Probabilistic Reasoning

Uncertainty and Bayesian networks

Peter Lucas

IRIS
Institute for Computing and Information Sciences
University of Nijmegen, The Netherlands

Uncertainty in Daily Life

- **Empirical evidence:**

  "If symptoms of fever, shortness of breath (dyspnoea), and coughing are present, and the patient has recently visited China, then the patient has *probably* SARS"

- **Subjective belief:**

  "The Balkenende II government is *likely* to resign soon"

- **Temporal dimension:**

  "There is less than *10% chance* that the Dutch economy will recover in the next two years"

Uncertainty Representation and Manipulation

- Methods for dealing with uncertainty are not new:
  - 17th century: Fermat, Pascal, Huygens, Leibniz, Bernoulli
  - 18th century: Laplace, De Moivre, Bayes
  - 19th century: Gauss, Boole

  ⇒ you could have contributed too if you had been around
- Most important research question in early AI (1970–1987):
  - How to incorporate uncertainty reasoning in logical deduction?

  ⇒ you could have contributed too if you had been around

Early AI Methods of Uncertainty

- **Rule-based uncertainty representation:**

  $$\textit{fever} \wedge \textit{dyspnoea} \Rightarrow \text{SARS}_{\text{CF}=0.4}$$

- **Uncertainty calculus** (certainty-factor (CF) model, subjective Bayesian method):
  - $\text{CF}(\textit{fever}, B) = 0.6$; $\text{CF}(\textit{dyspnoea}, B) = 1$ (B is background knowledge)
  - Combination functions:

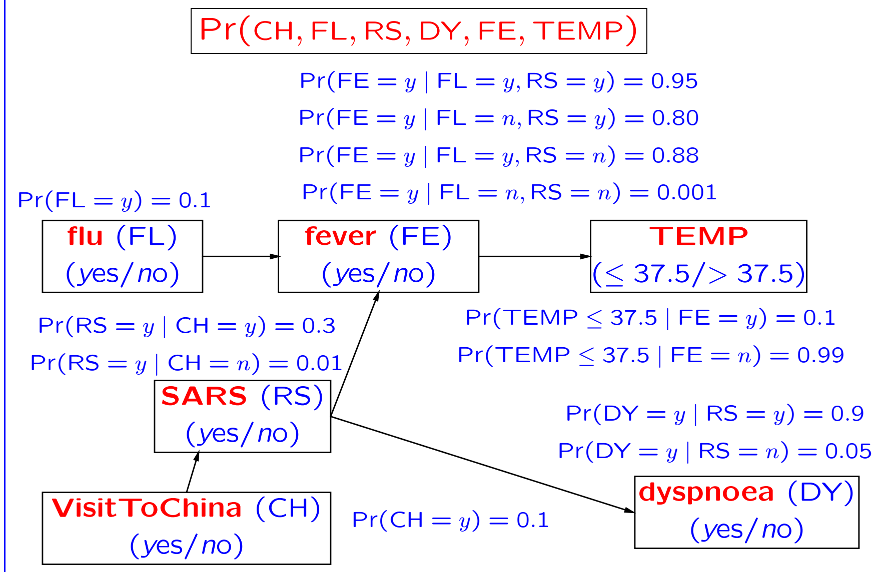
    $$\begin{aligned}\text{CF}(\text{SARS}, \{\textit{fever}, \textit{dyspnoea}\} \cup B) &= 0.4 \cdot \max\{0, \min\{\text{CF}(\textit{fever}, B), \text{CF}(\textit{dyspnoea}, B)\}\} \\ &= 0.4 \cdot \max\{0, \min\{0.6, 1\}\} = 0.24\end{aligned}$$

