X_i \mid Parents(X_i))$* or *$Pr(X_i)$* if there are no parents
  - parameterizes the independence structure

(1988)

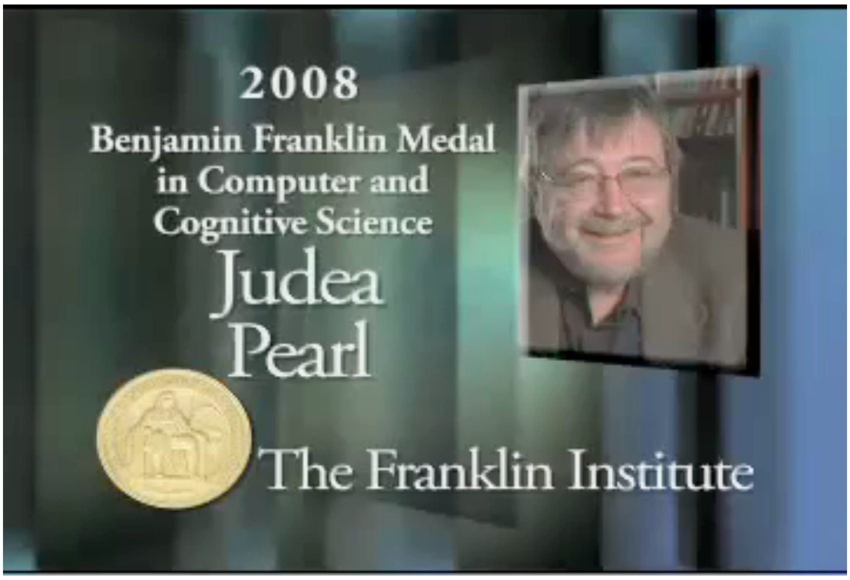


(2000)



Judea Pearl coined the term Bayesian networks to emphasize:

- the the subjective nature of the input information
- the reliance on Bayes's conditioning as the basis for updating beliefs
- the distinction between causal and evidential modes of reasoning



Bayesian networks

Bayesian networks

- rely on insight that independence forms a significant aspect of beliefs
- a compact representation of a full belief state (= joint distribution)
- also called probabilistic networks or DAG models

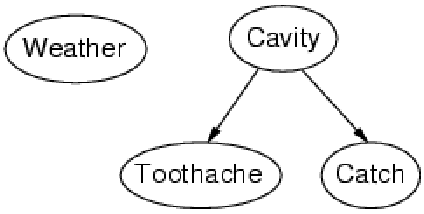
Each Bayesian network defines a set of cond. indep. statements:

$I(V, Parents(V), NonDescendants(V))$

- every variable is conditionally indep. of its nondescendants given its parents
  - Markovian assumptions: $Markov(G)$
- $Parents(V)$ are direct causes, $Descendants(V)$ are effects of V
  - given direct causes of *V*, beliefs in *V* are no longer influenced by any other variable, except possibly by its effects

Bayesian networks

Example:



- *Weather* is (even absolutely) independent of all other variables
- *Cavity* causally influences *Toothache* and *Catch*
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Bayesian networks

Example:
„I'm at work, neighbor John calls to say my burglar alarm is ringing. Sometimes it's set off by minor earthquakes. John sometimes confuses the alarm with a phone ringing. Real earthquakes usually are reported on radio. This would increase my belief in the alarm triggering and in receiving John‘s call.“

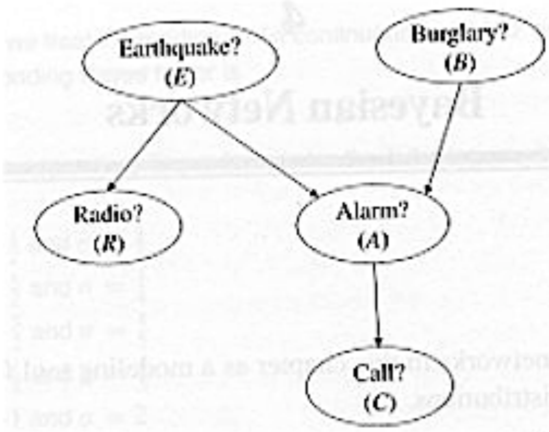
Variables: *Burglary, Earthquake, Alarm, Call, Radio*

