Reasoning and Decision-Making under Uncertainty

5. Session: Bayesian (Belief) Networks

Prof. Dr.-Ing. Stefan Kopp
Center of Excellence „Cognitive Interaction Technology“
AG Sociable Agents

CITEC

Sociable Agents

Independence

A given state of beliefs finds a belief independent of another belief **iff**

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

- Equivalent definition: $Pr(\alpha \wedge \beta) = Pr(\alpha) \cdot Pr(\beta)$

State of belief finds α conditionally independent of β given event γ **iff**

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

Allows to simplify the factorization of joint distribution (with chain rule):

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

→ For example: $Pr(\alpha_1 \wedge \ldots \wedge \alpha_n) = Pr(\alpha_1|\alpha_3 \wedge \alpha_8)Pr(\alpha_2|\ldots)\ldots$

Bayesian (belief) networks

Basic idea:

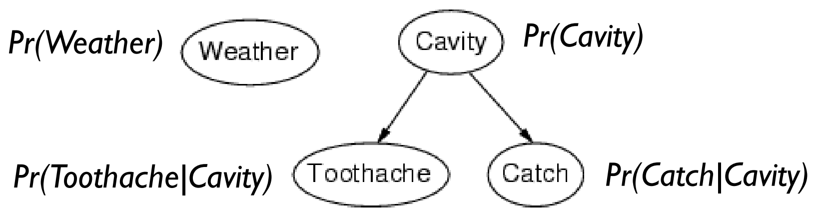
- rely on insight that independence forms a significant aspect of beliefs
- helps to form a compact representation of full belief state
- describe in terms of graphical probabilistic model

Definition: *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: $Pr(X_i \mid Parents(X_i))$, or $Pr(X_i)$ if no parents
  - parameterizes the independence structure

Bayesian (belief) networks

Example:



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

Bayesian (belief) networks

Formally: A Bayesian network defines a set of cond. indep. statements:

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

- Each variable is conditionally independent of its non-descendants, given its parents
  - Markovian assumption: $Markov(G)$
- $Parents(V)$ are direct causes, $Descendants(V)$ are effects of V
- Given the direct causes (parents) of *V*, the belief in *V* is no longer influenced by any other variable, except possibly by the effects (direct and indirect children) if *V*

Bayesian (belief) networks

Notation:

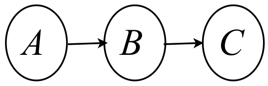
- $\Theta_{X|\mathbf{U}}$ denotes the CPT for variable X and parents U
- $\theta_{x|\mathbf{u}}$ denotes the cond. prob. $Pr(x|\mathbf{u})$ (network parameter)
  - must hold: $\sum_x \theta_{x|\mathbf{u}} = 1$
  - compatible with a network instantiation $\mathbf{z}$ if they agree on all values assigned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

As mathematical model of joint distribution:

- 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**:

$$Pr(\mathbf{z}) = \prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}} = \prod_{Pr(x|\mathbf{u}) \sim \mathbf{z}} Pr(x|\mathbf{u}), \text{ with } \mathbf{u} \text{ parents of } x$$

Example:



$$Pr(a,b,c) \;= Pr(c|b,a)Pr(b|a)Pr(a) \;= Pr(c|b)Pr(b|a)Pr(a)$$

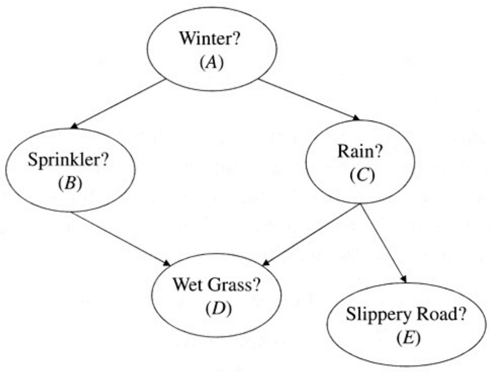
requires 8 rows
(exponential)

requires 2 rows
(+ normalization)

a	b	c	Pr(.)
T	T	T	…
T	T	F	…
T	F	T	…
T	F	F	…
F	T	T	…
F	T	F	…
F	F	T	…
F	F	F	…

a	b	c	Pr(a)	Pr(b|a)	Pr(c|b)
T	T	T	…	…	…
F	F	T	…	…	…

Example:



