Syllabus, Chapter 6:

Bringing Bayesian Networks into Practice

Inaccuracy versus robustness

Consider a BN $\mathcal{B} = (G, \Gamma)$. Assessments obtained (from data or human experts) for the model-parameters $\gamma_V \in \Gamma$ tend to be inaccurate or uncertain.

Robustness: pertains to stability of some output in terms of variation of model-parameter:

- output is robust if varying model-parameters reveals little effect on the output;
- if varying model-parameters shows a considerable effect, then the output is not robust and may be unreliable.

Inaccuracy, therefore, does not necessarily imply a lack of robustness.

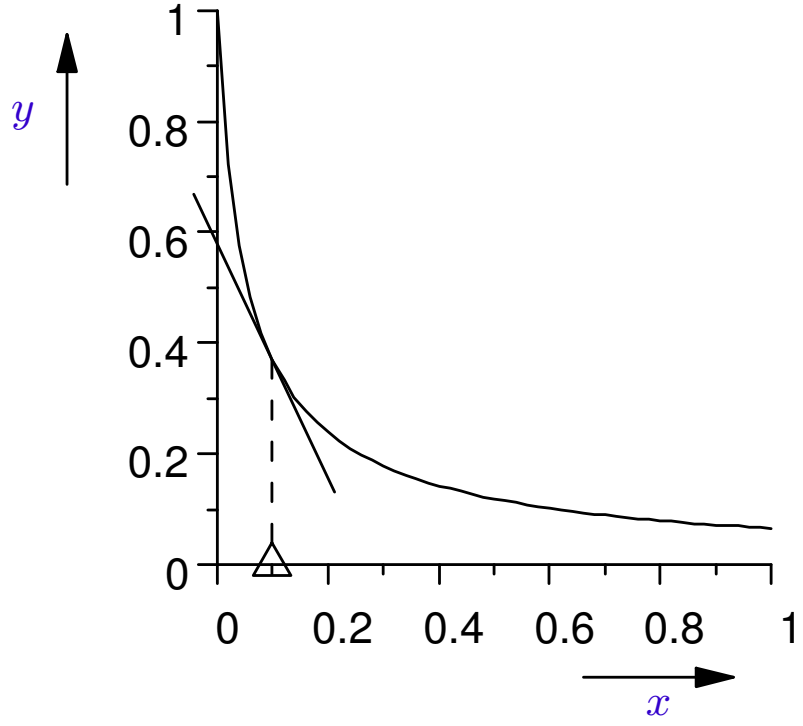
Analysing the robustness of a Bayesian network

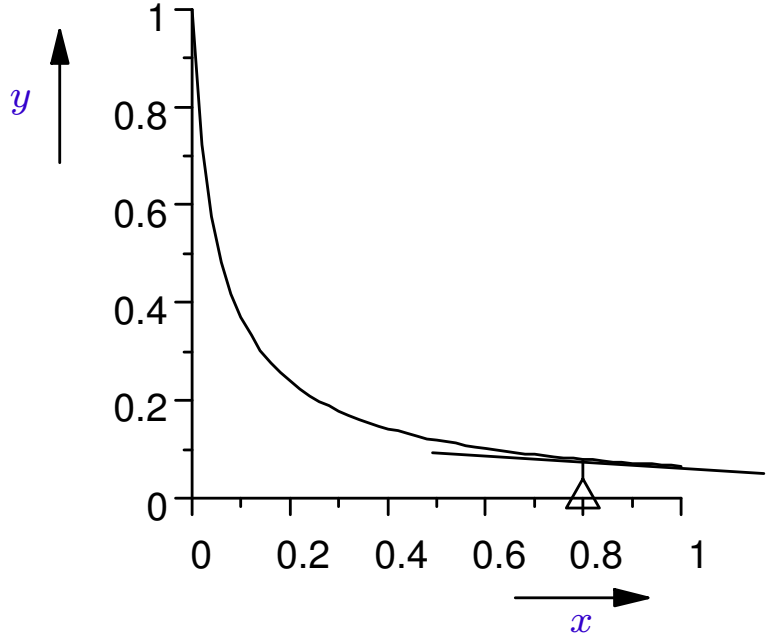
Various techniques are available for analysing the robustness of a Bayesian network.

- sensitivity analysis
  - systematically vary model-parameters and study the effect on the output;
  - in an n-way sensitivity analysis, n model-parameters are varied simultaneously;
- uncertainty analysis
  - repeatedly draw model-parameters from sample distributions and study the effect.

A one-way sensitivity analysis

A one-way sensitivity analysis for a network-parameter $x = \gamma(c_{V_i} \mid c_{\boldsymbol{\rho}(V_i)})$ results in a sensitivity curve, describing an output probability $y = \Pr(c_{V_o} \mid c_{\boldsymbol{E}})$ in terms of x:



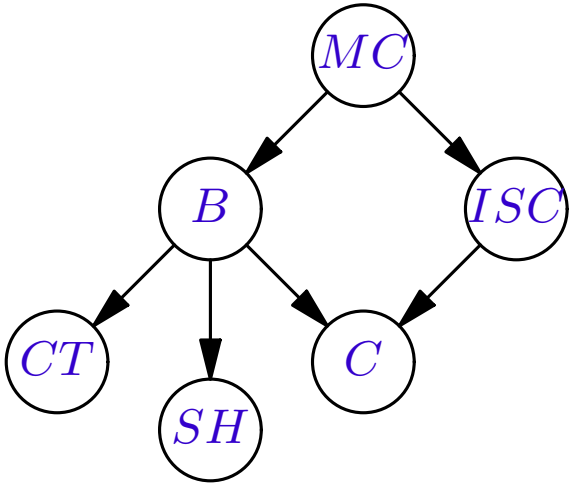


The effect of small variations in x on the output depends on the original assessment x_0 for network-parameter x.

The computational burden involved

Straightforward sensitivity analysis is highly time consuming:

- for the following network, a single analysis[8] requires 130 network propagations:



$$\gamma(b \mid mc) = 0.20 \quad \gamma(mc) = 0.20$$
$$\gamma(b \mid \neg mc) = 0.05$$
$$\gamma(c \mid b, isc) = 0.80$$
$$\gamma(sh \mid b) = 0.80 \quad \gamma(c \mid \neg b, isc) = 0.80$$
$$\gamma(sh \mid \neg b) = 0.60 \quad \gamma(c \mid b, \neg isc) = 0.80$$
$$\gamma(c \mid \neg b, \neg isc) = 0.05$$
$$\gamma(ct \mid b) = 0.95$$
$$\gamma(ct \mid \neg b) = 0.10 \quad \gamma(isc \mid mc) = 0.80$$
$$\gamma(isc \mid \neg mc) = 0.20$$

- for the medium-sized classical swine fever network, a single analysis requires approximately 20.000 network propagations.

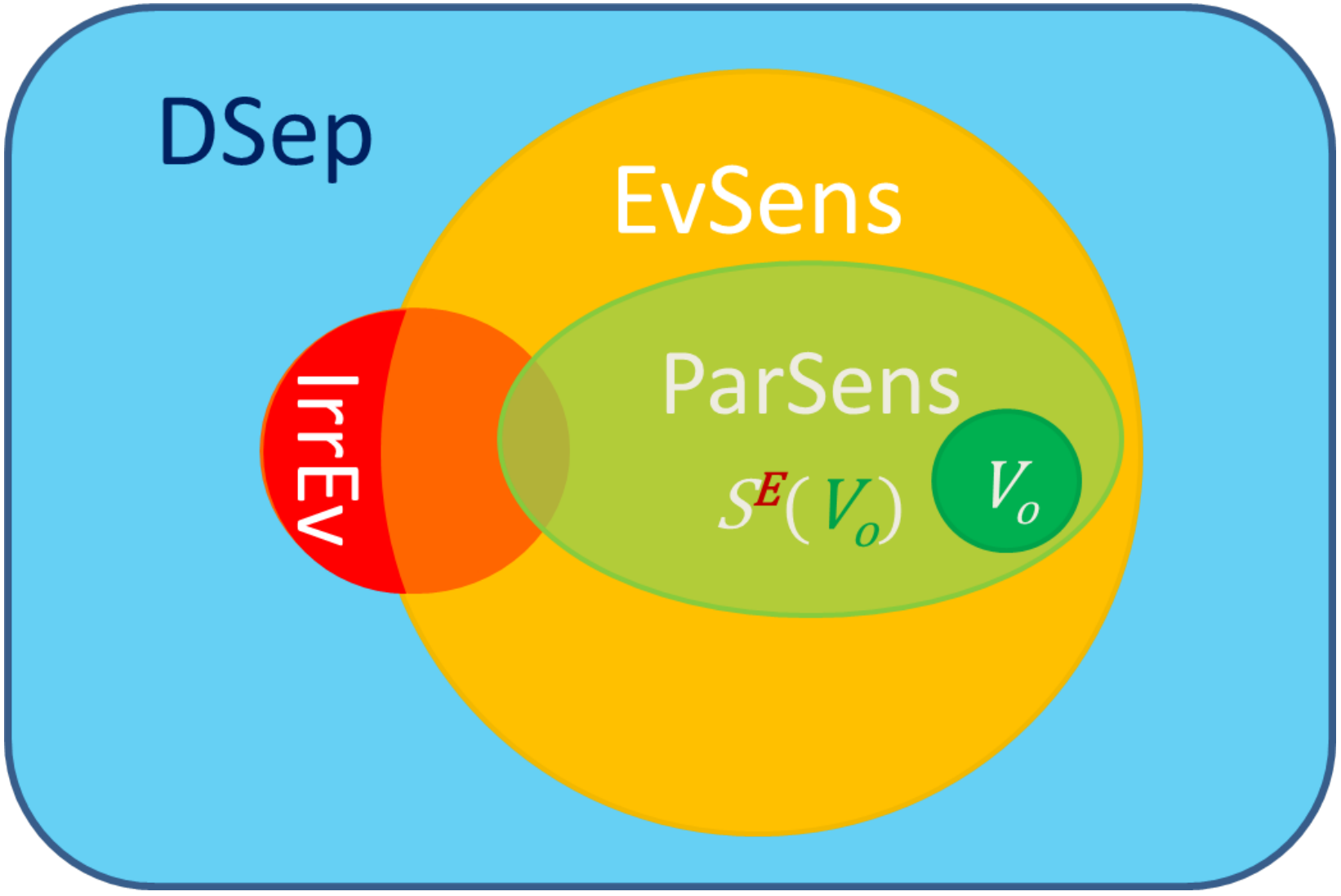
[8] assuming we compute 10 points per curve

Reducing the computational burden

The computational burden of a sensitivity analysis can be reduced by exploiting the following BN properties:

- various network-parameters cannot affect, upon variation, the output probability of the network;
- the output probability relates to any network-parameter under study as a quotient of two (multi-)linear functions.

(Un)influential parameters – an overview



(See Meekes, Renooij & van der Gaag: Relevance of evidence in Bayesian networks. (ECSQARU 2015))

Influential parameters – the basics

Consider $\mathcal{B} = (G, \Gamma)$ with output variable of interest $V_o \in \boldsymbol{V}_G$ and evidence for the set $\boldsymbol{E} \subseteq \boldsymbol{V}_G$.

Let $S^{\boldsymbol{E}}(V_o) \subseteq \boldsymbol{V}_G$ denote the set of variables whose assessments may affect, upon variation, the output distribution of interest $\Pr^{\boldsymbol{e}}(V_o)$.

Which $V_i \in \boldsymbol{V}_G$ belong to $S^{\boldsymbol{E}}(V_o)$?

Basically: each V_i for which a change in one of its network-parameters $\gamma(c_{V_i} \mid c_{\boldsymbol{\rho}(V_i)})$ will eventually result in a change in the messages computed for/at V_o upon inference.

$S^{\boldsymbol{E}}(V_o)$ is called the sensitivity set for V_o under evidence for $\boldsymbol{E}$.

(Un)influential parameters – introduction

Let $U^{\boldsymbol{E}}(V_o) = \boldsymbol{V}_G \setminus S^{\boldsymbol{E}}(V_o)$ capture the variables for which a change in an assessment will certainly not affect $\Pr^{\boldsymbol{e}}(V_o)$, i.e. the uninfluential ones.

- Suppose $\boldsymbol{E} = \emptyset$.
  Which $V_i \in \boldsymbol{V}_G$ belong to $S^{\emptyset}(V_o)$ and $U^{\emptyset}(V_o)$?

- Suppose $\boldsymbol{E} \neq \emptyset$. How can $V_i \in S^{\emptyset}(V_o)$ become uninfluential?

answers: see slide 326

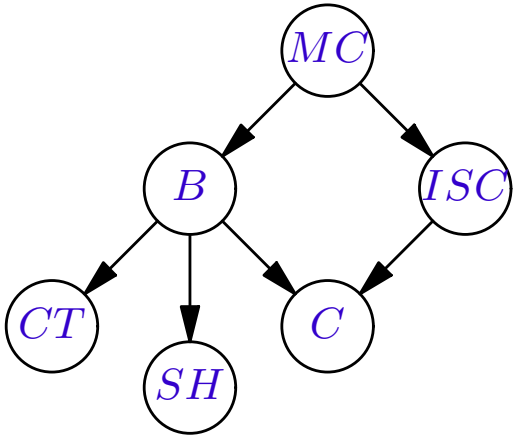
Uninfluential parameters: ancestors

The network-parameters for any variable V_i with

$$V_i \in \boldsymbol{\rho}^*(V_o) \text{ and } \langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid E \mid \{V_o\} \rangle^d$$

are uninfluential.

Example:



- Can assessments for *MC* or *B* affect the output probability $\Pr(sh \mid \neg b)$?
- Can assessments for *B* affect the output probability $\Pr(c \mid \neg b)$?

■

answers: 1) no; 2) possibly

(Un)influential parameters – introduction cntd

- Suppose $\boldsymbol{E} = \emptyset$. Then
  $S^{\emptyset}(V_o) = \boldsymbol{\rho}^*(V_o)$ and $U^{\emptyset}(V_o) = \{V_i \mid V_i \notin \boldsymbol{\rho}^*(V_o)\}$
- Suppose $\boldsymbol{E} \neq \emptyset$. Then
  $S^{\emptyset}(V_o) \cap U^{\boldsymbol{E}}(V_o) =$
  $\{V_i \mid V_i \in \boldsymbol{\rho}^*(V_o) \wedge \langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid \boldsymbol{E} \mid \{V_o\} \rangle^d\}$
- Suppose $\boldsymbol{E} \neq \emptyset$. Which $V_i \in U^{\emptyset}(V_o)$ remain uninfluential?

answer: see slide 328

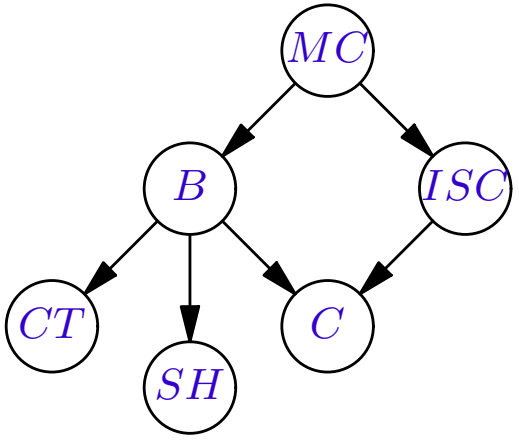
Uninfluential parameters: non-ancestors without evidence for descendants

The network-parameters for any variable V_i with

$$V_i \notin \boldsymbol{\rho}^*(V_o) \text{ and } \boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} = \emptyset$$

are uninfluential.