Network topology reflects believed causal knowledge about domain:

- burglar and earthquake can set the alarm off
- alarm can cause John to call
- earthquake can cause a radio report
- + *independence assumptions......?*

Bayesian networks

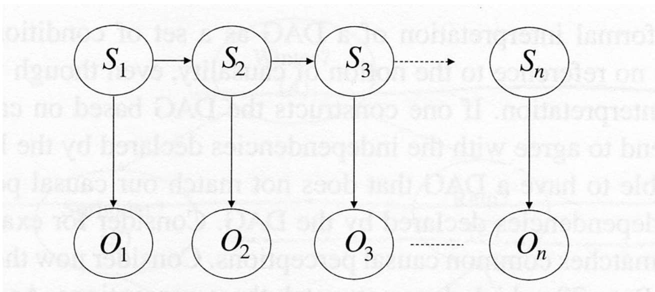
$$I(C, A, \{B, E, R\})$$
$$I(R, E, \{A, B, C\})$$
$$I(A, \{B, E\}, R)$$
$$I(B, \{\}, \{E, R\})$$
$$I(E, \{\}, B)$$



- given *Alarm, Call* is cond. indep. of *Earthquake, Burglary, Radio*
- given *Earthquake, Radio* is cond. indep. of *Alarm, Burglary, Call*
- given *Earthquake* and *Burglary, Alarm* is cond. indep. of *Radio*
- *Earthquake* and *Burglary* are indep. of each non-descendant

Bayesian networks

Hidden Markov Models (HMM)



- S_i represent state of a dynamic system at times i
- O_i represent sensor readings at times i

$$I(S_t, S_{t-1}, \{S_1, ..., S_{t-2}, O_1, ..., O_{t-1}\})$$

- given last state of the system, our belief in present system state is indep. of any other information from the past

Bayesian networks

Notation:

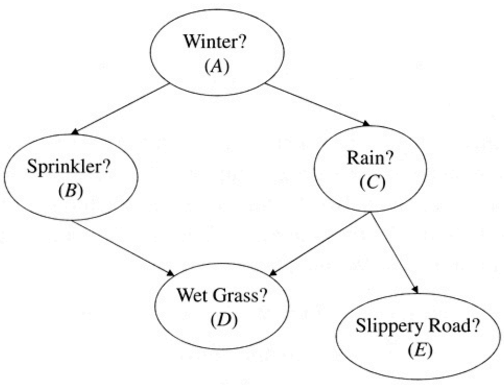
- $\Theta_{X|\mathbf{U}}$ denotes the CPT for variable X and parents U
- $\theta_{x|\mathbf{u}}$ denotes the cond. prob. $Pr(x|\mathbf{u})$ (network parameter)
  - must hold: $\sum_x \theta_{x|\mathbf{u}} = 1$
  - compatible with a network instantiation $\mathbf{z}$ if they agree on all values assigned to common variable: $\theta_{x|\mathbf{u}} \sim \mathbf{z}$

Properties:

- the network structure and parametrization of a network instantiation are satisfied by *one and only one* prob. distribution given by the chain rule for Bayesian networks:
  - product of all parameters compatible with $\mathbf{z}$

$$Pr(\mathbf{z}) = \prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}}$$

Example:



$$Pr(a, b, \bar{c}, d, \bar{e}) = \theta_a \theta_{b|a} \theta_{\bar{c}|a} \theta_{d|b,\bar{c}} \theta_{\bar{e}|\bar{c}} = (.6)(.2)(.2)(.9)(1) = .0216$$

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

$$Pr(\bar{a}, \bar{b}, \bar{c}, \bar{d}, \bar{e}) = \theta_{\bar{a}} \theta_{\bar{b}|\bar{a}} \theta_{\bar{c}|\bar{a}} \theta_{\bar{d}|\bar{b},\bar{c}} \theta_{\bar{e}|\bar{c}} = (.4)(.25)(.9)(1)(1) = .09$$

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

Probabilistic independence

distribution *Pr* specified by a Bayesian network satisfies the indep. assumptions

$$I(V, Parents(V), NonDescendants(V))$$

$$Markov(G)$$

...plus some others that implicitly *follow* from the above ones!!