$Pr(a,b,\bar{c},d,\bar{e}) =$

=

$Pr(\bar{a},\bar{b},\bar{c},\bar{d},\bar{e})$

=

=

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

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*

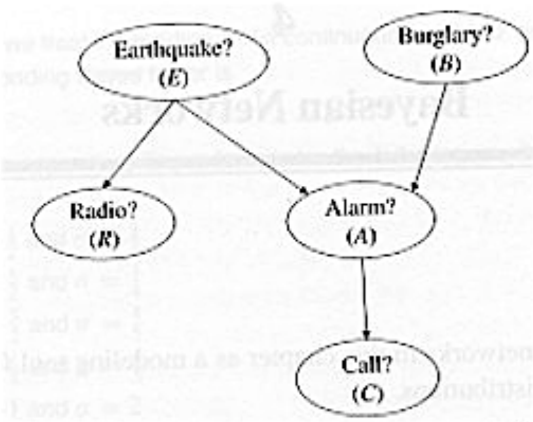
Network topology reflects believed causal structure of the domain:

- burglar and earthquake can set the alarm off
- alarm can cause John to call
- earthquake can cause a radio report

- + independence assumptions?

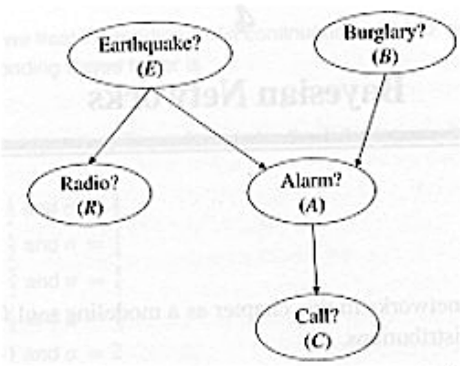
Example:

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



- 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*
- given no descendant, *Earthquake* and *Burglary* are indep.

Example:



(chain rule of prob. calculus / repeated Bayesian cond.) $$Pr(c,a,r,b,e)$$

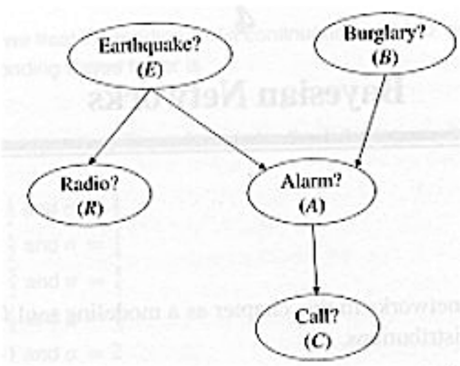
$$= Pr(c|a,r,b,e)Pr(a|r,b,e)Pr(r|b,e)Pr(b|e)Pr(e)$$

(decomposition / independence assumptions encoded in the graph)

$$= Pr(c|a)Pr(a|b,e)Pr(r|e)Pr(b)Pr(e)$$

$$= \theta_{c|a}\theta_{a|b,e}\theta_{r|e}\theta_b\theta_e$$

Example:



$$Pr(e|a) = ?$$

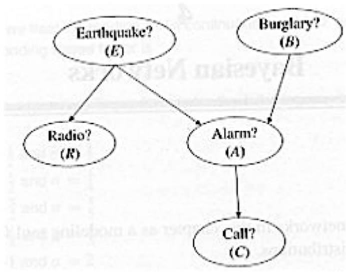
(Bayesian conditioning & case analysis)

$$= \frac{Pr(e,a)}{Pr(a)} = \frac{Pr(a|e)Pr(e)}{Pr(a)} = \frac{Pr(a,e,b) + Pr(a,e,\bar{b})}{Pr(a)}$$

$$= \frac{Pr(a,e,b) + Pr(a,e,\bar{b})}{Pr(a|e,b) + Pr(a|e,\bar{b}) + Pr(a|\bar{e},b) + Pr(a|\bar{b},\bar{e})}$$

$$= \frac{0.072}{0.2442} = 0.307$$

Bayesian networks



Besides the mathematical meaning of the network (cf. chain rule), independence relations can be derived from the graph topology