However · · ·

fever ∧ *dyspnoea* ⇒ $\text{SARS}_{\text{CF}=0.4}$

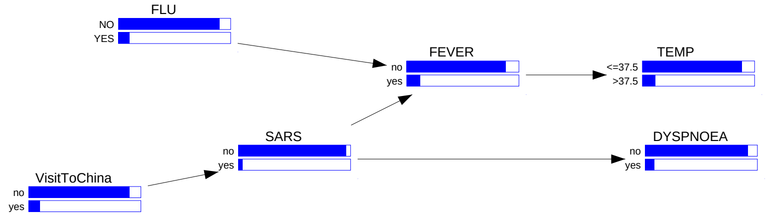
- How likely is the occurrence of *fever* or *dyspnoea* given that the patient has *SARS*?
- How likely is the occurrence of *fever* or *dyspnoea* in the absence of *SARS*?
- How likely is the presence of *SARS* when just *fever* is present?
- How likely is *no SARS* when just *fever* is present?

Bayesian Networks

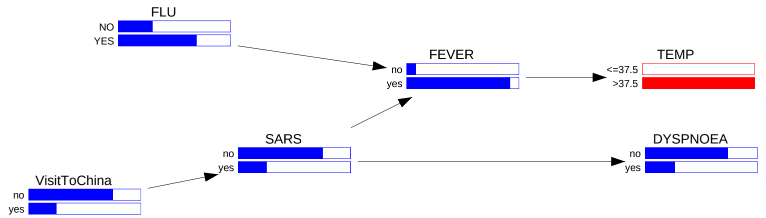


Reasoning: Evidence Propagation

- **Nothing known:**

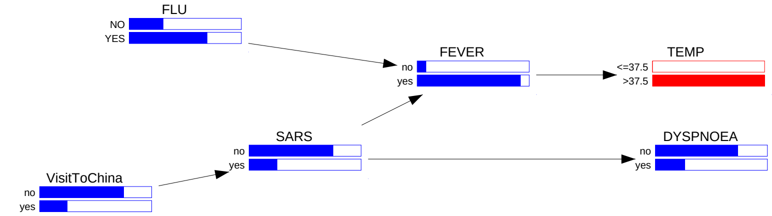


- **Temperature >37.5 °C:**

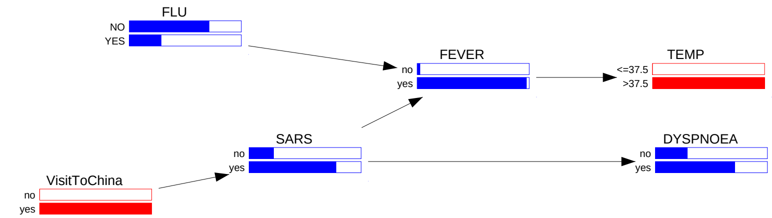


Reasoning: Evidence Propagation

- **Temperature >37.5 °C:**



- **I just returned from China:**



Bayesian Network Formally

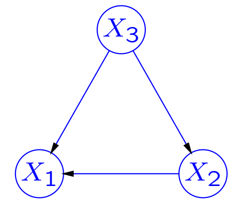
A Bayesian network (BN) is a pair $\mathcal{B} = (G, \text{Pr})$, where:

- $G = (V(G), A(G))$ is an acyclic directed graph, with
  - $V(G) = \{X_1, X_2, \ldots, X_n\}$, a set of vertices (nodes); $X \in V(G)$ corresponds to a random variable X
  - $A(G) \subseteq V(G) \times V(G)$ a set of arcs reflecting (conditional) independences among variables
- $\text{Pr} : \wp(V(G)) \to [0,1]$ is a joint probability distribution, such that

  $$\text{Pr}(V(G)) = \prod_{i=1}^{n} \text{Pr}(X_i \mid \pi_G(X_i))$$

  where $\pi_G(X_i)$ denotes the set of immediate ancestors (parents) of vertex X_i in G