Example:



- Can assessments for *SH* or *CT* affect the output probability $\Pr(c \mid \neg isc)$?
- Can assessments for *SH* affect the output probability $\Pr(c \mid sh)$?

■

answers: 1) no; 2) possibly

(Un)influential parameters – introduction cntd

- Suppose $\boldsymbol{E} = \emptyset$. Then
  $S^{\emptyset}(V_o) = \boldsymbol{\rho}^*(V_o)$ and $U^{\emptyset}(V_o) = \{V_i \mid V_i \notin \boldsymbol{\rho}^*(V_o)\}$
- Suppose $\boldsymbol{E} \neq \emptyset$. Then
  $S^{\emptyset}(V_o) \cap U^{\boldsymbol{E}}(V_o) =$
  $\{V_i \mid V_i \in \boldsymbol{\rho}^*(V_o) \wedge \langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid \boldsymbol{E} \mid \{V_o\} \rangle^d\}$
- Suppose $\boldsymbol{E} \neq \emptyset$. Then
  $U^{\emptyset}(V_o) \cap U^{\boldsymbol{E}}(V_o) \supseteq \{V_i \mid V_i \notin \boldsymbol{\rho}^*(V_o) \wedge \boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} = \emptyset\}$

- Suppose $\boldsymbol{E} \cap \boldsymbol{\sigma}^*(V_i) \neq \emptyset$. Which V_i remain in $U^{\emptyset}(V_o) \cap U^{\boldsymbol{E}}(V_o)$?

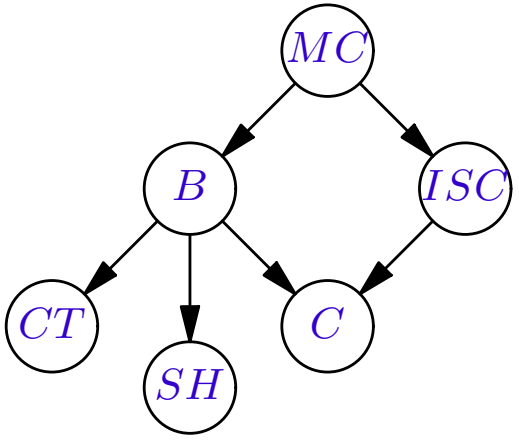
Uninfluential parameters: non-ancestors with evidence for descendants

The network-parameters for any variable V_i with

$$V_i \notin \boldsymbol{\rho}^*(V_o),\ \langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid \boldsymbol{E} \mid \{V_o\} \rangle^d \text{ and } \boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} \neq \emptyset$$

are uninfluential.

Example:



- Can assessments for *B* affect the output probability $\Pr(isc \mid \neg\, ct)$?
- Can assessments for *B* affect the output $\Pr(isc \mid mc \wedge \neg\, ct)$?

■

answers: 1) possibly; 2) no

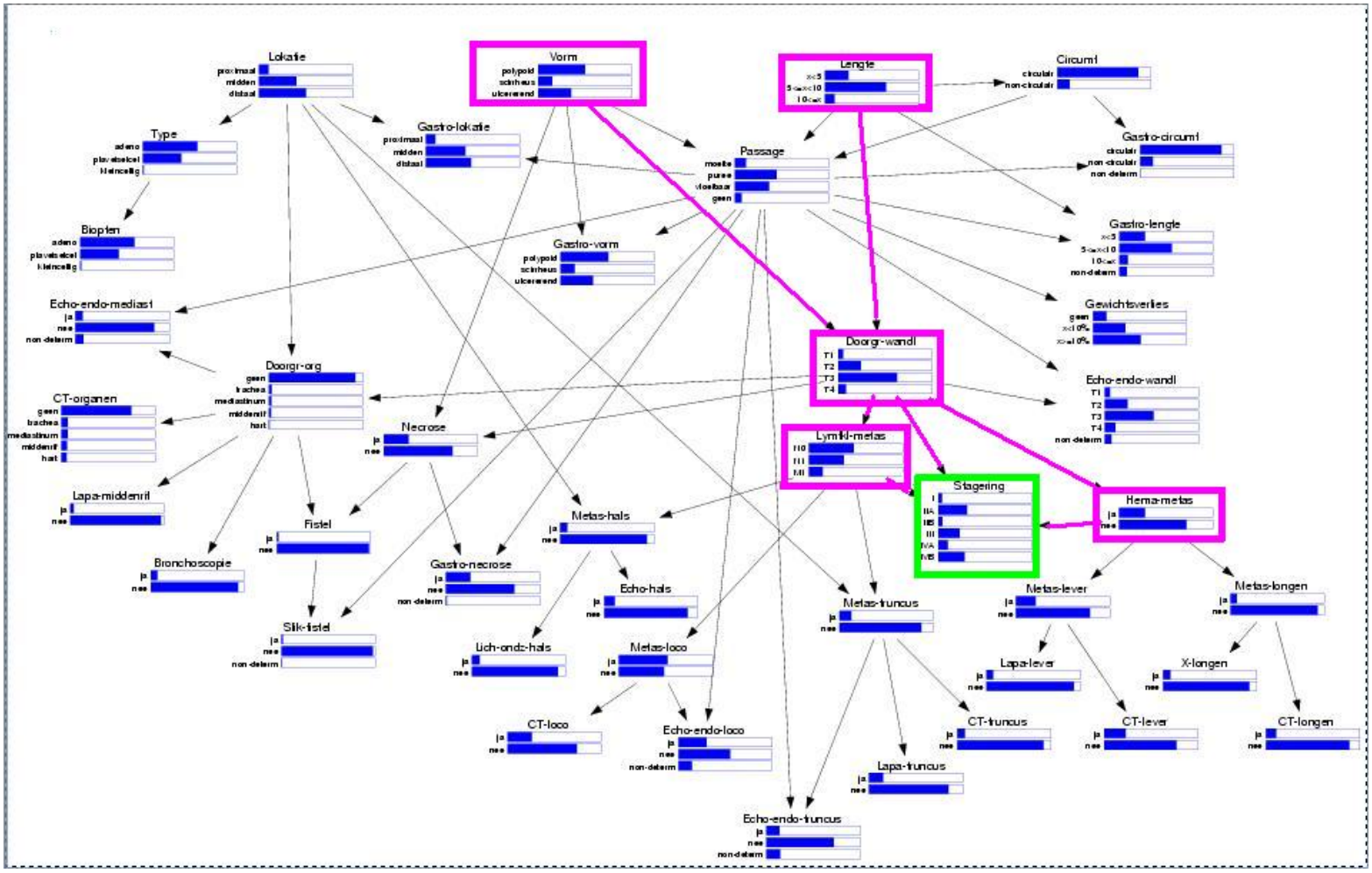
The sensitivity set – definition

The sensitivity set $S^{\boldsymbol{E}}(V_o)$ is the set of variables V_i for which none of the following holds:

- $V_i \in \boldsymbol{\rho}^*(V_o)$ and $\langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid \boldsymbol{E} \mid \{V_o\} \rangle^d$;
- $V_i \notin \boldsymbol{\rho}^*(V_o)$ and $\boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} = \emptyset$;
- $V_i \notin \boldsymbol{\rho}^*(V_o)$, $\langle \{V_i\} \cup \boldsymbol{\rho}(V_i) \mid \boldsymbol{E} \mid \{V_o\} \rangle^d$ and $\boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} \neq \emptyset$;

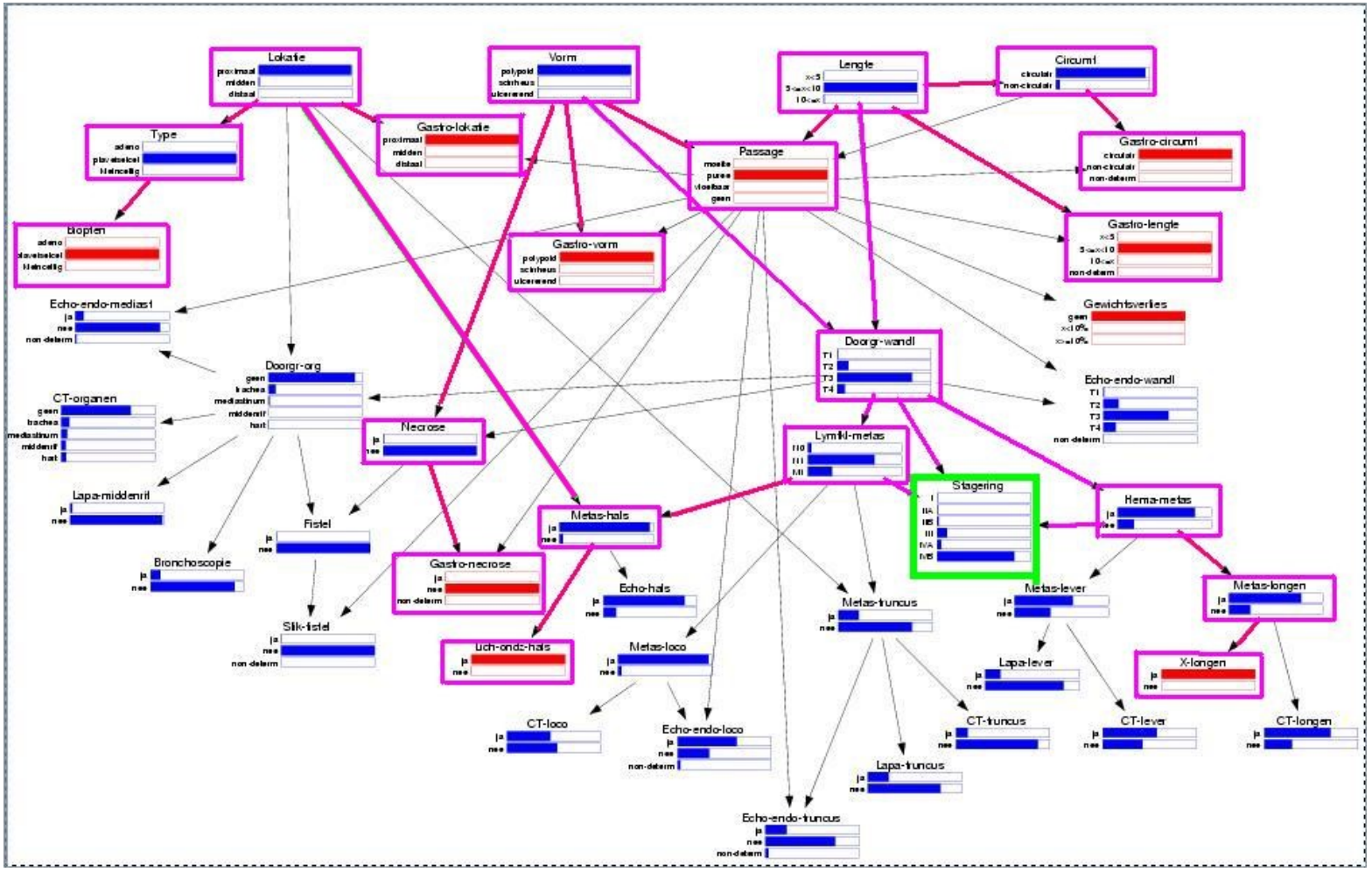
Only the network-parameters for the variables in the sensitivity set may affect, upon variation, the network's output probability.

Example: the prior sensitivity set for variable $Stage$



The sensitivity set $S^{\emptyset}(Stage)$ in the prior network consists of 6 variables, together specifying 206 model-parameters.

Example: a posterior sensitivity set for variable $Stage$



The sensitivity set $S^{\boldsymbol{E}}(Stage)$ in this posterior network consists of 21 variables, together specifying 527 model-parameters.

Computing the sensitivity set (I)

The sensitivity set $S^{\boldsymbol{E}}(V_o)$ is identified as follows:

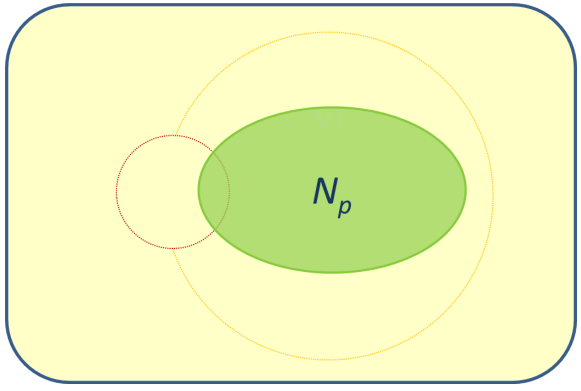
- construct, from the network's digraph G, a new digraph G^* by adding an auxiliary parent X_i to every $V_i \in \boldsymbol{V}_G$;
- determine all nodes V_i for which $\neg \langle \{X_i\} \mid \boldsymbol{E} \mid \{V_o\} \rangle^d_{G^*}$; these constitute the sensitivity set.

The sensitivity set can thus be identified in polynomial time ($O(|\boldsymbol{A}_{G^*}|)$) from just graphical considerations.

Computing the sensitivity set (II)

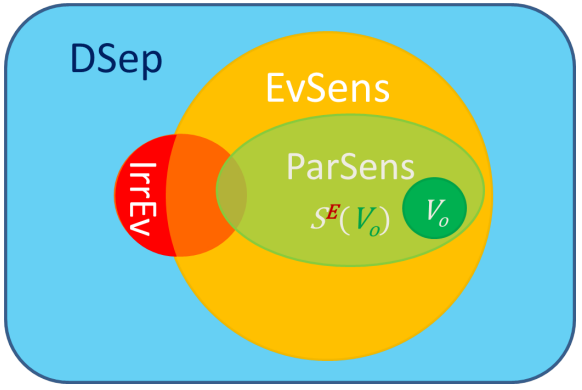
An alternative to identifying the sensitivity set $S^{\boldsymbol{E}}(V_o)$ is to use Bayes-Ball (BB) output (see Shachter, UAI 1998 for details):

BB terminology:
top mark, $N_p(V_o, \boldsymbol{E})$,
'Requisite $p()$*'*



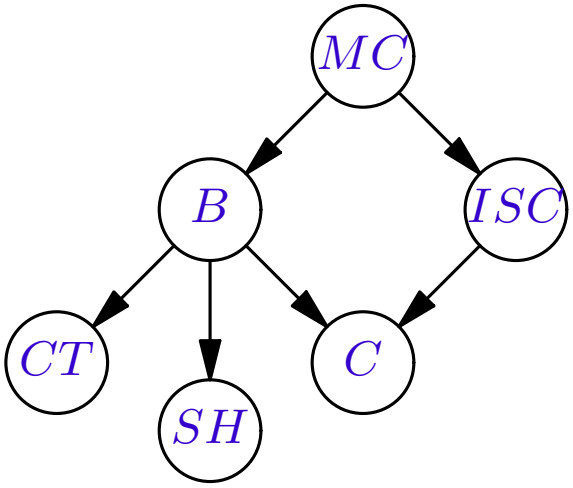
$S^{\boldsymbol{E}}(V_o) = N_p$

BB can also output
'Requisite e' ($\boldsymbol{E}\backslash$IrrEv) and
'Irrelevant' ($\boldsymbol{E}\cup$DSep)



The sensitivity set can be identified in $O(|\boldsymbol{V}_G| + |\boldsymbol{A}_G|)$ from just graphical considerations.