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

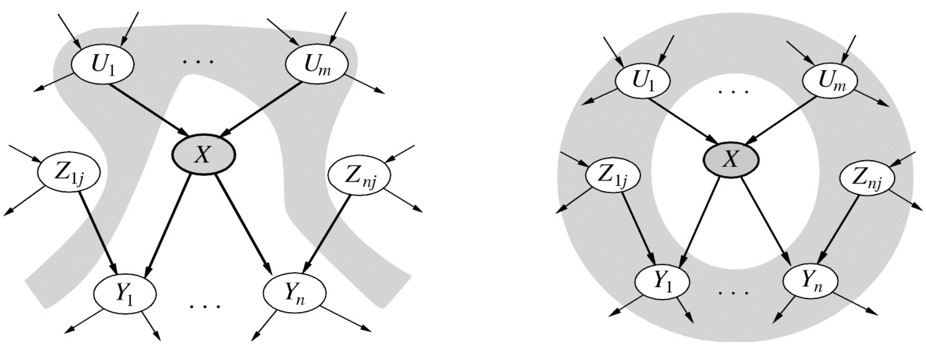
$I(V, Parents(V), NonDescendants(V))$

- $Parents(V)$ are direct causes, $Descendants(V)$ are effects of V
- Each node is conditionally indep. of its non-descendants given its parents
  - given full info about the direct causes of *V*, degree of belief in *V* is no longer influenced by information about any other variable, except from its effects

Bayesian networks

Two equivalent topological specifications:

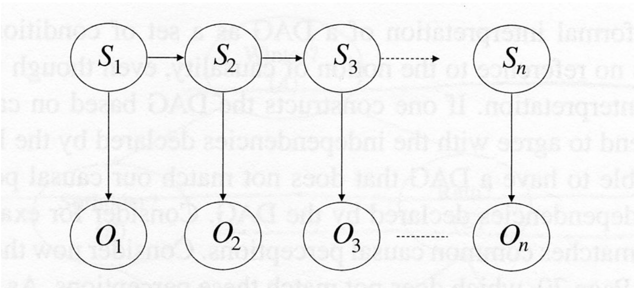
- *X* cond. indep. of all non-descendants given all parents
- *X* cond. indep. of all other nodes given its Markov blanket (= parents + children + children's parents)



(a) (b)

HMM as special case of Bayesian networks

Example: Hidden Markov Models ➔ Dynamic BNs



- 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

Probabilistic independence

Distribution *Pr* specified by a Bayesian network satisfies the independence assumptions

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

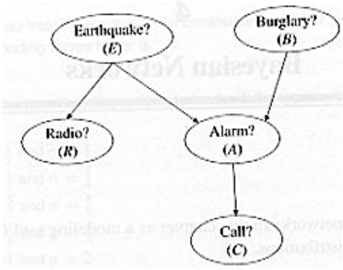
Plus some more that follow implicitly from the above ones!

This is due to some properties known as graphoid axioms:

- symmetry
- decomposition
- weak union
- contraction

Graphoid axioms I

- 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 II

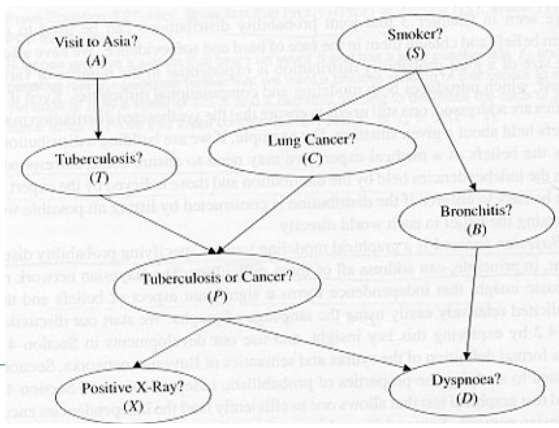
- decomposition

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

  - every variable X is indep. of any subset of its non-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*