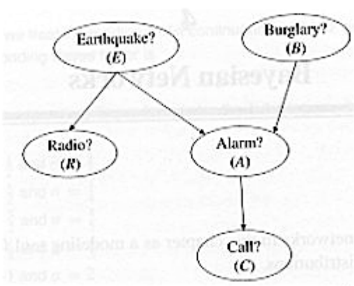
- e.g. in the previous example: $I_{Pr}(D, \{A, C\}, E)$

This is due to some properties of prob. independence known as graphoid axioms:

- symmetry
- decomposition
- weak union
- contraction

Graphoid axioms

- symmetry $I_{Pr}(X, Z, Y) \quad iff \quad I_{Pr}(Y, Z, X)$
  - if learning y doesn't change belief in x, then learning x doesn't change belief in y
  - Example:



$$I_{Pr}(A, \{B, E\}, R)$$
$$\rightarrow I_{Pr}(R, \{B, E\}, A)$$

Graphoid axioms

- decomposition

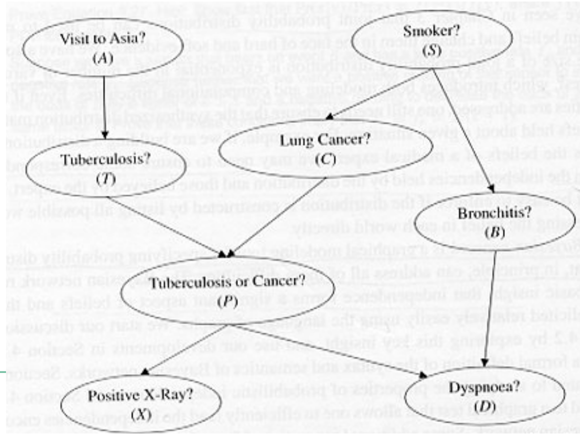
$$I_{Pr}(X, Z, Y \cup W) \text{ only if } I_{Pr}(X, Z, Y) \text{ and } I_{Pr}(X, Z, W)$$
$$I_{Pr}(X, Parents(X), W) \text{ for every } W \subseteq NonDescendants(X)$$

- every variable X is indep. of any subset of its descendants given its parents
- any part of irrelevant information is irrelevant too

- *Example*:

$$I(B, S, \{A, C, P, T, X\}) \rightarrow I(B, S, C)$$

once knowing *smoker,* belief in *bronchitis* no longer influenced by info about *cancer*



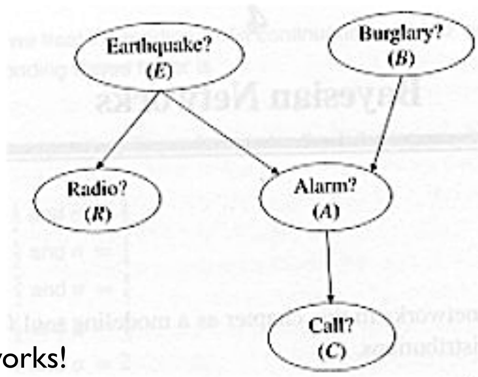
Graphoid axioms

- decomposition (cont'd)
  - allows to prove chain rule for Bayesian networks, given an appropriate „bottom-up" ordering of variables
  - implies a simple method to calculate degree of belief in an event:

Example:

$$Pr(c, a, r, b, e)$$
$$= Pr(c|a, r, b, e)Pr(a|r, b, e)Pr(r|b, e)Pr(b|e)Pr(e)$$
(chain rule of prob. calculus)
$$= Pr(c|a)Pr(a|b, e)Pr(r|e)Pr(b)Pr(e)$$
(decomp. / independencies)
$$= \theta_{c|a}\theta_{a|b,e}\theta_{r|e}\theta_b\theta_e$$