Computing an example sensitivity set

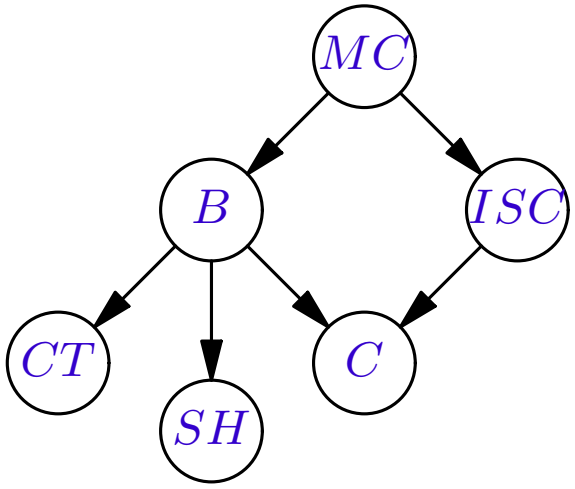


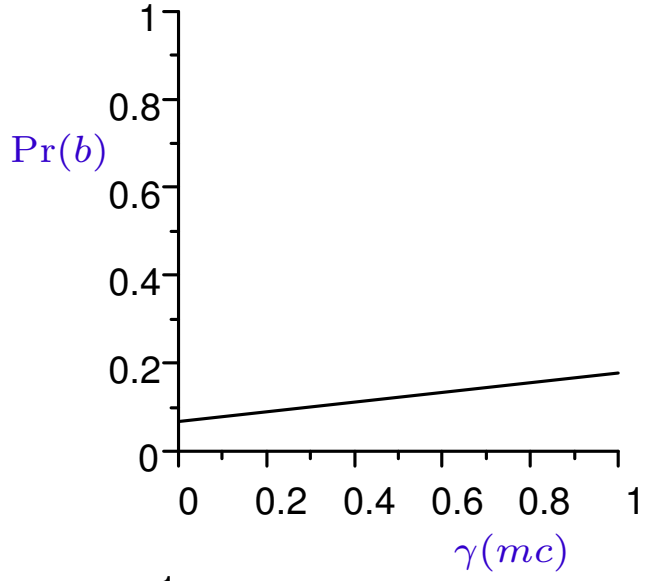
Assume that the graph is extended with auxiliary parents X_{CT}, X_{SH}, X_C, X_B, X_{ISC}, and X_{MC}.

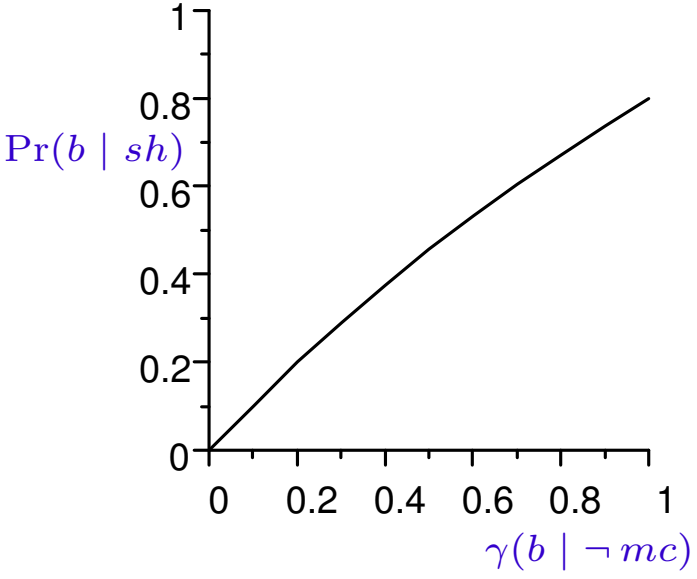
- the sensitivity set for *ISC* given *MC* and *CT* equals {*ISC*};
- the sensitivity set for *C* given *MC* and *CT* equals {*B*, *CT*, *C*, *ISC*}.

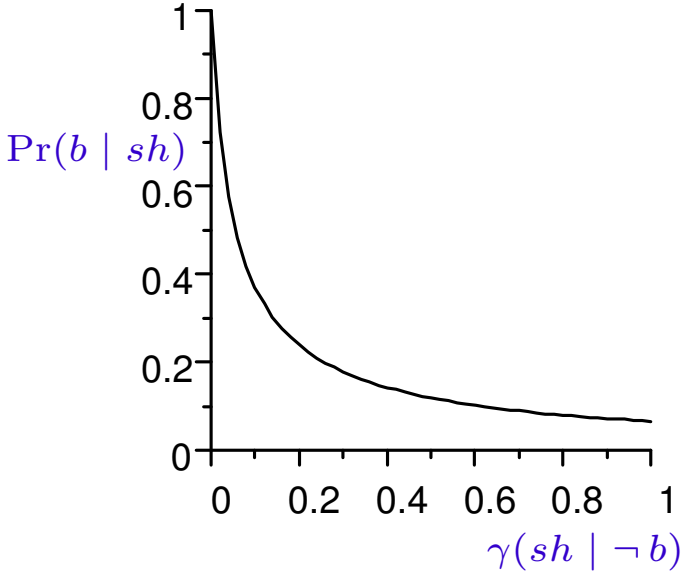
An introduction to the sensitivity function

In sensitivity analyses of Bayesian networks, any output probability is a function of the model-parameter under study:



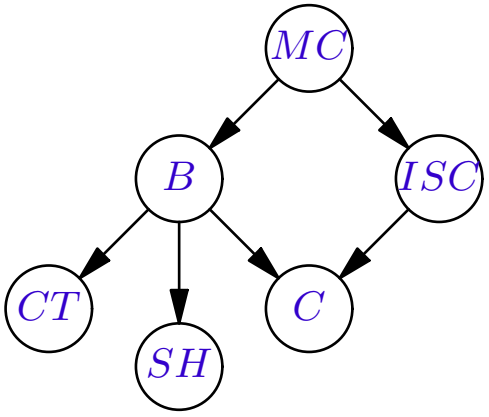






An example sensitivity function

A sensitivity function is strongly constrained by network $\mathcal{B}$. Consider the following Bayesian network:



$\gamma(b \mid mc) = 0.20$
$\gamma(b \mid \neg mc) = 0.05$

$\gamma(sh \mid b) = 0.80$
$\gamma(sh \mid \neg b) = 0.60$

$\gamma(ct \mid b) = 0.95$
$\gamma(ct \mid \neg b) = 0.10$

$\gamma(mc) = 0.20$

$\gamma(c \mid b, isc) = 0.80$
$\gamma(c \mid \neg b, isc) = 0.80$
$\gamma(c \mid b, \neg isc) = 0.80$
$\gamma(c \mid \neg b, \neg isc) = x$

$\gamma(isc \mid mc) = 0.80$
$\gamma(isc \mid \neg mc) = 0.20$

Output probability $\Pr(\neg mc \wedge \neg b \wedge \neg isc \wedge c)$, analytically expressed as a function of model-parameter $x = \gamma(c \mid \neg b \wedge \neg isc)$:

$$\Pr(\neg mc \wedge \neg b \wedge \neg isc \wedge c)(x) =$$
$$= \sum_{c_{CT}, c_{SH}} \Pr(\neg mc \wedge \neg b \wedge \neg isc \wedge c \wedge c_{CT} \wedge c_{SH})(x)$$
$$= \gamma(\overline{mc}) \cdot \gamma(\overline{b} \mid \overline{mc}) \cdot \gamma(\overline{isc} \mid \overline{mc}) \cdot \gamma(c \mid \overline{b} \wedge \overline{isc}) \cdot \sum_{c_{CT}} \gamma(c_{CT} \mid \overline{b}) \cdot \sum_{c_{SH}} \gamma(c_{SH} \mid \overline{b})(x)$$
$$= \gamma(\neg mc) \cdot \gamma(\neg b \mid \neg mc) \cdot \gamma(\neg isc \mid \neg mc) \cdot \gamma(c \mid \neg b \wedge \neg isc) \cdot 1(x)$$
$$= 0.80 \cdot 0.95 \cdot 0.80 \cdot x = 0.61 \cdot x$$

The (one-way) sensitivity function: in general

Consider a sensitivity analysis of $\mathcal{B} = (G, \Gamma)$ with output variable of interest V_o and evidence for set $\boldsymbol{E}$.

Consider an arbitrary network-parameter x from Γ. Then,

- the output probability of interest equals

$$\Pr(v_o \mid \boldsymbol{e})(x) = \frac{\Pr(v_o \wedge \boldsymbol{e})(x)}{\Pr(\boldsymbol{e})(x)} = \frac{a \cdot x + b}{c \cdot x + d}$$

  where a, b, c, and d are constants;

- if $c \neq 0$ is guaranteed, i.e. $\Pr(\boldsymbol{e})$ actually varies with x, then in essence only three constants are required:

$$\Pr(v_o \mid \boldsymbol{e})(x) == \frac{a/c \cdot x + b/c}{c/c \cdot x + d/c}$$

- The sensitivity function takes the form of (a fragment of) a rectangular hyperbola.

The (one-way) sensitivity function: specific case

Consider an network-parameter x from Γ. Then,

- if $x = \gamma(c_{V_i} \mid c_{\boldsymbol{\rho}(V_i)})$ is associated with a $V_i \in \boldsymbol{V}_G$ for which $\boldsymbol{\sigma}^*(V_i) \cap \boldsymbol{E} = \emptyset$, then the output probability of interest equals

  $$\Pr(v_o \mid \boldsymbol{e})(x) = a \cdot x + b$$

  where a and b are constants.

- The sensitivity function is linear.
- Note that this always holds in a prior network without evidence.

Proportional scaling of parameters

Upon varying a single model-parameter $x = \gamma(v_i \mid \boldsymbol{\rho})$ for a variable V, the other moedl-parameters $\gamma(v_j \mid \boldsymbol{\rho})$, $j \neq i$, for V are co-varied:

$$\gamma(v_j \mid \boldsymbol{\rho})(x) = \begin{cases} x & \text{if } j = i \\ \gamma(v_j \mid \boldsymbol{\rho}) \cdot \dfrac{1-x}{1-\gamma(v_i \mid \boldsymbol{\rho})} & \text{otherwise} \end{cases}$$

The scheme of proportional scaling keeps the proportions between the model-parameters $\gamma(v_j \mid \boldsymbol{\rho})$, $j \neq i$, constant.

The scheme results in the smallest distance[9] between the original and the new distribution.

[9] Chan & Darwiche (2003): A distance measure for bounding probabilistic belief change

Computing the sensitivity function $f(x)$

Building upon its general form, it suffices to compute the constants of a sensitivity function:

- a simple algorithm computes the output probability for a small number of values of the model-parameter under study and solves the resulting system of equations;[10]
- a more intricate algorithm establishes the constants in the function analytically through propagation;
- observing the relation between the constants and derivatives of $f(x)$, we can also use a differential approach.[11]

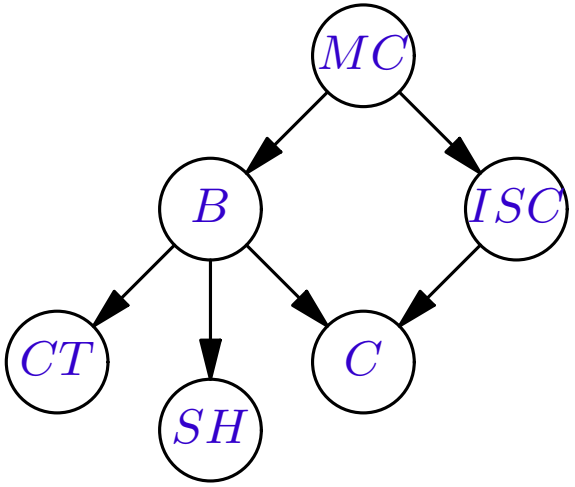
---

[10] The next slides illustrate this algorithm; if you need to compute a sensitivity function by hand, please use the analytic approach from slide 337!

[11] Darwiche (2000): A differential approach to inference in Bayesian networks.

Computing an example sensitivity function (1)

Consider once again the following Bayesian network:



$\gamma(b \mid mc) = 0.20$
$\gamma(b \mid \neg mc) = 0.05$

$\gamma(sh \mid b) = 0.80$
$\gamma(sh \mid \neg b) = 0.60$

$\gamma(ct \mid b) = 0.95$
$\gamma(ct \mid \neg b) = 0.10$

$\gamma(mc) = 0.20$

$\gamma(c \mid b, isc) = 0.80$
$\gamma(c \mid \neg b, isc) = 0.80$
$\gamma(c \mid b, \neg isc) = 0.80$
$\gamma(c \mid \neg b, \neg isc) = 0.05$

$\gamma(isc \mid mc) = x$
$\gamma(isc \mid \neg mc) = 0.20$

Compute the sensitivity function for output probability $\Pr(mc \mid isc)$ as a function of $x = \gamma(isc \mid mc)$:

1) compute the output probability from the network three (max four) times, for different values of x, using standard inference

For example, for $x = 0.2$, $x = 0.5$ and $x = 0.8$ we find:

$$\Pr(mc \mid isc)(0.2) = 0.200$$
$$\Pr(mc \mid isc)(0.5) = 0.385$$
$$\Pr(mc \mid isc)(0.8) = 0.500$$

Computing an example sensitivity function (2)

Compute the sensitivity function for output probability $\Pr(mc \mid isc)$ as a function of $x = \gamma(isc \mid mc)$:

2) establish a system of linear equations:

$$\begin{aligned} \Pr(mc \mid isc)(0.2) &= 0.200 \\ \Pr(mc \mid isc)(0.5) &= 0.385 \\ \Pr(mc \mid isc)(0.8) &= 0.500 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \frac{a' \cdot 0.2 + b'}{0.2 + d'} &= 0.200 \\ \frac{a' \cdot 0.5 + b'}{0.5 + d'} &= 0.385 \\ \frac{a' \cdot 0.8 + b'}{0.8 + d'} &= 0.500 \end{aligned}$$

Computing an example sensitivity function (3)

Compute the sensitivity function for output probability $\Pr(mc \mid isc)$ as a function of $x = \gamma(isc \mid mc)$:

3) solve the system of linear equations:

$a' \cdot 0.2 + b' = 0.200 \cdot 0.2 + 0.200 \cdot d'$ and
$a' \cdot 0.5 + b' = 0.385 \cdot 0.5 + 0.385 \cdot d'$

which together give $a' = 1.525/3 + 1.85/3 \cdot d'$.

Combining this with equation

$$a' \cdot 0.8 + b' = 0.500 \cdot 0.8 + 0.500 \cdot d'$$

gives $b' = -0.2/30 + 0.2/30 \cdot d'$.