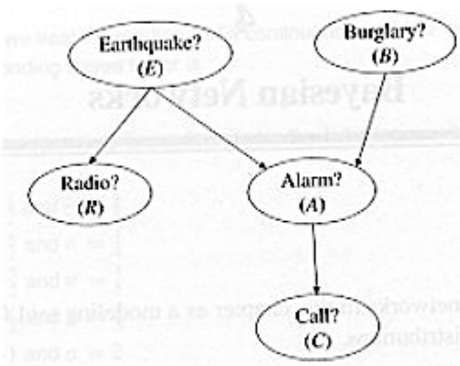
Graphoid axioms III

- 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* given *z*, 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 IV

- contraction

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

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

- [intersection]

  $I_{Pr}(X, Z \cup W, Y)$ and $I_{Pr}(X, Z \cup Y, W)$ 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 that assign non-zero prob. to every consistent event

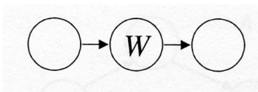
D-seperation

All independencies in *Pr* (implied by Graphoid axioms) can be derived from the graph structure, using a graphical test called d-separation

Idea: there are three types of causal structures („valves") in a graph

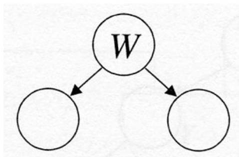
- a causal structure or valve can be either open or closed
- closed valves block a path in the graph, implying independence

Sequential, chain



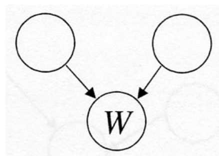
W intermediary betw. cause and effect

Divergent, fork



W common cause of two effects

Convergent, collider, inverted fork

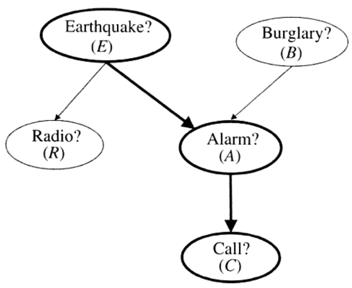


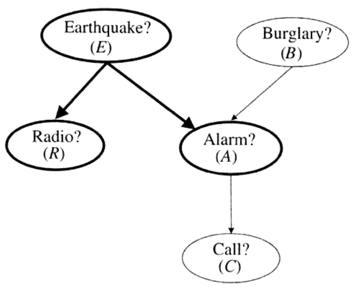
W common effect of two causes

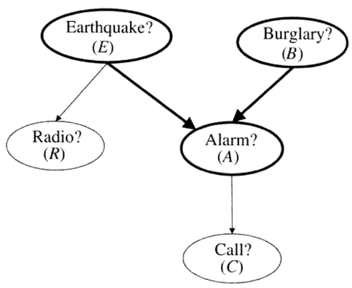
D-seperation

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

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







D-seperation

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 each 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

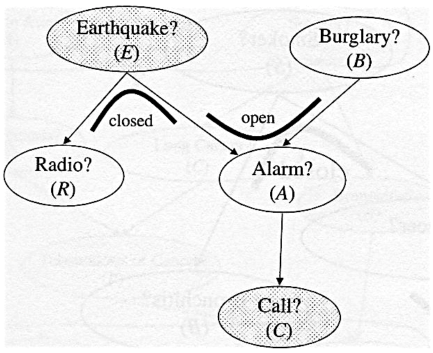
$$I_{Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \leftrightarrow dsep_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$$

It holds (see proofs in Darwiche) that

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

D-separation

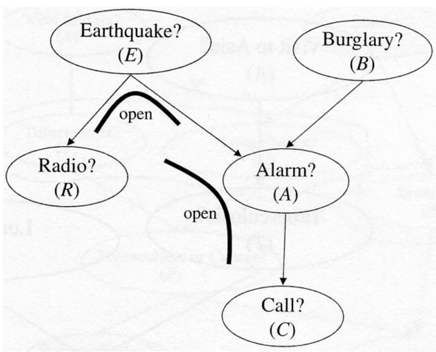
Examples:



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*



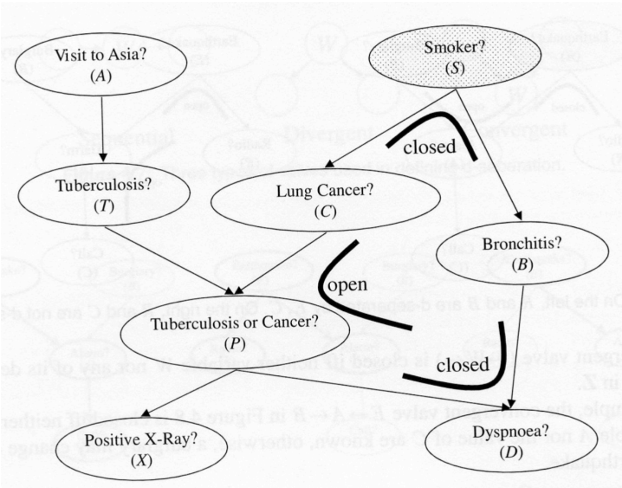
Two valves between R and C, both are open

➔ *R* and *C* are not d-separated

➔ learning about *C* changes degree of belief in *R* (and vice versa)

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*

Reasoning with Bayesian networks

Classical reasoning in (extensional) logics-based representations

- Modus Ponens: (a→b), a ➠ b
  - *probabilistic*: assuming a causal link between a and b, and given some evidence P(a) ➠ how does P(b) change to P(b|a)? (causal inference)
- Modus Tollens: (a→b), ¬b ➠ ¬a
  - *probabilistic*: assuming a causal link between a and b, and given some evidence P(¬b) ➠ how does P(a) change to P(a|¬b)? (causal)
- Abductive Reasoning: (a→b), b ➠ a
  - *probabilistic*: assuming a causal link between a and b, and given some evidence P(b) ➠ how does P(a) change to P(a|b)? (diagnostic)
- Inductive Reasoning: A(x1),A(x2),A(x3),A(x4),... ➠ ∀xA(x)
  - *probabilistic*: given some evidence P(A(xi))=1 for i=1..k ➠ what is P(A)?

Reasoning with Bayesian networks

Four general types of queries one can solve with Bayesian networks:

- **probability of evidence**:
  how likely is a complete variable instantiation **e** ➔ *Pr(***e***)=?*
- **prior and posterior marginals**: how probable is an instantiation of a *limited* set of variables ➔ $Pr(x_1,...,x_m)=?$ or $Pr(x_1,...,x_m|\mathbf{e})=?$
- **most probable explanation (MPE)**: what is the most probable instantiation of *all* n network var's given some evidence **e** ➔ **x** *with* $Pr(x_1,...,x_n|\mathbf{e})=max?$
- **maximum a posteriori hypothesis (MAP):** what is the most probable instantiation of a *subset* of m (m<n) var's given some evidence **e** ➔ **x** *with* $Pr(x_1,...,x_m|\mathbf{e})=max?$

➔ Let's try out with *Samlam*

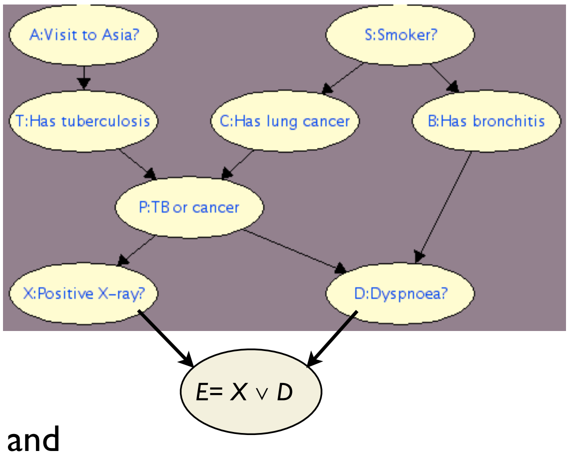
Probability of evidence

Query: How likely is some variable instantiation **e** ➔ *Pr(***e***)=?*

Example: *Pr(X=yes, D=no)=?*

Example: *Pr(X=yes ∨ D=yes)=?*

can be computed indirectly with the auxiliary-node technique:



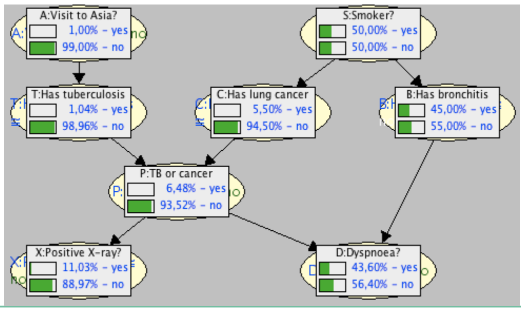
- add node *E* with *X,D* as parents and *Pr(e|x,d)=1 iff e=1 and (d=1 or x=1)*
- possible when not too many evidence var's

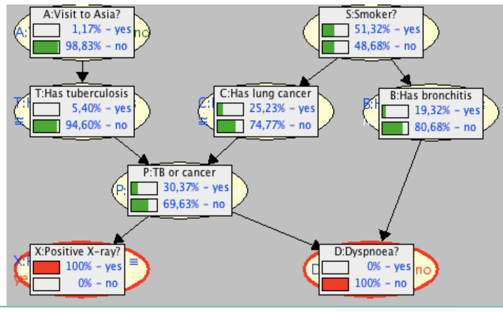
Prior and posterior marginals

Query: How probable is an instantiation of a limited set of variables
➔ *Pr(x1,...,xm)=? or Pr(x1,...,xm|***e***)=?*

Definition: Given a joint distribution *Pr(x1,...xn)* and a limited number *m* of variables,

- prior marginal : $$Pr(x_1, ..., x_m) = \sum_{x_{m+1},...x_n} Pr(x_1, ..., x_n)$$
- posterior marginal given **e** : $$Pr(x_1, ..., x_m|\mathrm{e}) = \sum_{x_{m+1},...x_n} Pr(x_1, ..., x_n|\mathrm{e})$$

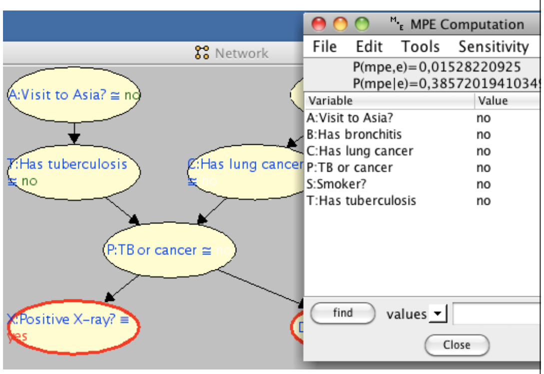




Most probable explanation (MPE)

Query: What is the most probable instantiation of all network var‘s given some evidence **e** ➔ ***x*** *with Pr(x1,...,xn|***e***)=max?*

Example: MPE for positive x-ray and not dyspnoea?



Cannot be computed directly from the maximal posterior marginals

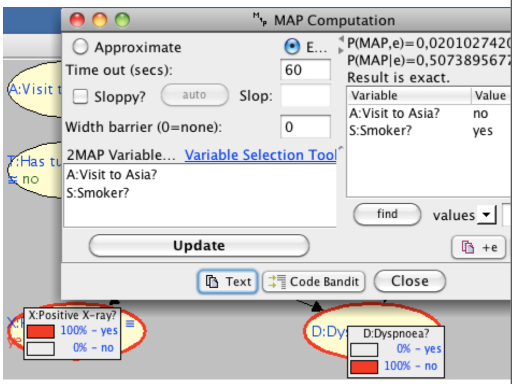
- choosing x_i such that *Pr(*x_i*|***e***)=max* yields expl. *p* with *smoker=true and Pr(p|e)=20.03%* whereas *Pr(mpe|e)=38.57%*

Maximum a posteriori hypothesis (MAP)

Query: What is the most probable instantiation of a subset of var's $\mathbf{M}=X_1,\dots,X_m$ given some evidence $\mathbf{e} \rightarrow \mathbf{m}$ *with* $Pr(\mathbf{m}|\mathbf{e})=max$?

- MPE is a special case of MAP, easier to compute algorithmically

Example: Given *X=yes*, *D=no*, what is the most probable instantiation of **M**={A,S}?



Approximative method to find MAP:

- compute MPE and return values for MAP variables (projecting MPE on MAP var's)
- but, leads to *A=no, S=yes* here with prob ~48%, while *A=no, S=no* is MAP with prob ~50%