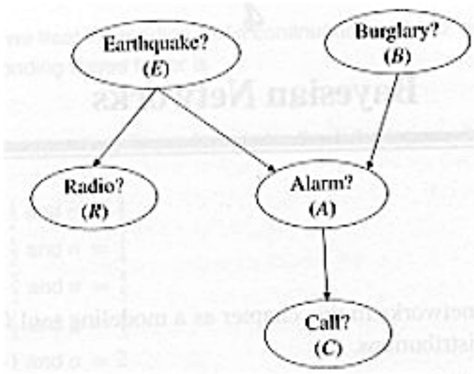
equals results given by chain rule of Bayesian networks!



Graphoid axioms

- weak union $I_{Pr}(X, Z, Y \cup W)$ onlfy if $I_{Pr}(X, Z \cup Y, W)$
  - if info *yw* is not relevant to our belief in *x*, then the partial info *y* will not make the rest of the info *w* relevant
  - *Example*:

$$I(C, A, \{B, E, R\})$$
$$\rightarrow I(C, \{A, E, B\}, R)$$



Graphoid axioms

- contraction

$$I_{Pr}(X, Z, Y) \text{ and } I_{Pr}(X, Z \cup Y, W) \text{ only if } I_{Pr}(X, Z, Y \cup W)$$

  - if after learning irrelevant info *y* the info *w* is found to be irrelevant to belief in x, then combined info *yw* must have been irrelevant from beginning

- [intersection]

$$I_{Pr}(X, Z \cup W, Y) \text{ and } I_{Pr}(X, Z \cup Y, W) \text{ only if } I_{Pr}(X, Z, Y \cup W)$$

  - if info *w* is irrelevant given *y* and info *y* is irrelevant given *w*, then the combined info *yw* is irrelevant to start with
  - holds only for strictly positive prob. distributions (assign non-zero prob. to every consistent event)

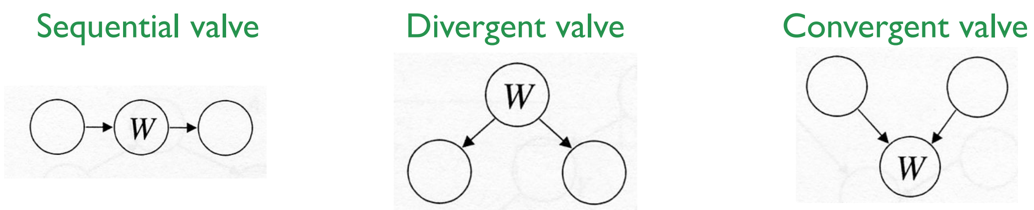
Graphical test of independence

Bayesian network induces a belief state/prob distribution *Pr*

All independencies in *Pr* (implied by Graphoid axioms) can be derived efficiently using a graphical test called d-separation

<u>*Idea:*</u> there are three types of causal structures („valves") in a graph

- a valve can be either open or closed
- closed valves block a path in the graph, implying independence



W *intermediary* between cause and effect

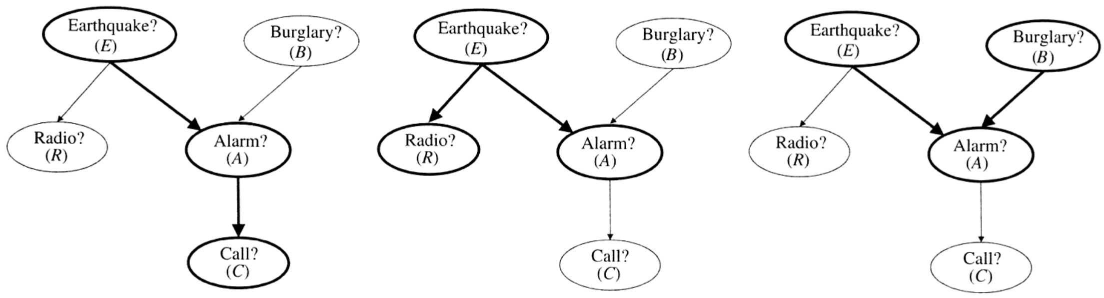
W *common cause* of two effects

W *common effect* of two causes

Graphical test of independence

Given a set of variables ***Z***, a valve with variable *W* is closed iff...