Substituting a' and b' in the first equation gives $d' = 1.65/2.1 \approx 0.786$ and therefore $a' \approx 0.993$ and $b' \approx -0.001$.

Practicable sensitivity analysis

Straightforward sensitivity analysis of a Bayesian network is infeasible. The digraph of the network, however, induces

- algebraic independence of the output probability of various network-parameters;
- simple mathematical functions that relate the output probability to the potentially influential network-parameters.

By exploiting these properties, sensitivity analysis of a Bayesian network is rendered practicable.

Still, the number of sensitivity functions returned from all potentially influential network-parameters can be quite large.

How do we select the network-parameters that we consider sensitive and that require further study ?

Selection of sensitive assessments

A sensitivity analysis results in a large amount of data.

Example: the oesophageal cancer network:

In the prior network, 206 parameters potentially influence the 6 probabilities of $\Pr(Stage) \rightarrow$ 1236 sensitivity functions.

Given patient evidence (156), the number of potentially influential network-parameters may become 826. ■

Various selection criteria can be employed to select network-parameters that deserve attention.

Selection criteria

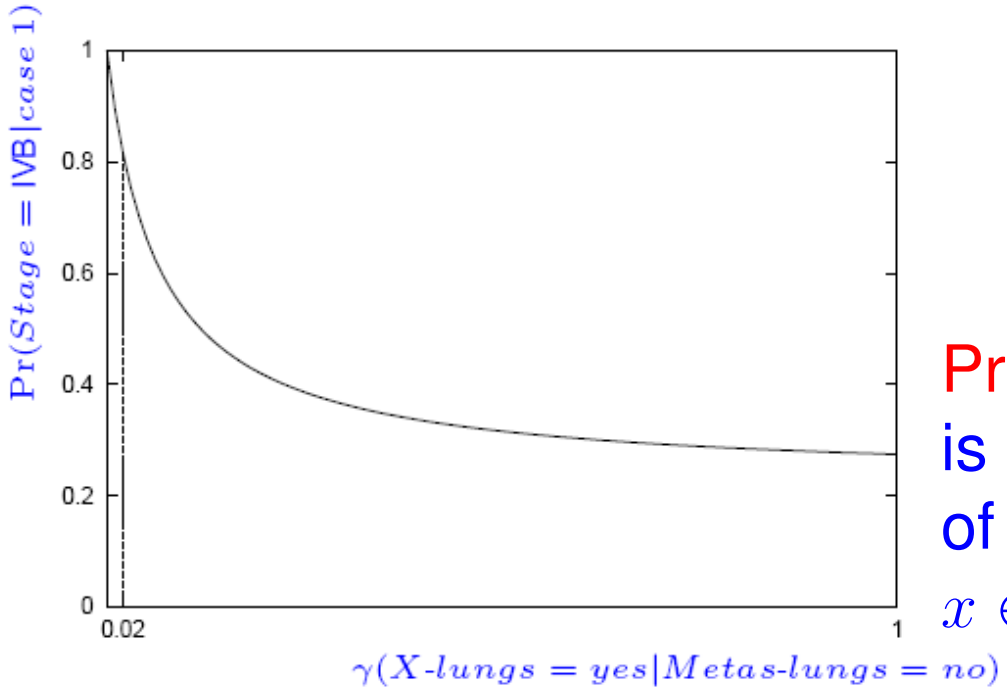
Parameter assessments that may require further study can be selected based upon:

- absolute effect of variation on output probability: $|f(0) - f(1)|$;
- plausible effect on output probability;
- the sensitivity value, i.e. the absolute value of the first derivative of the sensitivity function at original assessment;
- the vertex proximity, i.e the distance between the original assessment of the network-parameter and the vertex ("shoulder") of the function;
- the admissible deviation, i.e. the variation allowed in the network-parameter without changing the most likely value of the variable of interest.

The sensitivity value as selection criterion

Consider sensitivity function $f(x)$ for network-parameter x.Let x_0 be the original assessment for x.

The absolute value of the first derivative of $f(x)$ in $(x_0, f(x_0))$, also called the sensitivity value, captures how sensitive the output is to varying x.



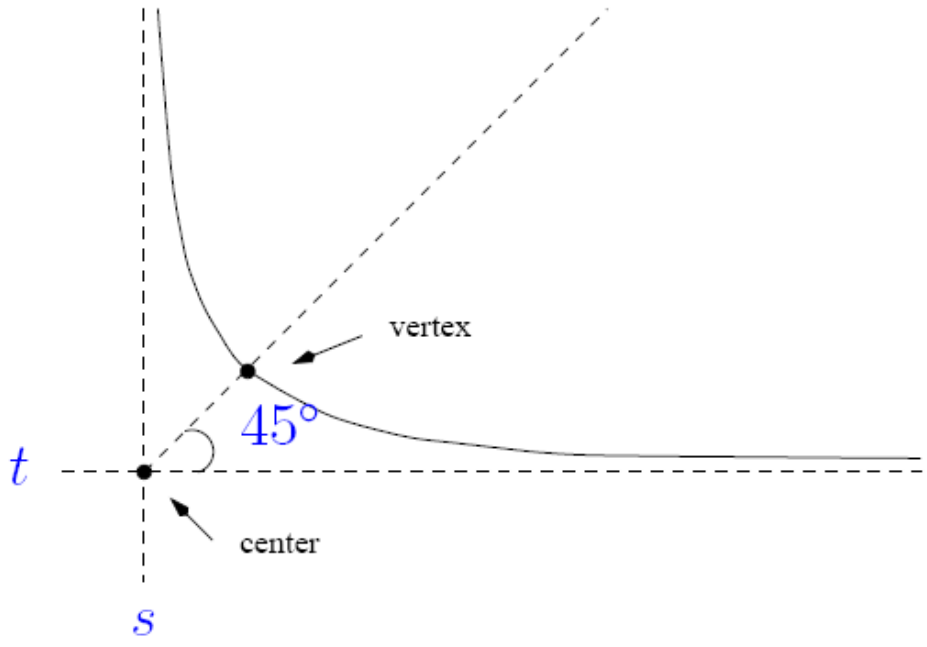
$$\left|\frac{\partial f}{\partial x}(0.02)\right| = 6.97$$

Problem: the first derivative is a good approximation of the function only for $x \in [x_0 - \epsilon, x_0 + \epsilon]$.

Vertex proximity

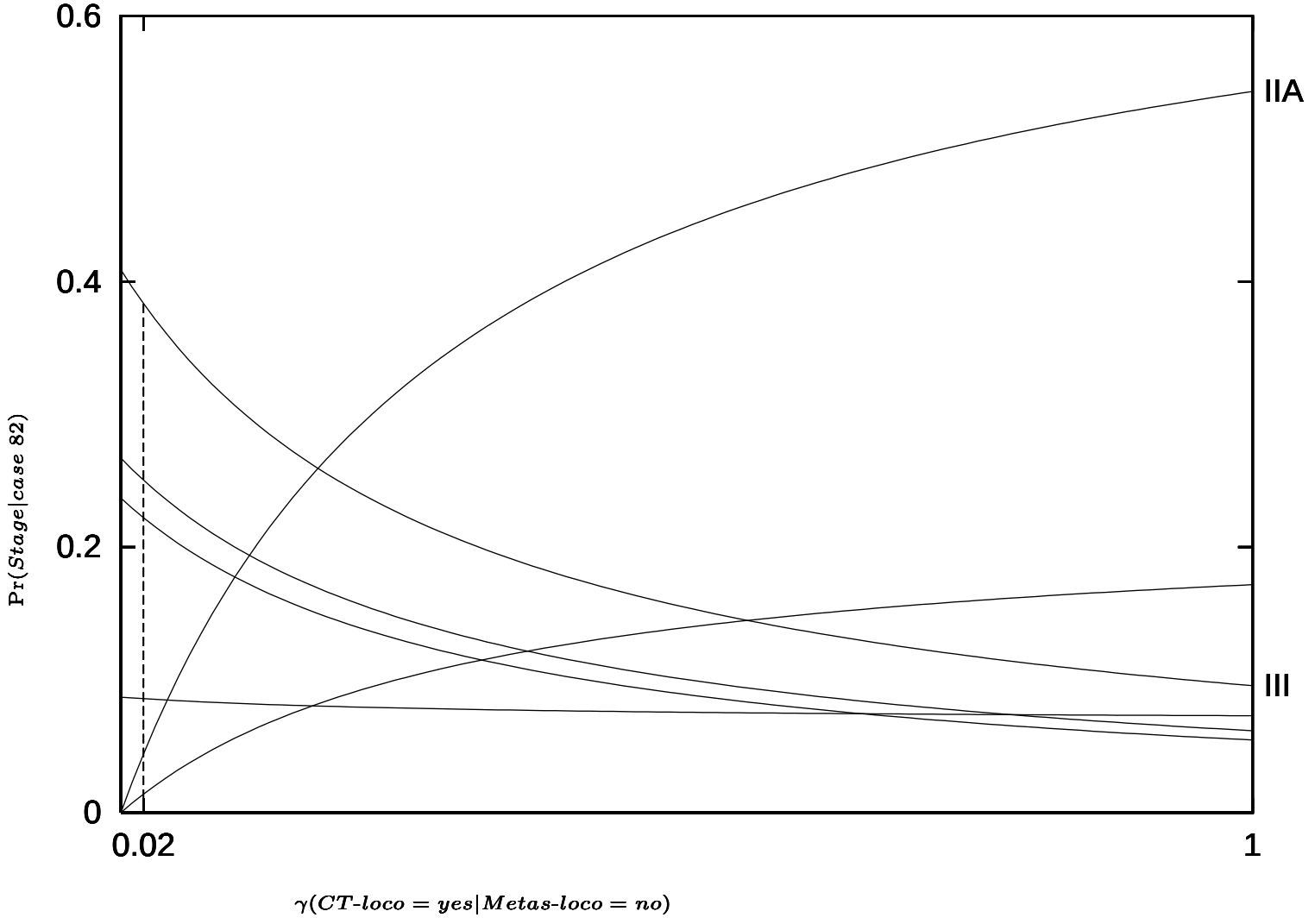
The sensitivity value in x_0 may be small near the vertex (shoulder) of a sensitivity function.

Yet, slight variation of the parameter around x_0 can have a large effect on the outcome probability.



Solution: if x_0 is close to x_{vertex}, then select x for further study, regardless of the sensitivity value.

The admissible deviation



small sensitivity value, smaller admissible deviation

More elaborate sensitivity analyses

Properties of an n-way analysis for $n > 1$:

- all n model-parameters are varied simultaneously.
- reveals possible interactions, or synergistic effects.
- sensitivity function is a fraction of two multi-linear functions in the model-parameters under study.
- hardly any research into shapes and properties of n-way sensitivity functions for $n \geq 2$.
- interpretation of results is hard, especially for $n > 2$.

Two-way sensitivity analyses

With a two-way sensitivity analysis, two model-parameters are varied simultaneously:

$$f(x,y) = \frac{c_1 \cdot x \cdot y + c_2 \cdot x + c_3 \cdot y + c_4}{c_5 \cdot x \cdot y + c_6 \cdot x + c_7 \cdot y + c_8}$$

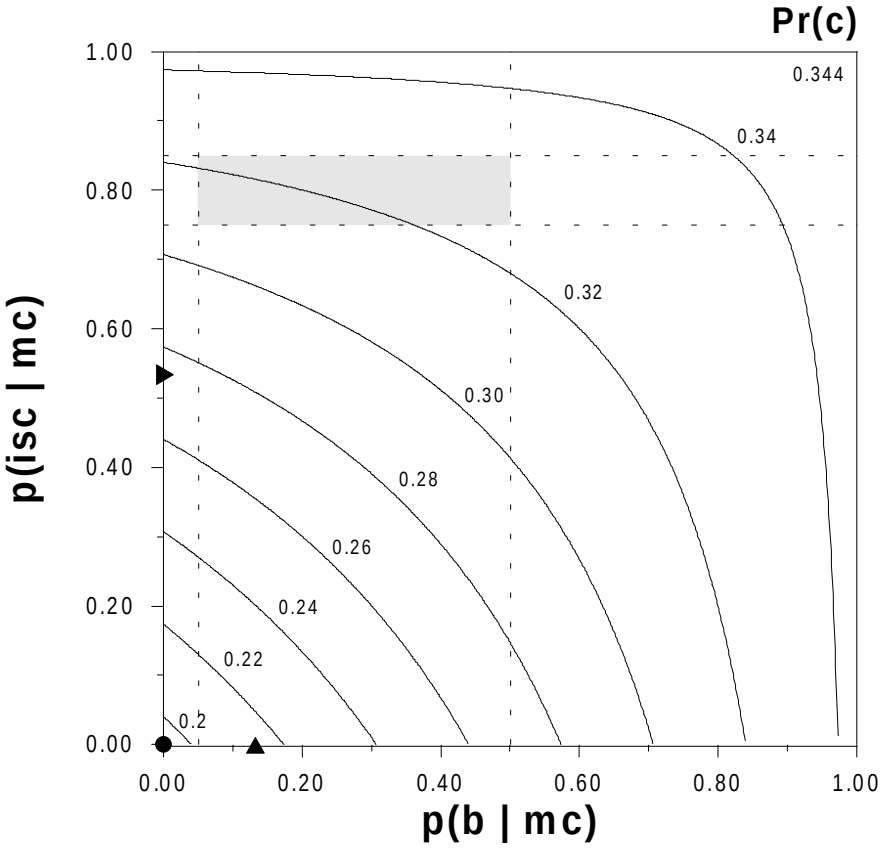
A two-way analysis reveals possible synergistic effects (c_1, c_5) not found from two one-way analyses.

<u>Selection criteria</u>: Parameter assessments that may require further study can be selected based upon:

- absolute effect of variation on output probability;
- plausible effect on output probability;
- the (max) sensitivity value: $\sqrt{(\frac{\partial f}{\partial x}(x_0,y_0))^2 + (\frac{\partial f}{\partial y}(x_0,y_0))^2}$
- contour distances, i.e the distances between iso-probability lines in a 2D projection of the sensitivity function.

Contour distance

A two-way analysis reveals *synergistic* effects.



- absolute distance: the smaller the distance, the more sensitive the output probability is to parameter variation;
- relative distance: varying distances indicate interaction effects.

The iso-probability contours here are not equi-distant due to non-zero interaction terms in the sensitivity function.

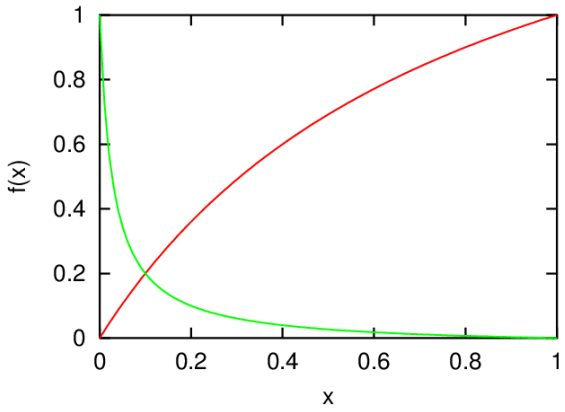
Intermezzo

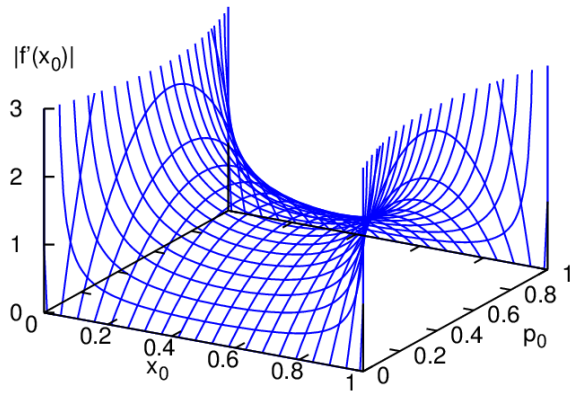
De following slides briefly summarize more research related to sensitivity analysis done in our Department.