Factorisation



Conditional probability distribution:

$$\text{Pr}(X_1 \mid X_2, X_3) = \frac{\text{Pr}(X_1, X_2, X_3)}{\text{Pr}(X_2, X_3)}$$

$$\Rightarrow \text{Pr}(X_1, X_2, X_3) = \text{Pr}(X_1 \mid X_2, X_3)\,\text{Pr}(X_2 \mid X_3)\,\text{Pr}(X_3)$$

Chain rule yields a factorisation:

$$\text{Pr}(\bigwedge_{i=1}^{n} X_i) = \prod_{i=1}^{n} \text{Pr}(X_i \mid \bigwedge_{k=i+1}^{n} X_k)$$

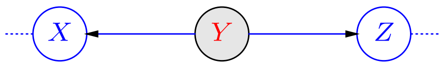
Independence Representation in Graphs

The set of variables X is conditionally independent of the set Z *given* the set Y, notation $X \perp\!\!\!\perp Z \mid Y$, iff

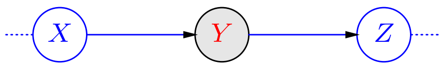
$$\boxed{\text{Pr}(X \mid Y, Z) = \text{Pr}(X \mid Y)}$$

Three flavours of graph-representation of (in)dependence:

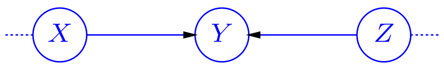
Diverging: Y **blocks** X and Z: $X \perp\!\!\!\perp Z \mid Y$



Serial: Y **blocks** X and Z: $X \perp\!\!\!\perp Z \mid Y$



Converging: Y **connects** X and Z: $X \not\perp\!\!\!\perp Z \mid Y$



Use of Independence Information

General:

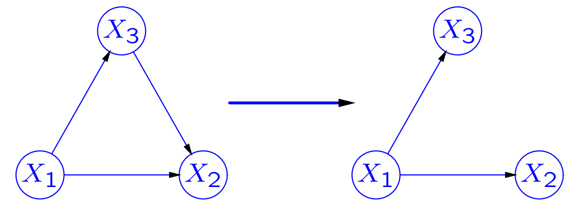
$$\text{Pr}(X_1, X_2, X_3) = \text{Pr}(X_2 \mid X_1, X_3)\,\text{Pr}(X_3 \mid X_1)\,\text{Pr}(X_1)$$

Assume that $X_2 \perp\!\!\!\perp X_3 \mid X_1$, then:

$$\text{Pr}(X_2 \mid X_1, X_3) = \text{Pr}(X_2 \mid X_1)$$

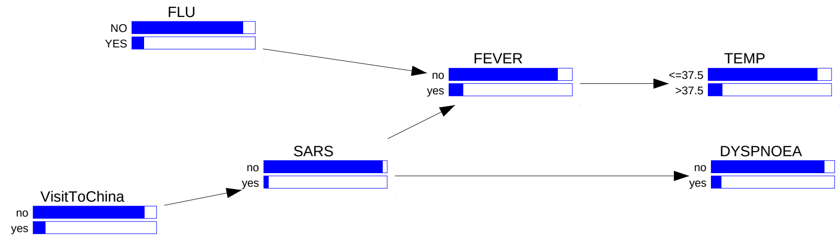
and

$$\text{Pr}(X_3 \mid X_1, X_2) = \text{Pr}(X_3 \mid X_1)$$



Only $5 = 2 + 2 + 1$ probabilities needed for $\text{Pr}(X_1, X_2, X_3)$ (instead of 7)

Find the Independences



Examples:

- FLU $\perp\!\!\!\perp$ VisitToChina $\mid \varnothing$
- FLU $\perp\!\!\!\perp$ SARS $\mid \varnothing$
- FLU $\not\perp\!\!\!\perp$ SARS | FEVER, also FLU $\not\perp\!\!\!\perp$ SARS | TEMP
- SARS $\perp\!\!\!\perp$ TEMP | FEVER
- VisitToChina $\perp\!\!\!\perp$ DYSPNOEA | SARS