- sequential valve: *W* appears in ***Z***
  - Example: E -> A -> C closed if A given, E and C become cond. indep.
- divergent valve: *W* appears in ***Z***
  - Example: R <- E -> A closed if E given, R and A become cond. indep.
- convergent valve: neither *W* nor any of its descendants appears in ***Z***
  - Example: E -> A <- B closed if neither A nor C given



d-separation

Definition:

Variable sets ***X*** and ***Y*** are d-separated by ***Z*** iff every path between a node in ***X*** and a node in ***Y*** is blocked by ***Z*** (at least one valve on the path is closed given ***Z***).

$$dsep_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$$

Theorem:

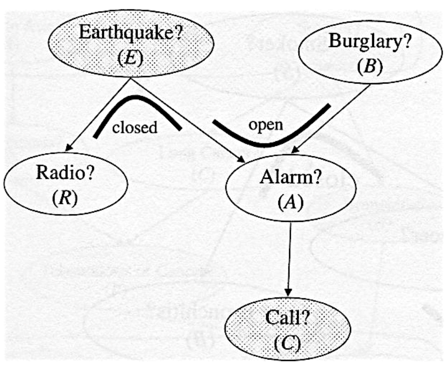
For every network graph G there is a parametrization Θ such that

$$I_{Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \leftrightarrow dsep_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$$

- *dsep* is correct (sound)
- *dsep* is complete for a suitable parametrization (but not for every!)

d-separation

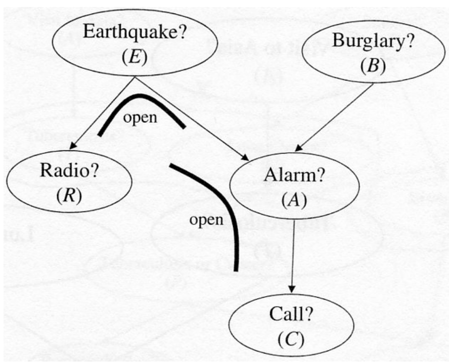
Examples:



Two valves between *R* and *B*, first valve (divergent) is closed given *E*

➔ *R* and *B* are d-separated by *E*

➔ *R* and *B* are cond. indep. given *E*

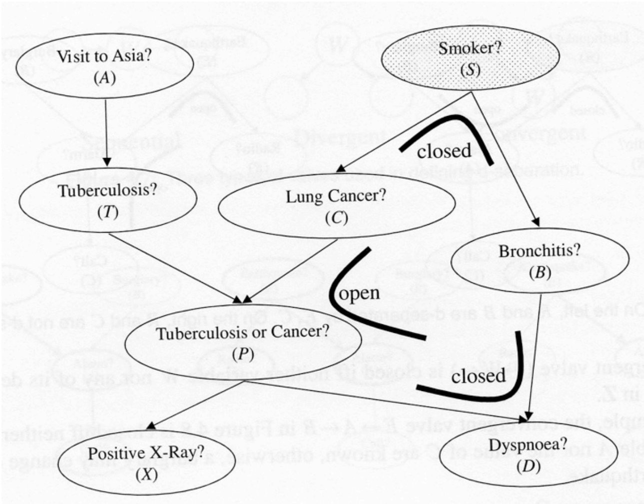


Two valves between R and C, both are open

➔ *R* and *C* are not d-separated

d-separation

Examples:



Are *B* and *C* d-separeted by S?

Two paths:

- 1st one closed valve (C<-S->B) because S given
- 2nd one closed valve (B->D<-P) because D not given

➔ *B* and *C* are d-separated by *S*

➔ *B* and *C* are cond. indep. given *S*

Next week(s)

- How to build a Bayesian network?
- How to use it for inferencing?
- Inference algorithms
  - exact
  - approximative