This does not have to be studied for the exam;
you can also skip to the topic of evaluation.

Brief: robustness to parameter inaccuracies II

We can provide general bounds on sensitivity functions through (x_0, p_0) and on their properties[12]





which can be further bounded[13] given $f_{\Pr(\boldsymbol{e})}(x) = c \cdot x + d$:

$$f_{\Pr(h|\boldsymbol{e})}(x) = \frac{r}{x - s} + t, \qquad r = (x_0 - s) \cdot (p_0 - t)$$

for asymptotes $x = s = -\frac{d}{c}$ and $y = t$.

[12] S. Renooij, L.C. van der Gaag (2004). Evidence-invariant sensitivity bounds. In: UAI 2004.

[13] S. Renooij, L.C. van der Gaag (2005). Exploiting evidence-dependent sensitivity bounds. In: UAI 2005.

Brief: robustness to structure changes

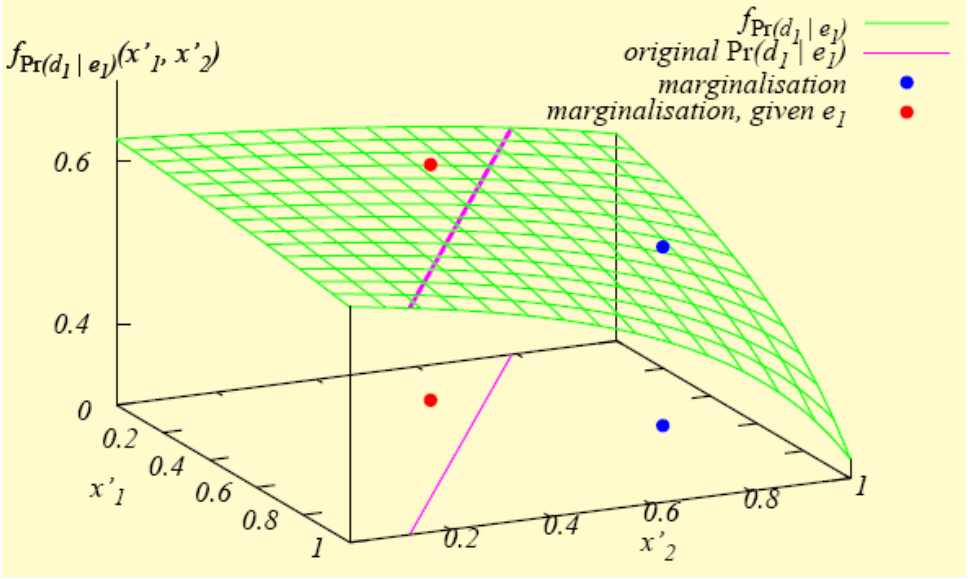
We can simulate the removal of an arc by posing constraints on an n-way sensitivity function[14]

Original CPT for node B:

	c_1		c_2	
	a_1	a_2	a_1	a_2
b_1	0.7	0.1	0.9	0.6
b_2	0.3	0.9	0.1	0.4

For removing $A \to B$:

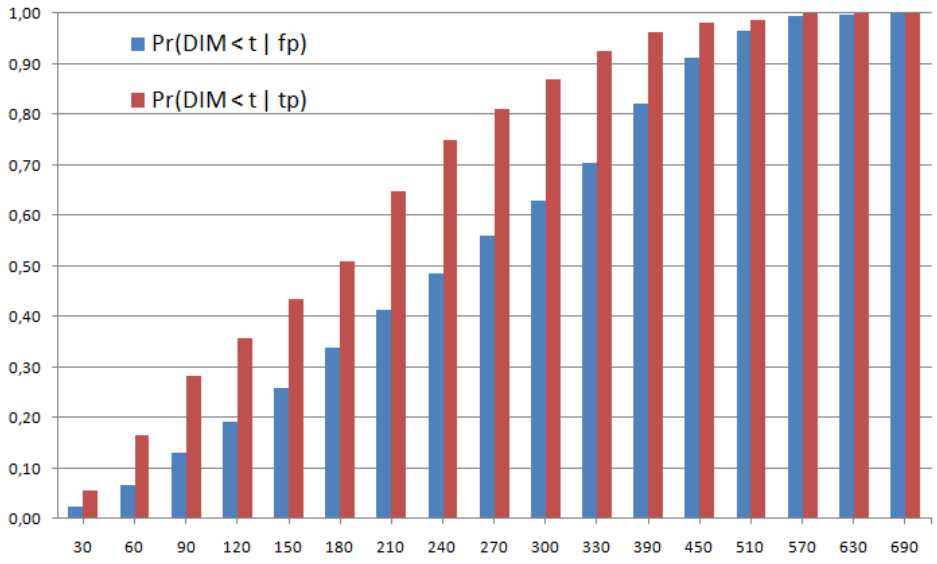
	c_1		c_2	
	a_1	a_2	a_1	a_2
b_1	x'_1		x'_2	
b_2	$1 - x'_1$		$1 - x'_2$	



[14] S. Renooij (2010). Bayesian network sensitivity to arc-removal. In: PGM 2010

Brief: robustness to discretisation

We can study the effect of choosing a different discretisation[15]



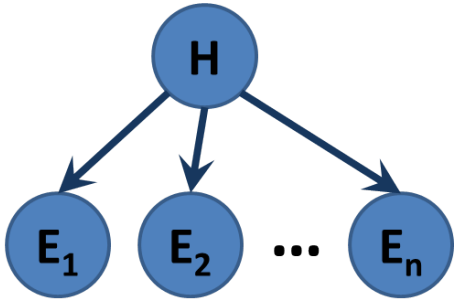
- changing a discretisation threshold is like varying a network-parameter

[15] R. Bertens, L.C. van der Gaag, S. Renooij (2012). Discretisation effects in naive Bayesian networks. In: IPMU 2012

Brief: sensitivity to model assumptions

We can gain understanding about the behaviour of

- networks of restricted topology
  - naive Bayesian network classifiers[16]



  - multi-dimensional Bayesian network classifiers[17]
- causal interaction models[18]

---

[16] S. Renooij, L.C. van der Gaag (2008). Evidence and scenario sensitivities in naive Bayesian classifiers. IJAR vol 49.

[17] J.H. Bolt, S. Renooij (2014). Sensitivity of multi-dimensional Bayesian classifiers. In: ECAI 2014.
& J.H. Bolt, S. Renooij (2015). Robustness of multi-dimensional Bayesian network classifiers. In: BNAIC 2015.

[18] S.P.D. Woudenberg, L.C. van der Gaag (2015), Propagation effects of model-calculated probability values in Bayesian networks, IJAR vol 61.

Brief: results applied in other contexts

Rather than using sensitivity functions as analysis tools, we can exploit their properties in other contexts[19]

- parameter tuning [20]
- pre-processing inference in credal networks[21]
- ...?

[19] J.H. Bolt, S. Renooij (2017). Structure-based categorisation of Bayesian network parameters. In: ECSQARU 2017

[20] J.H. Bolt, S. Renooij (2014). Local sensitivity of Bayesian networks to multiple simultaneous parameter shifts. PGM 2014

[21] J.H. Bolt, J. De Bock, S. Renooij (2016). Exploiting Bayesian network sensitivity functions for inference in credal networks. In: ECAI 2016

End of Intermezzo

Evaluation of Bayesian networks

An evaluation of the practical value of a Bayesian network consists of the following steps:

1) select realistic cases to evaluate
   (for example from data or scenarios);
2) select the outcome variable(s) of interest;
3) choose a standard of validity;
4) compute, from the network, the outcome for each case;
5) compare the outcome to your standard of validity.

Evaluation of Bayesian networks: an example

Consider the evaluation of the practical value of the oesophageal cancer network.

- data: symptoms and test-results for 156 patients (average: 14.8 of the 25, per patient);
- outcomes of interest: *Stage* of the tumour: I, IIA, IIB, III, IVA, IVB;
- standard of validity: assessment of the *stage*, given by the physicians.

From the oesophageal cancer network we now compute the *stage* for each of the 156 patients.

Patient file for Patient X

Passage:	can pass mashed food
Weightloss:	none
Physical exam:	swollen lymph nodes neck

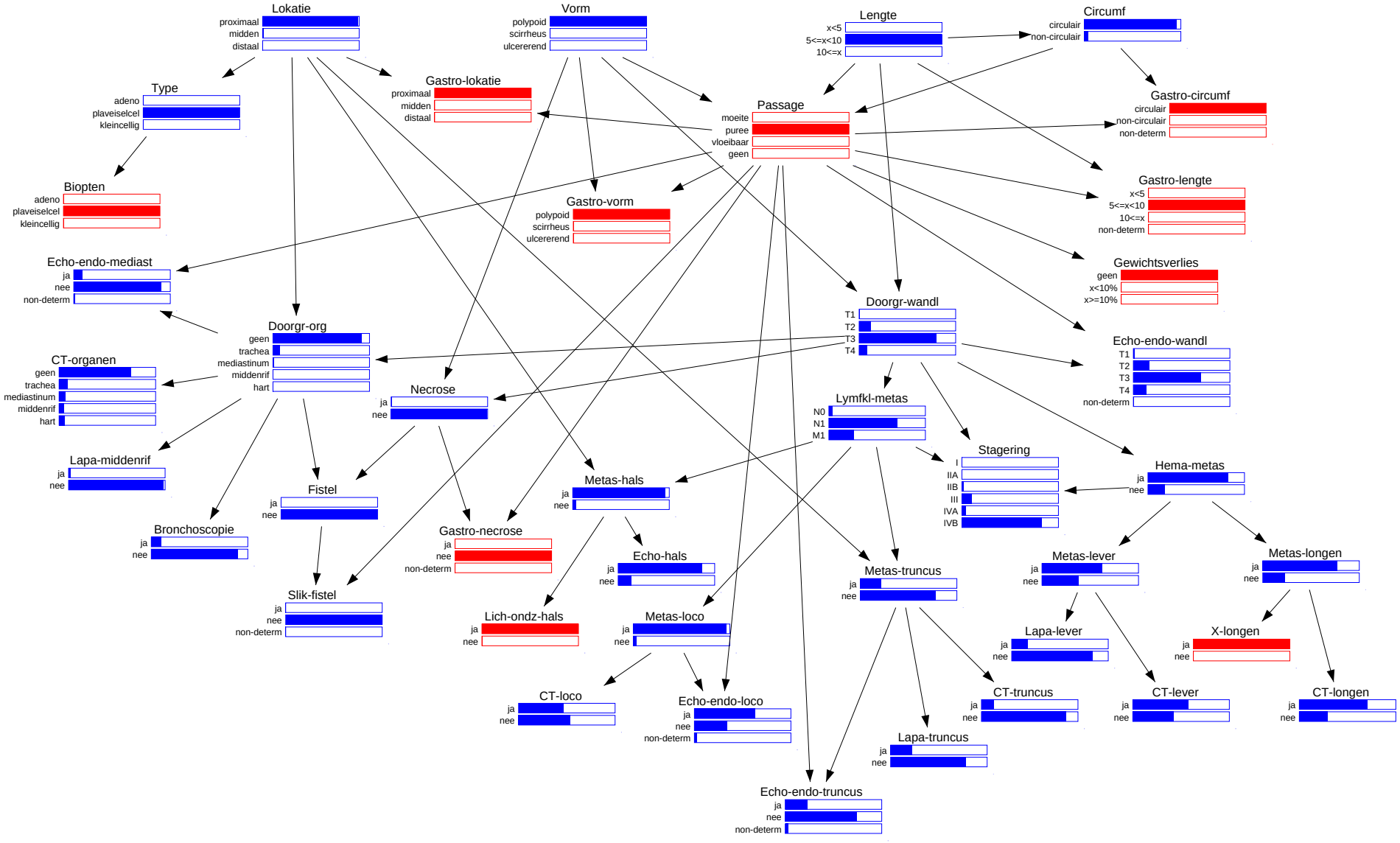
Biopsy:	squamous
X-lungs:	metastases
Bronchoscopy:	×
Sono-cervix:	×
Barium swallow:	×

Gastroscopy:	circumf:	length:	location:	necrosis:	shape:
	circular	7 cm	proximal	absent	polypoid

CT-scan (liver, locoregion, lungs, organs, truncus):	×
Endosonography (locoregion, mediastinum, truncus, wall):	×
Laparascopy (liver, diaphragm, truncus):	×

Diagnosis: stage = I/IIA/IIB/III/IVA/IVB

Diagnosing Patient X



The percentage correct

After processing evidence, a Bayesian network gives a posterior probability distribution for the outcome variable.

The standard of validity, however, usually consists of a single value for the outcome variable.

- The most likely value of the outcome variable is chosen as *the* outcome of the network;
- *the* outcome is compared against the standard: the outcome is either correct or incorrect.

The percentage of cases where the outcome predicted by the network is correct according to the standard of validity is called the percentage correct (or: accuracy).

The percentage correct: an example

Compare for each patient the *stage* predicted by the network against the *stage* assessed by the physicians.