Probabilistic Reasoning

- Interested in marginal probability distributions:

  $$\Pr(V_i \mid \mathcal{E}) = \Pr^{\mathcal{E}}(V_i)$$

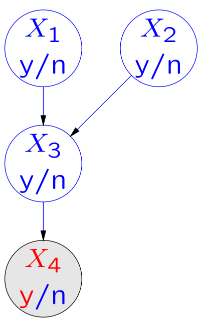
  for (possibly empty) evidence $\mathcal{E}$ (instantiated variables)
- Joint probability distribution $\Pr(V)$:
  - marginalisation:

    $$\Pr(\mathcal{W}) = \sum_{V \setminus \mathcal{W}} \Pr(V) = \sum_{V \setminus \mathcal{W}} \prod_{X \in V} \Pr(X \mid \pi(X))$$

  - conditional probabilities and Bayes' rule:

    $$\Pr(Y, Z \mid X) = \frac{\Pr(X \mid Y, Z)\Pr(Y, Z)}{\Pr(X)}$$
- Many efficient Bayesian reasoning algorithms exist

Naive Probabilistic Reasoning: Evidence



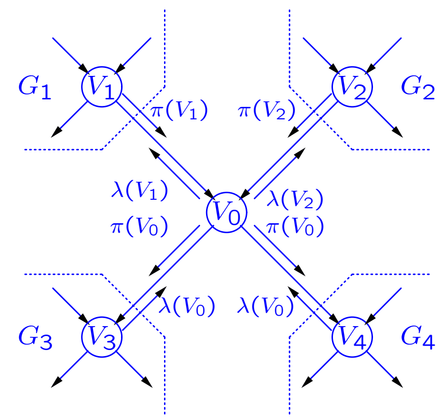
$\Pr(x_4 \mid x_3) = 0.4$
$\Pr(x_4 \mid \neg x_3) = 0.1$
$\Pr(x_3 \mid x_1, x_2) = 0.3$
$\Pr(x_3 \mid \neg x_1, x_2) = 0.5$
$\Pr(x_3 \mid x_1, \neg x_2) = 0.7$
$\Pr(x_3 \mid \neg x_1, \neg x_2) = 0.9$
$\Pr(x_1) = 0.6$
$\Pr(x_2) = 0.2$

$$\Pr^{\mathcal{E}}(x_2) = \Pr(x_2 \mid x_4) = \frac{\Pr(x_4 \mid x_2)\Pr(x_2)}{\Pr(x_4)} \text{ (Bayes' rule)}$$

$$= \frac{\sum_{X_3} \Pr(x_4|X_3) \sum_{X_1} \Pr(X_3|X_1, x_2)\Pr(X_1)\Pr(x_2)}{\sum_{X_3} \Pr(x_4 \mid X_3) \sum_{X_1, X_2} \Pr(X_3 \mid X_1, X_2)\Pr(X_1)\Pr(X_2)}$$

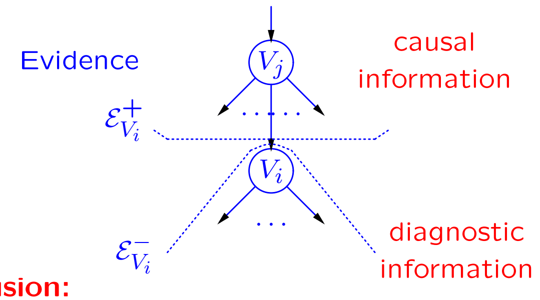
$$\approx 0.14$$

Judea Pearl's Algorithm



- Object-oriented approach: vertices are objects, which have local information and carry out local computations
- Updating of probability distribution by message passing: arcs are communication channels

Data Fusion Lemma



Data fusion:

$$\Pr{}^{\mathcal{E}}(V_i) = \Pr(V_i \mid \mathcal{E})$$
$$= \alpha \cdot \text{causal info for } V_i \cdot \text{diagnostic info for } V_i$$
$$= \alpha \cdot \pi(V_i) \cdot \lambda(V_i)$$

where:

- $\mathcal{E} = \mathcal{E}^+_{V_i} \cup \mathcal{E}^-_{V_i}$: evidence
- α: normalisation constant

Problem Solving

Bayesian networks are declarative, i.e.:

- mathematical basis
- problem to be solved determined by (1) entered evidence $\mathcal{E}$ (may include decisions); (2) given hypothesis H:

  $\Pr(H \mid \mathcal{E})$ (cf. $\text{KB} \wedge H \vDash \mathcal{E}$)

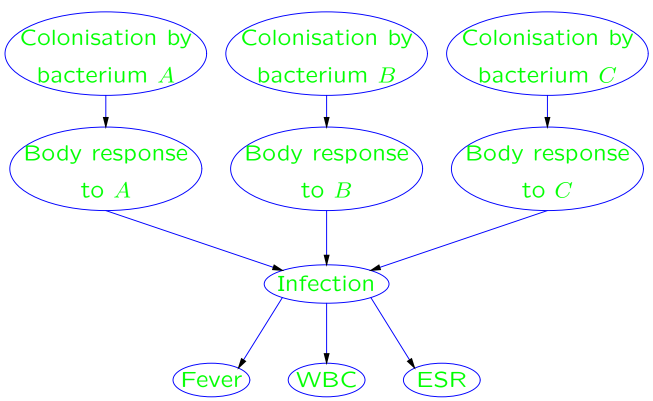
Examples:

- Description of populations
- Classification and diagnosis: $D = \arg\max_H \Pr(H \mid \mathcal{E})$
- Temporal reasoning, prediction, *what-if* scenarios
- Decision-making based on decision theory

$$\text{MEU}(D \mid \mathcal{E}) = \max_{d \in D} \sum_{x \in X_{\pi(U)}} u(x) \Pr(x \mid d, \mathcal{E})$$

Manual Construction of Bayesian Networks

Qualitative modelling:



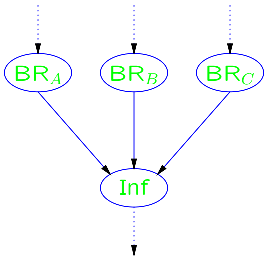
People become colonised by bacteria when entering a hospital, which may give rise to infection

Bayesian-network Modelling

Qualitative — *causal modelling*

Cause → Effect

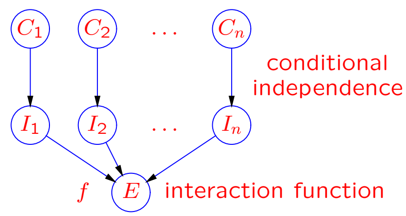
Quantitative — *interaction modelling*



$\Pr(\text{Inf} \mid \text{BR}_A, \text{BR}_B, \text{BR}_C)$

	BR_A							
	t				f			
	BR_B				BR_B			
	t		f		t		f	
	BR_C		BR_C		BR_C		BR_C	
Inf	t	f	t	f	t	f	t	f
t	0.8	0.6	0.5	0.3	0.4	0.2	0.3	0.1
f	0.2	0.4	0.5	0.7	0.6	0.8	0.7	0.9

Causal Independence

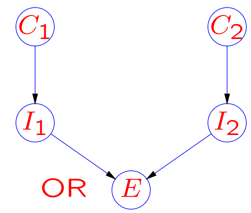


$$\Pr(e \mid C_1,\ldots,C_n) = \sum_{I_1,\ldots,I_n} \Pr(e \mid I_1,\ldots,I_n) \prod_{k=1}^{n} \Pr(I_k \mid C_k)$$
$$= \sum_{f(I_1,\ldots,I_n)=e} \prod_{k=1}^{n} \Pr(I_k \mid C_k)$$