For 133 of the 156 patients, the network gives an accurate prediction:

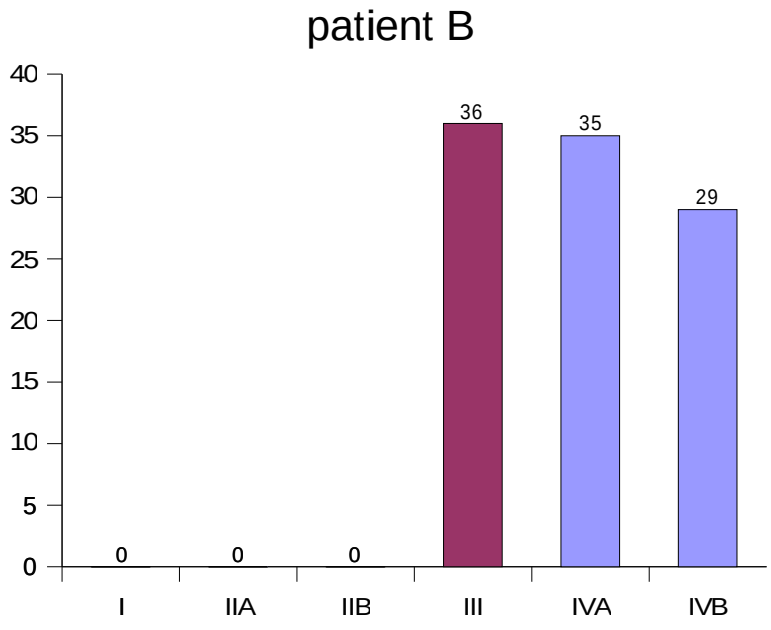
		network						
		I	IIA	IIB	III	IVA	IVB	*total*
	I	**2**	0	0	0	0	0	2
	IIA	0	**37**	0	1	0	0	38
phys.	IIB	0	1	**0**	3	0	0	4
	III	1	10	0	**36**	0	0	47
	IVA	0	0	0	4	**35**	0	39
	IVB	0	0	0	3	0	**23**	26
	total	3	48	0	47	35	23	156

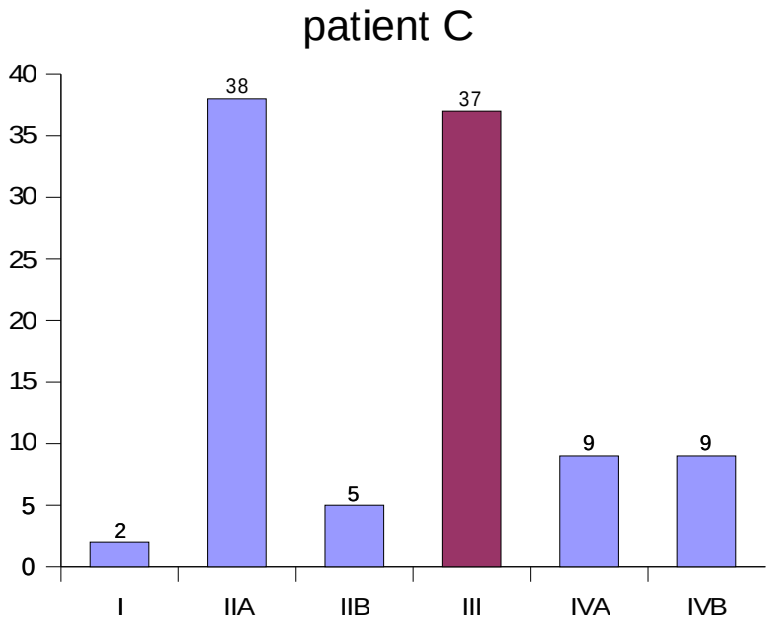
The percentage correct is therefore 85%.

Explaining the differences

Differences between the outcomes of a network and the standard of validity can originate from several sources:

- modelling errors;
- errors in the standard, or in the data;
- random variation:





Evaluation scores: the *Brier* score

The uncertainty expressed in the predicted distribution can be taken into account in the evaluation.

Let $p_{ij} = \Pr(v_j \mid \boldsymbol{e_i})$ be the predicted (network) probability for case i and value j of the outcome variable.

$$\text{Let } s_{ij} = \begin{cases} 1 & \text{if outcome } j \text{ is correct outcome for case } i \\ & \text{(according to standard of validity);} \\ 0 & \text{otherwise} \end{cases}$$

The Brier score for the predicted distribution for case i now is

$$B_i = \sum_j (p_{ij} - s_{ij})^2$$

The Brier score lies within the interval $[0, 2]$, where 0 indicates a perfect prediction.

The Brier score: an example

Consider evaluating the oesophageal cancer network, where

- p_{ij} is the network probability computed for patient i and stage $j \in \{I, \ldots, IVB\}$;
- s_{ij} returns 1 if patient i's medical file states stage j, and 0 otherwise.

The Brier score for patient i now is $B_i = \sum_{j=\mathrm{I},\ldots,\mathrm{IVB}} (p_{ij} - s_{ij})^2$

For patients X, B and C we find, respectively:

$$B_X = (0-0)^2 + (0.01-0)^2 + (0.04-0)^2 + (0.14-0)^2 +$$
$$+ (0.06-0)^2 + (0.75-1)^2 = 0.09$$
$$B_B = 3 \cdot (0-0)^2 + (0.36-1)^2 + (0.35-0)^2 + (0.29-0)^2 = 0.62$$
$$B_C = (0.02-0)^2 + (0.38-0)^2 + (0.05-0)^2 + (0.37-1)^2 +$$
$$+ (0.09-0)^2 + (0.09-0)^2 = 0.56$$

Average Brier score

We can compute an average Brier score over n 'forecasts':

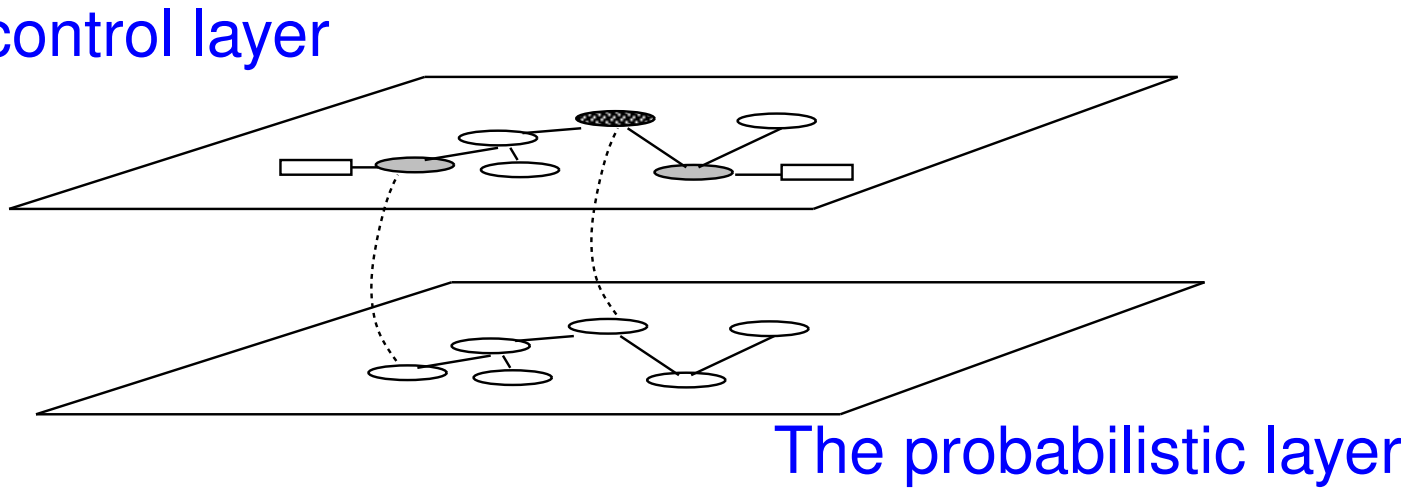
$$B = \frac{1}{n} \sum_{i=1,\ldots,n} B_i$$

An example: The average Brier score over all patients per predicted-stage / actual-stage combination:

		network					
		I	IIA	IIB	III	IVA	IVB
	I	**0.21**	–	–	–	–	–
	IIA	–	**0.28**	–	1.52	–	–
phys.	IIB	–	1.17	–	0.98	–	–
	III	1.40	0.89	–	**0.26**	–	–
	IVA	–	–	–	0.75	**0.08**	–
	IVB	–	–	–	0.87	–	**0.06**

The average Brier score over all 156 patients is: 0.29 ■

Decision support: a two-layer problem solving architecture



Probabilistic layer for probabilistic reasoning:

- stores: a Bayesian network;
- tasks: receive evidence, propagate it, and return requested probabilities.

Control layer for (intelligent) control over reasoning

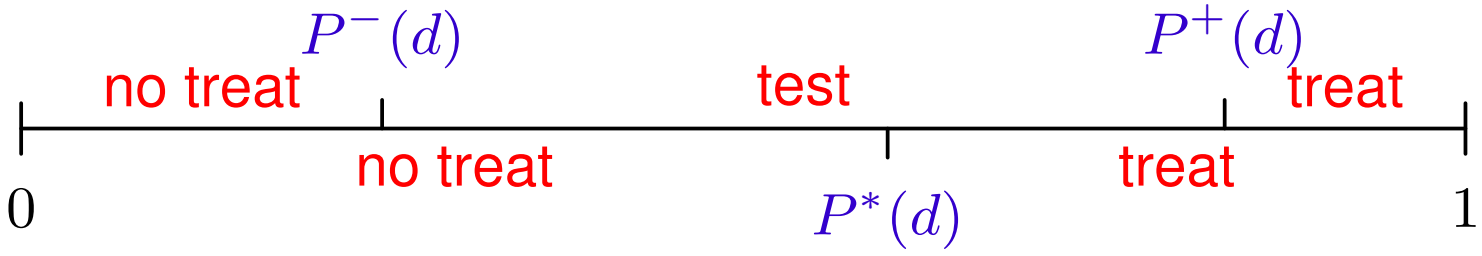
- stores: non-probabilistic information;
- tasks: make strategic decisions by sending evidence, requesting probabilistic information, computing non-probabilistic information.

Problem solving: Threshold decision making

The purpose of threshold decision making is supporting the choice between therapeutic decision alternatives.

A system for threshold decision making has the following tasks:

- Diagnostic reasoning: compute the probability $\Pr(d)$ of some hypothesis (diagnosis), based upon the available findings.



- Treatment advisement: give advise concerning treatment, based upon $\Pr(d)$ and the threshold values for the treatment options.

Threshold decision making

A simple strategy for threshold decision making using a Bayesian network $\mathcal{B} = (G, \Gamma)$:

PROCEDURE THRESHOLDDECISION($\mathcal{B}$,$c_{\boldsymbol{E}}$,P,A):
  PROPAGATE-EVIDENCE($\mathcal{B}$,$c_{\boldsymbol{E}}$);
  ADVISE(P, A)
END

The procedure is called with

- evidence $c_{\boldsymbol{E}}$ for a set of nodes $\boldsymbol{E} \subset \boldsymbol{V}_G$, and
- a set of threshold values P for the diagnosis under consideration.

The procedure returns a treatment alternative of $A \notin \boldsymbol{V}_G$.

Expected utility of treatment

The choice between two treatment alternatives depends on their expected benefit. Benefit can be defined in terms of utility.

Consider hypothesis node H and evidence $\boldsymbol{e}$ for a nodes $\boldsymbol{E}$; variable A models different treatment alternatives.

- the desirability of each c_{AH} of A and H is given by a subjective utility $u(c_{AH})$;
- the expected utility of each treatment alternative c_A then is

$$\hat{u}(c_A) = \sum_{c_H} u(c_A \wedge c_H) \cdot \Pr^{\boldsymbol{e}}(c_H), \quad \text{where} \quad c_A \wedge c_H \equiv c_{AH}$$

Advise: treatment alternative with highest expected utility.

Drawback: each $\hat{u}(c_A)$ has to be recomputed every time a different value for $\Pr^{\boldsymbol{e}}(c_H)$ is encountered...

Expected utility for setting thresholds

Let H, $\boldsymbol{e}$ and A be as before. Expected utility can be written as a function of $\Pr^{\boldsymbol{e}}(h)$ for value of interest h of H.

In case of a binary-valued H this function equals:

$$\begin{aligned}
\hat{u}(c_A) &= \sum_{c_H} u(c_A \wedge c_H) \cdot \Pr^{\boldsymbol{e}}(c_H) \\
&= u(c_A \wedge h) \cdot \Pr^{\boldsymbol{e}}(h) + u(c_A \wedge \neg h) \cdot \Pr^{\boldsymbol{e}}(\neg h) \\
&= (u(c_A \wedge h) - u(c_A \wedge \neg h)) \cdot \Pr^{\boldsymbol{e}}(h) + u(c_A \wedge \neg h)
\end{aligned}$$

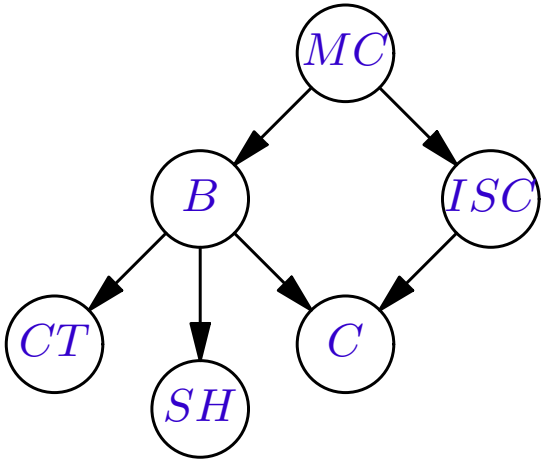
Therefore, with $x = \Pr^{\boldsymbol{e}}(h)$ we have

$$\hat{u}(c_A)(x) = (u(c_A \wedge h) - u(c_A \wedge \neg h)) \cdot x + u(c_A \wedge \neg h)$$

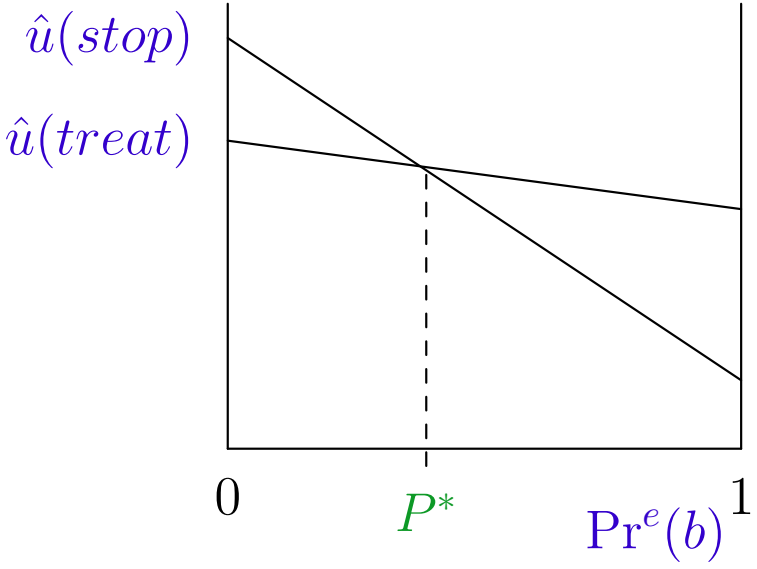
Threshold probabilities are computed by solving x (for each pair of alternatives a_i and a_j, $i \neq j$, for A) from $\hat{u}(a_i)(x) = \hat{u}(a_j)(x)$.

An example

Consider the following network and utilities $u(c_A \wedge c_H)$:



$$u(stop \wedge b) = 0.02$$
$$u(stop \wedge \neg b) = 1.00$$
$$u(treat \wedge b) = 0.50$$
$$u(treat \wedge \neg b) = 0.92$$



Threshold value $P^* \approx 0.143$ is computed from:

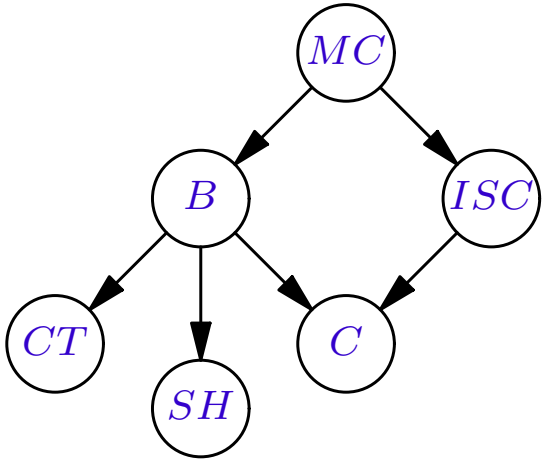
$$\hat{u}(treat)(x) = (0.50 - 0.92) \cdot x + 0.92$$
$$\hat{u}(stop)(x) = -0.98 \cdot x + 1.00$$

where $x = \Pr^{\boldsymbol{e}}(h) = \Pr(b)$

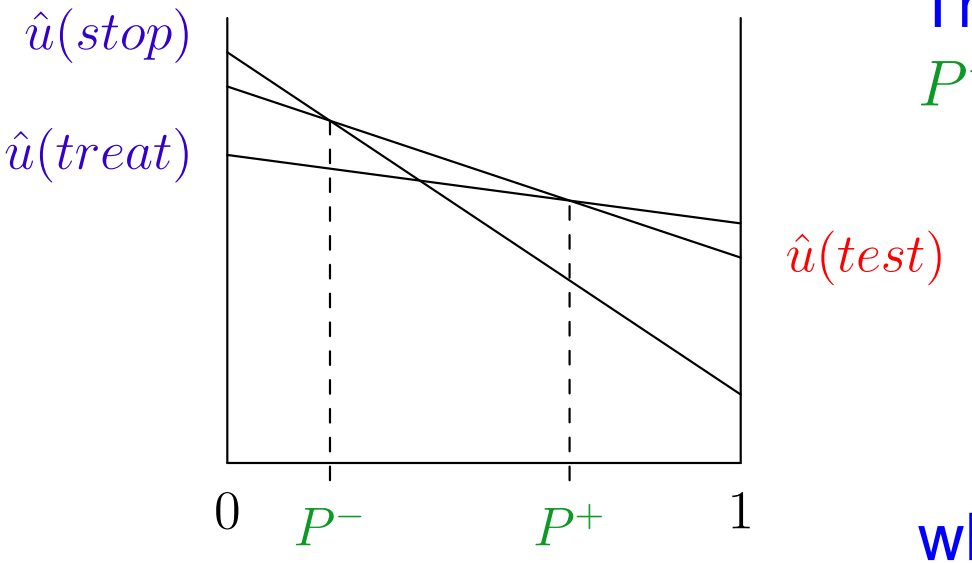
Should a patient with $\Pr(b) = 0.10$ be treated or not?

An example

Consider the following network and utilities $u(c_A \wedge c_H)$:



$$u(stop \wedge b) = 0.02$$
$$u(stop \wedge \neg b) = 1.00$$
$$u(test \wedge b) = 0.45$$
$$u(test \wedge \neg b) = 0.98$$
$$u(treat \wedge b) = 0.50$$
$$u(treat \wedge \neg b) = 0.92$$



Threshold values $P^- \approx 0.044$ and $P^+ \approx 0.545$ are computed from:

$$\hat{u}(stop)(x) = -0.98 \cdot x + 1.00$$
$$\hat{u}(treat)(x) = -0.42 \cdot x + 0.92$$
$$\hat{u}(test)(x) = -0.53 \cdot x + 0.98$$

where $x = \Pr^{\boldsymbol{e}}(h) = \Pr(b)$

Should a CT-scan be ordered for a patient with $\Pr(b) = 0.10$?

Threshold decision making: summary

For threshold decision making, the probabilistic layer and the control layer have the following functionality:

Probabilistic layer:

- propagates evidence and returns requested probabilities

Control layer:

- stores utility functions
- computes and stores threshold probabilities for different treatment choices;
- compares probabilities with appropriate thresholds and returns a treatment advise based upon the comparisons.

Problem solving: Diagnostication

Diagnostication: determine the most likely hypothesis (diagnosis), at the lowest possible costs (a.k.a adaptive testing in Intelligent Tutoring Systems).

A system for diagnostication has the following tasks:

- Diagnostic reasoning: determine most likely problem cause from available information about its manifestations.
- Test selection: select appropriate tests to gain more information about the manifestations.
- Stopping criterion evaluation: check whether the current diagnosis is sufficiently reliable.

Simple diagnostication

A simple strategy for diagnostication using a Bayesian network $\mathcal{B} = (G, \Gamma)$:

```
PROCEDURE DIAGNOSTICATION(𝓑,E,H):
   SUFFICIENT ← FALSE;
   WHILE E ≠ ∅ AND NOT SUFFICIENT DO
         E_i ← SELECT-TEST(E);
         e_i ← GATHER-EVIDENCE(E_i);
         PROPAGATE-EVIDENCE(B,e_i);
         E ← E \ {E_i};
         SUFFICIENT ← EVALUATE-STOP
   OD;
   DIAGNOSE(H)
END
```

The procedure is called with the set $\boldsymbol{E} \subset \boldsymbol{V}_G$ of all evidence nodes. It returns a sufficiently reliable hypothesis for $H \in \boldsymbol{V}_G$.

Test-selection measures

Gathering evidence has benefit for diagnostication, as it may decrease uncertainty concerning the diagnosis.

Most often information measures are used to establish the expected benefit:

- Shannon entropy;
- Gini index;
- misclassification error;
- Kullback-Leibler divergence (uses cross entropy);
- expected utility

These measures all measure uncertainty only; it is possible to include different types of cost as well.

Expected utility for selecting tests

Consider binary hypothesis node H. Let $\boldsymbol{e}$ denote the processed evidence and let E_i be a relevant uninstantiated evidence node.

- The utility of the value c_{E_i} for node E_i is defined as

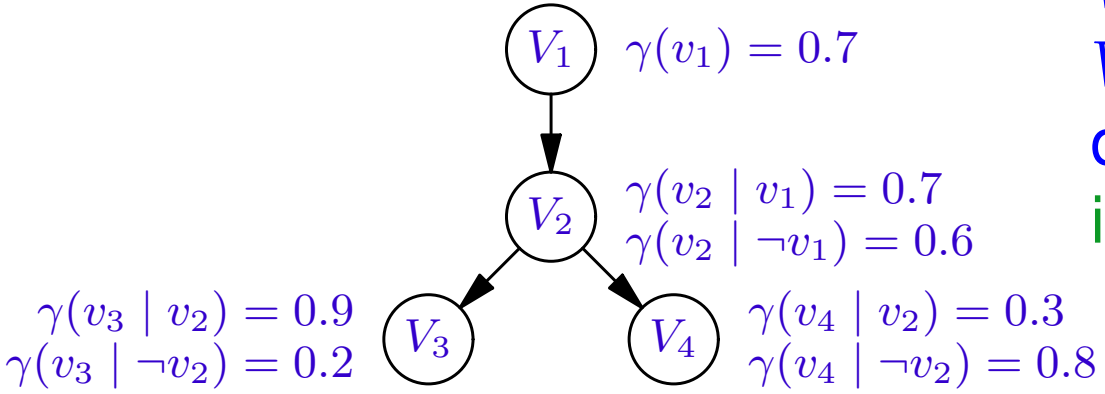
$$u(c_{E_i}) = |\Pr^{\boldsymbol{e}}(h) - \Pr^{\boldsymbol{e}}(h \mid c_{E_i})|$$

- the expected utility of observing a value for node E_i (i.e. doing the test) then is

$$\hat{u}(E_i) = \sum_{c_{E_i}} u(c_{E_i}) \cdot \Pr^{\boldsymbol{e}}(c_{E_i})$$

SELECT-TEST($\boldsymbol{E}$) now returns a node $E_i \in \boldsymbol{E}$ with highest expected utility.

An example



V_2 is an hypothesis node; V_1, V_3 and V_4 are evidence nodes; all are un-instantiated.

$$\Pr^{\boldsymbol{e}}(h) = \Pr(v_2) = 0.67$$

For V_3: $u(v_3) = |\Pr(v_2) - \Pr(v_2 \mid v_3)| = |0.67 - 0.901| = 0.231$

$u(\neg v_3) = |\Pr(v_2) - \Pr(v_2 \mid \neg v_3)| = |0.67 - 0.202| = 0.468$

The expected benefit of obtaining V_3's value is:

$$\begin{aligned} \hat{u}(V_3) &= u(v_3) \cdot \Pr(v_3) + u(\neg v_3) \cdot \Pr(\neg v_3) \\ &= 0.231 \cdot 0.669 + 0.468 \cdot 0.331 = 0.309 \end{aligned}$$

For V_1 and V_4 we similarly find $\hat{u}(V_1) = 0.042$ and $\hat{u}(V_4) = 0.223$.

$\hat{u}(V_3)$ is highest $\rightarrow$ user is prompted for value of V_3.

Some assumptions

To reduce computational complexity two simplifying assumptions are made:

- the myopia assumption: tests are selected and performed one at a time;
- the single-disorder assumption: all hypotheses are mutually exclusive.

Both assumptions, however, can be somewhat relaxed.

Stopping criteria

After processing newly obtained evidence, a stopping criterion is evaluated: if this criterion is met, the selection of tests is halted.
Some examples of stopping criteria:

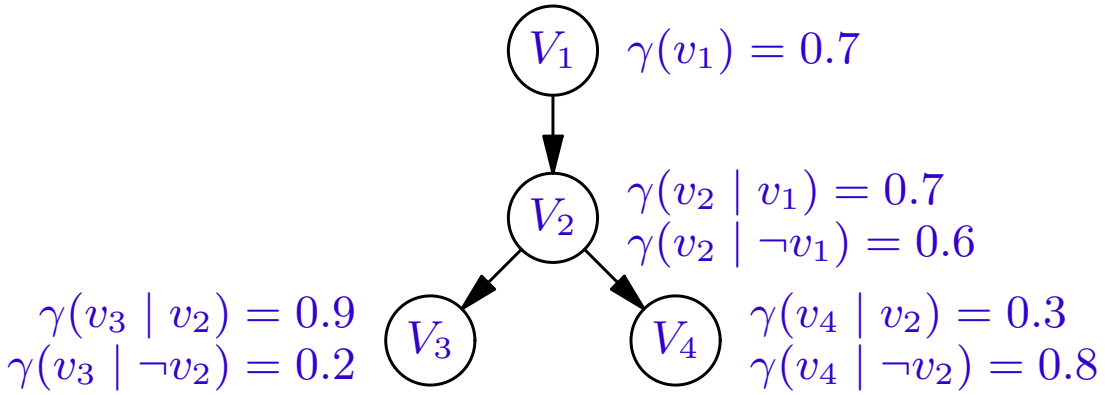
- sufficiency of confirmation: the probability of the hypothesis is above (below) a given threshold value;

  (or: take the entire distribution over the hypothesis node into consideration)

- sufficiency of information: the expected utilities of the relevant uninstantiated evidence nodes are below a given threshold value;

  (or: take the maximum utility instead of expected utility into consideration).

An example



V_2 is an hypothesis node; V_1, V_3 and V_4 are evidence nodes.

Suppose the stopping criterion for selecting tests is 'sufficiency of information' with a threshold value of 0.1.

With evidence $V_3 = true$, we find $\Pr^{\boldsymbol{e}}(h) = \Pr^{v_3}(v_2) = 0.90$.

The expected utilities for V_1 and V_4 are now updated for $\boldsymbol{e} = v_3$:

$$\hat{u}(V_1) = 0.017 \text{ and } \hat{u}(V_4) = 0.089$$

Both expected utilities are below 0.1 so selection of tests is halted.

Diagnostication: summary

For diagnostication, the probabilistic layer and the control layer have the following functionality:

Probabilistic layer:

- propagates evidence and returns requested probabilities

Control layer:

- stores knowledge concerning the roles of different variables (hypothesis, evidence, intermediate);
- stores and computes (expected) utilities of the different tests available;
- selects the most appropriate tests;
- evaluates the stopping criterion.

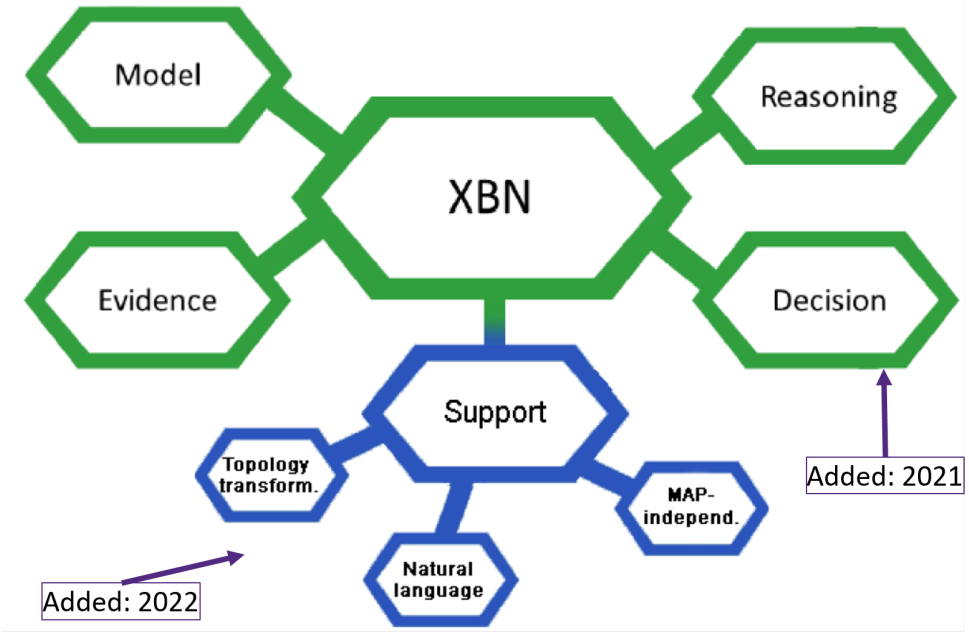
Explanation of Bayesian networks

The ability to explain a Bayesian network and its predictions is crucial for its acceptance (explainable AI)!

- what can and should we explain?
- for whom is the explanation intended?
  - BN expert / domain expert / user
- how to explain?
- ...