Boolean functions: $P(E \mid I_1,\ldots,I_n) \in \{0,1\}$

Interaction function f, defined such that $f(I_1,\ldots,I_n) = e$ if $P(e \mid I_1,\ldots,I_n) = 1$

Example: noisy OR

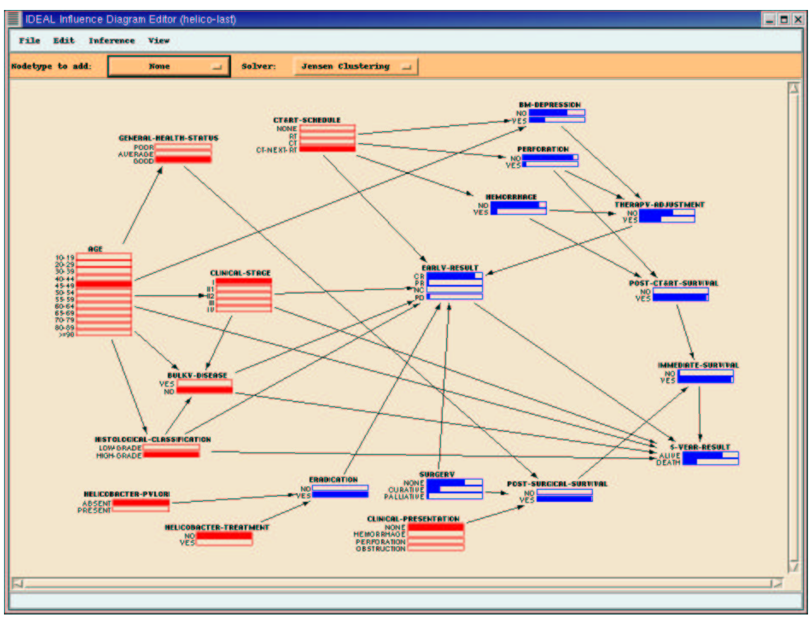


- Interactions among causes: logical OR
- Meaning: presence of any one of the causes C_i with absolute certainty will cause the effect e (i.e. $E = \mathit{true}$)

$$\Pr(e|C_1,C_2) = \sum_{I_1 \vee I_2 = e} \Pr(e|I_1,I_2) \prod_{k=1,2} \Pr(I_k \mid C_k)$$
$$= \Pr(i_1|C_1)\Pr(i_2|C_2) + \Pr(\neg i_1|C_1)\Pr(i_2|C_2) + \Pr(i_1|C_1)\Pr(\neg i_2|C_2)$$

- Assessment of $O(n)$ instead of $O(2^n)$ probabilities

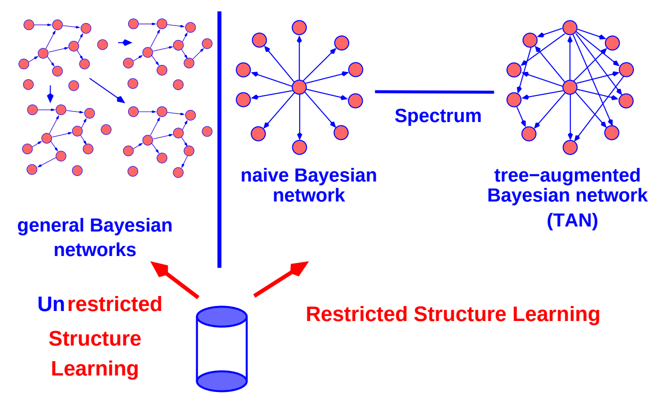
Example BN: non-Hodgkin Lymphoma



Bayesian Network Learning

Bayesian network $\mathcal{B} = (G, \Pr)$, with

- digraph $G = (V(G), A(G))$, and
- probability distribution Pr



Learning Bayesian Networks

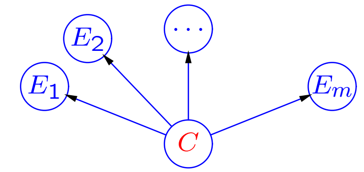
Problems:

- for many BNs too many probabilities have to be assessed
- complex BNs do not necessarily yield better classifiers
- complex BNs may yield better estimates of a probability distribution

Solution:

- use simple probabilistic models for classification:
  - *naive* (independent) form BN
  - *T*ree-*A*ugmented Bayesian *N*etwork (TAN)
  - *F*orest-*A*ugmented Bayesian *N*etwork (FAN)
- use background knowledge and clever heuristics

Naive (independent) form BN



- C is a **class variable**
- The **evidence variables** E_i in the evidence $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$ are conditionally independent given the class variable C

This yields:

$$P(C \mid \mathcal{E}) = \frac{P(\mathcal{E} \mid C)P(C)}{P(\mathcal{E})} = \frac{\prod_{E \in \mathcal{E}} P(E \mid C)}{\sum_C P(\mathcal{E} \mid C)P(C)}$$

as $E_i \perp\!\!\!\perp E_j \mid C$, for $i \neq j$

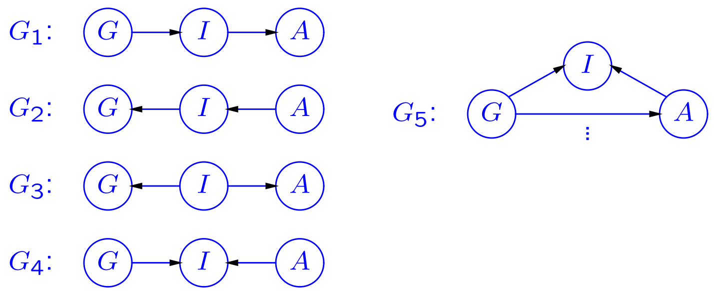
Classifier: $c_{\max} = \arg\max_C P(C \mid \mathcal{E})$

Learning Structure from Data

Given the following dataset D:

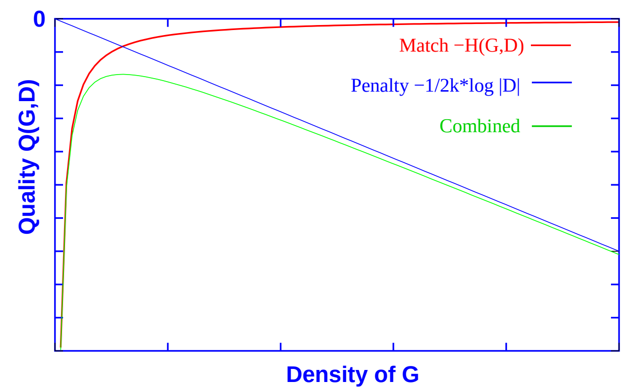
Student	Gender	IQ	High Mark for Maths
1	male	low	no
2	female	average	yes
3	male	high	yes
4	female	high	yes

and the following Bayesian networks:



Which one is the best?

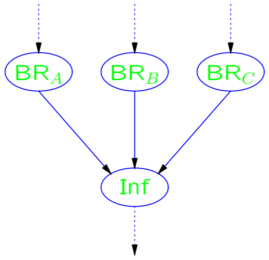
Quality Measure Q



$Q(G, D) = \log \Pr(G) - |D| \cdot H(G, D) - \frac{1}{2}k \cdot \log |D|$, where:

- $\Pr(G)$: prior probability of G
- $-H(G, D)$: negative value of match
- $-\frac{1}{2}k \cdot \log |D|$: penalty term

Research Issues



Qualitative modelling:

- To determine the structure of a network
- Assessment of $\Pr(V_i \mid \pi(V_i))$
- Enhancement of logical semantics

Learning

- Structure learning: determine the 'best' graph topology
- Parameter learning: determine the 'best' probability distribution (discrete or continuous)

$\Rightarrow$ you can contribute too $\cdots$