Explaining Bayesian networks

- 1992: *Explanation in Bayesian belief networks* (Stanford PhD thesis by H.J. Suermondt)
- 2001: *A Review of Explanation Methods for Bayesian Networks* (KER paper by C. Lacave and F.J. Díez)



2021: *A taxonomy of explainable Bayesian networks* (I.P. Derks, A. de Waal)

2022: *Extending MAP-independence for Bayesian network explainability* (E. Valero-Leal, P. Larrañaga, C. Bielza)

Analysis for explaining decisions

Derks & De Waal (2021):

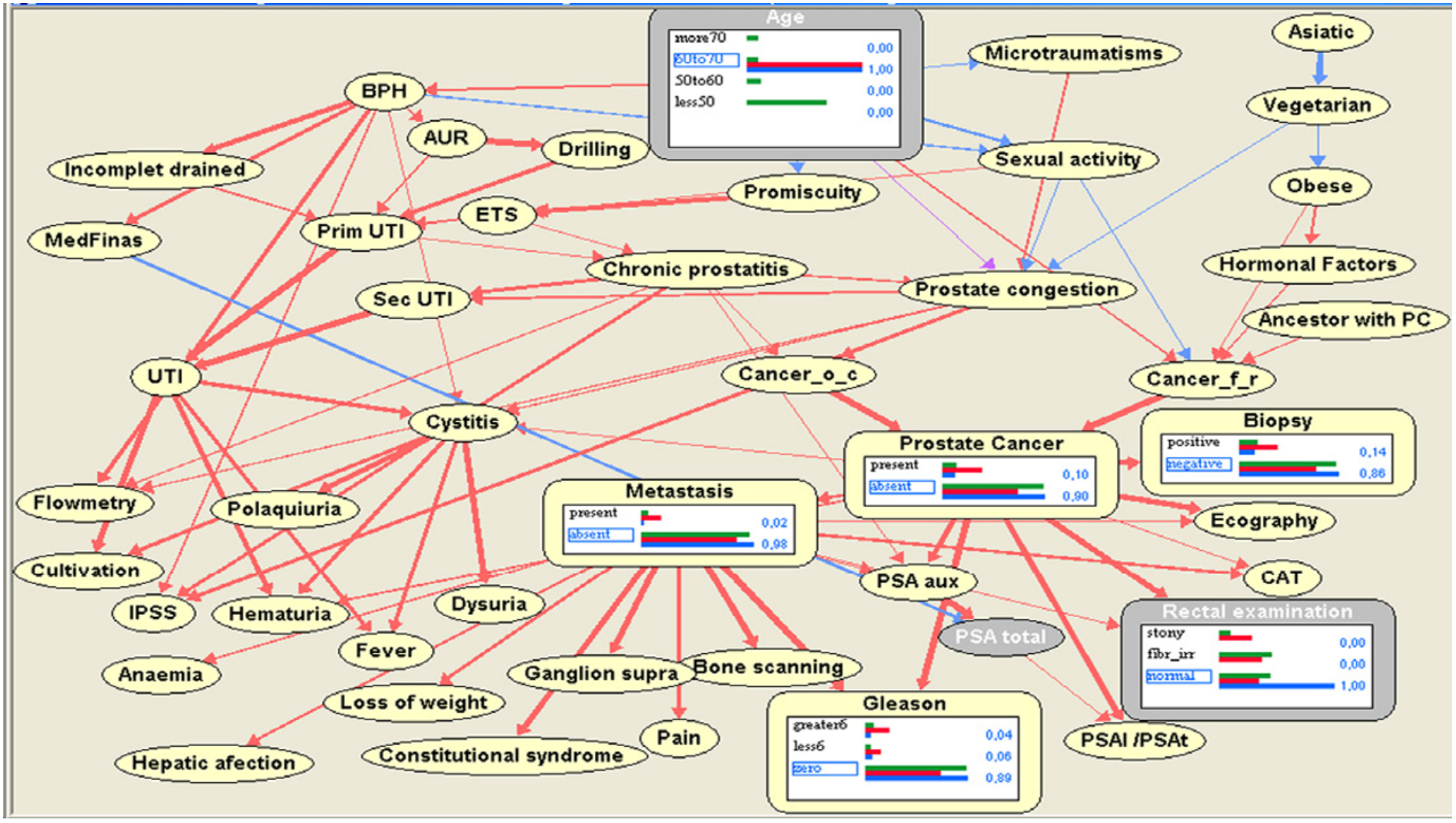
Explanation of decisions supports the following questions:

- "Given the available information, are we ready to make a decision?", and if not
- " What additional information do we require to make an informed decision?"

using threshold-based solutions:

- SDP: probability that same decision is made upon obtaining additional evidence (2012 –)
- sensitivity analysis: to what extent does the outcome depend on the specified conditional probabilities? (1995 –)

Explanation of reasoning: monotonicity (visual)



Img: *Explanation of Bayesian Networks and Influence Diagrams in Elvira* (C. Lacave, M. Luque, F.J. Díez, IEEE Trans., 2007)

Explanation of reasoning: scenarios (textual)

1991:

```
The following scenario(s) are
compatible with cold:
A. Cold and no cat hence no
   allergy                          0.47
   Other less probable
   scenario(s)                      0.06

The following scenario(s) are
incompatible with cold:
B. No Cold and cat causing
   allergy                          0.48

Scenario A is about as likely as
scenario B (0.47/0.48)
because cold in A is a great deal
less likely than no cold in B
(0.08/0.92),
although no cat in A is a great deal
more likely than cat in B (0.9/0.1).

Therefore cold is slightly more
likely than not (p=0.52).
```

2016:

Scenario 2: Sylvia and Tom committed the burglary. (prior probability: 0.0001, posterior probability: 0.2326)

> **Scenario: Sylvia and Tom committed the burglary**: Sylvia and Tom had debts and a window was already broken. Then, Sylvia and Tom climbed through the window. Then, Tom stole a laptop.

Scenario 2 is complete and consistent. It contains the evidential gap 'Sylvia and Tom had debts' and the supported implausible element 'A window was already broken'.

Evidence for and against scenario 2:

* Broken window: moderate evidence to support scenario 2.
* Statement: Tom sold laptop: moderate evidence to support scenario 2.
* Testimony: window was already broken: weak evidence to support scenario 2.
* All evidence combined: very strong evidence to support scenario 2.

1991: *Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning* (M. Henrion, M.J. Druzdzel, UAI)
2016: *When stories and numbers meet in court* (C.S. Vlek, PhD Thesis, RUG)

Explanation of reasoning: relevance of evidence

1997:

Before presenting any evidence, the probability of GALLSTONES being present is 0.128.

The following pieces of evidence are considered important (in order of importance):

- Presence of GUARDING results in a posterior probability of 0.175 for GALLSTONES.
- AGE of 41 results in a posterior probability of 0.172 for GALLSTONES.

Their influence flows along the following paths:

- GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
- AGE influences GALLSTONES.

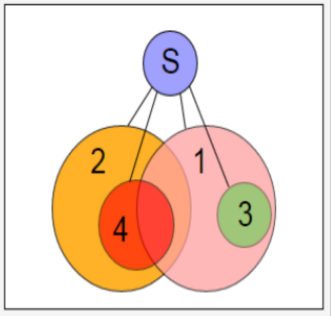
Presentation of the evidence results in a posterior probability of 0.227 for the presence of GALLSTONES.

2015:

The value **scirrheus** of node **Shape** is certain (P = 1.00).

We were able to construct four arguments based on the evidence associated with the value **scirrheus** for node **Shape (S)**.
The arguments are ordered by how influential they are for the value of the node **Shape (S)**

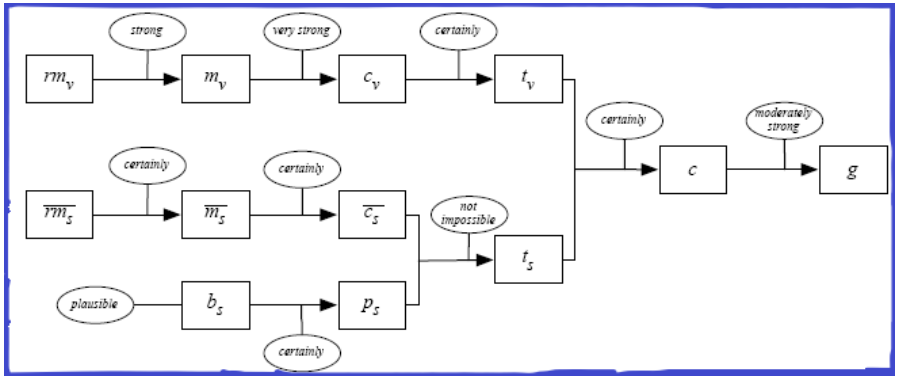
- Argument 1: Node **Endosono-mediast** has value **no**
  Node **Bronchoscopy** has value **no**
  Node **Lapa-diagragm** has value **no**
  Node **CT-organs** has value **none**
  Node **X-fistula** has value **no**
  Node **CT-liver** has value **no**
  Node **X-lungs** has value **no**
  Node **CT-lungs** has value **no**
  Node **Endosono-wall** has value **T3**
- Argument 2: Node **Gastro-shape** has value **scirrheus**
  Node **Gastro-circumf** has value **circulair**
  Node **Gastro-length** has value **5 <= x < 10**
  Node **Weightloss** has value **x<10%**
  Node **Endosono-wall** has value **T3**
  Node **Endosono-truncus** has value **non-determ**
  Node **Endosono-loco** has value **yes**
  Node **Gastro-necrosis** has value **no**
  Node **X-fistula** has value **no**
  Node **Endosono-mediast** has value **no**
  Node **Gastro-location** has value **distal**
- Argument 3: Node **Gastro-shape** has value **scirrheus**
- Argument 4: Node **X-fistula** has value **no**
  Node **Gastro-necrosis** has value **no**



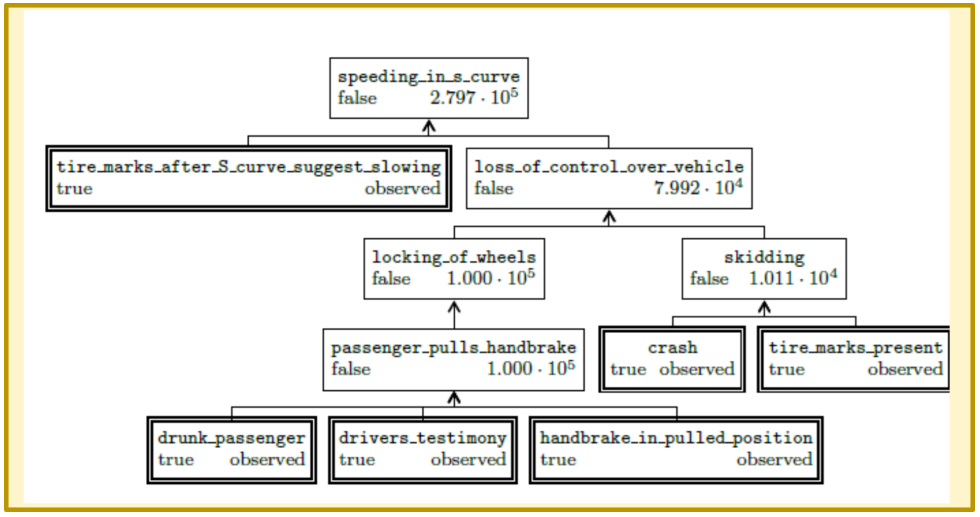
1997: *BANTER: a Bayesian network tutoring shell* (P. Haddawy, J. Jacobson, Ch.E. Kahn Jr., AI in Med.)
2015: *Explaining the reasoning of Bayesian networks with intermediate nodes and clusters* (J. van Leersum, MSc Thesis, UU)

Explanation of reasoning: argument graphs

2011:



2017:



2011: *On extracting arguments from Bayesian network representations of evidential reasoning* (J. Keppens, ICAIL)
2017: *Designing and understanding forensic Bayesian networks using argumentation* (S.T. Timmer, PhD Thesis, UU)

Persuasive contrastive explanation
(explanation of reasoning: classification)

Consider evidence e, resulting in output t instead of t'.

A persuasive contrastive explanation combines

- sufficient explanation s
  ▶ *minimal* sub-configuration of evidence e that suffices for concluding t, regardless of the values for $E \setminus S$

  " evidence s would already be enough to conclude t "

- counterfactual explanation c
  ▶ *minimal* sub-configuration of unobserved values $\overline{e}$ that in combination with the remaining evidence for $E \setminus C$ suffices to conclude t'

  " t' would result if the evidence contains c instead "

Persuasive contrastive Explanations for Bayesian networks (T. Koopman, S. Renooij, ECSQARU 2021)

Explanation support: MAP-independence

Recall: MAP $h^* = \arg\max_{\boldsymbol{h}} P(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{E} = \boldsymbol{e})$.

h^* is MAP-*independent* of subset $\boldsymbol{R}$ of intermediate variables, if for all $\boldsymbol{r}$: (Kwisthout, 2021)

$$\operatorname*{argmax}_{h'} \Pr(h' \wedge \boldsymbol{r} \mid \boldsymbol{e}) = h^*$$

If $\operatorname{argmax} h' \neq h^*$ for some $\boldsymbol{r}$ then

- $\boldsymbol{r}$ provides for a *"counterfactual"*;
- that *contrasts* outputs h^* and h'.

Note that the explanation concerns the effects of possible future observations rather than current!

Explainable AI using MAP-independence (J. Kwisthout, ECSQARU 2021)

Relevance for Robust Bayesian Network MAP-Explanations (S. Renooij, PGM 2022)

Interactive explanation

S	Based on the given evidence, what is your expected value for *Acccident*?
U	I expect a moderate accident.
S	Based on the given evidence, it is found that *severe* is the most likely value for *Accident*. Do you prefer understandable or convincing explanations for this finding?
U	I prefer understandable explanations.
S	We found the following sufficient and counterfactual explanations:
	1. If only $Age = adolescent, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence. If $Model = normal$ was observed instead of the actual values and all other values would stay the same, *moderate* would be the most probable value for Accident
	2. If only $Mileage = 2000, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence. [+ same counterfactual]
S	Do you want to see more explanations?
U	No.
S	On what observation is your expectation for a moderate accident based?
U	On the observation that I have a cautious driving style.
S	You probable have the right expectations of how E_i influences the target. However, this influence was suppressed by the other observations.

sufficient

counterfactual

Computing contrastive, counterfactual explanations for Bayesian networks (T. Koopman, MSc. Thesis, UU, 2020)

Explanation: what & how

- structure alone
- probabilistic relations in the graph
  - signs on arcs (QPNs), thickness of arcs
- relation between evidence and outcome
  - reasoning chains: from graphs, verbal explanations (text), arguments
  - sufficient and counterfactual explanations
- evidence itself
  - MAP/MPE (= a configuration of maximum probability)
  - conflict / surprise
- outcome distribution/probability
  - verbal explanation: text + verbal probability expression

Any widely adopted solutions after 30 years? No. . .

(but see MSc thesis by J.R. Koiter for examples)

Syllabus, Chapter 7:

Conclusions

Concluding observations about P(G)Ms

The state of the art as far as Probabilistic (Graphical) Models are concerned is as follows:

- P(G)Ms and their associated algorithms offer a useful framework for representing and manipulating probabilistic information;
- the framework combines mathematical correctness with expressiveness and efficiency;
- advances in research enable and facilitate applicability of P(G)Ms in increasingly more practical situations;
- P(G)Ms are becoming more and more important due to their interpretability.

Current Research into P(G)Ms

Research aims mostly at supporting their practical application:

- approximate inference;
- learning from data;
- confounding variables, causality, and interventions;
- representation and manipulation of continuous distributions;
- representation and manipulation of time;
- incremental model-construction;
- relevance of variables, values, arcs and probabilities;
- model-complexity vs accuracy;
- model-checking and repairing;
- design of methods for knowledge acquisition and explanation;
- building actual applications;
- design of software for builders and users;
- ...

Interested in more?

For further information on research on the subject of this course, see:

- links on the course website, also for info about graduation projects;
- (proceedings of) the annual UAI conference on Uncertainty in Artificial Intelligence;
- (online proceedings of) the BMAW workshop linked to UAI: Bayesian Modeling Applications Workshop (more);
- (proceedings of) the bi-annual PGM conference on Probabilistic Graphical Models;
- authors' homepages
- ...

What's next?

- opportunity to ask your remaining questions about the course
- the exam
  - see www.cs.uu.nl/docs/vakken/prob/beoordeling.html for details
  - see studymanual for expectations
- Please fill out the Caracal course evaluation!