Machine Intelligence

Lecture 5: Bayesian networks

Thomas Dyhre Nielsen

Aalborg University

Tentative course overview

Topics:

- Introduction
- Search-based methods
- Constrained satisfaction problems
- Logic-based knowledge representation
- Representing domains endowed with uncertainty.
- **Bayesian networks**
- Machine learning
- Planning
- Reinforcement learning
- Multi-agent systems

Bayesian Networks

Example

Random variables (all Boolean):

Tampering	fire alarm has been tampered with
Fire	fire in the building
Alarm	fire alarm ringing
Smoke	smoke in the building
Leaving	people leaving the building
Report	report of people leaving the building

Joint distribution according to chain rule:

$$
\begin{aligned}
&P(\mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}, \mathit{Smoke}, \mathit{Leaving}, \mathit{Report}) = \\
&\quad P(\mathit{Tampering}) \cdot \\
&\quad P(\mathit{Fire} \mid \mathit{Tampering}) \cdot \\
&\quad P(\mathit{Alarm} \mid \mathit{Tampering}, \mathit{Fire}) \cdot \\
&\quad P(\mathit{Smoke} \mid \mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}) \cdot \\
&\quad P(\mathit{Leaving} \mid \mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}, \mathit{Smoke}) \cdot \\
&\quad P(\mathit{Report} \mid \mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}, \mathit{Smoke}, \mathit{Leaving})
\end{aligned}
$$

Example continued

Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$

$$P(\textit{Fire} \mid \textit{Tampering}) =$$

Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$

$$P(\textit{Fire} \mid \textit{Tampering}) = P(\textit{Fire})$$

$$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) =$$

Example continued

Conditional independence assumptions

$P(\textit{Tampering}) \; = P(\textit{Tampering})$

$P(\textit{Fire} \mid \textit{Tampering}) \; = \; P(\textit{Fire})$

$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) \; = \; P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire})$

$P(\textit{Smoke} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}) \; =$

Example continued

Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$

$$P(\textit{Fire} \mid \textit{Tampering}) = P(\textit{Fire})$$

$$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) = P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire})$$

$$P(\textit{Smoke} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}) = P(\textit{Smoke} \mid \textit{Fire})$$

$$P(\textit{Leaving} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}) =$$

Example continued

Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$

$$P(\textit{Fire} \mid \textit{Tampering}) = P(\textit{Fire})$$

$$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) = P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire})$$

$$P(\textit{Smoke} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}) = P(\textit{Smoke} \mid \textit{Fire})$$

$$P(\textit{Leaving} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}) = P(\textit{Leaving} \mid \textit{Alarm})$$

$$P(\textit{Report} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}, \textit{Leaving}) =$$

Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$

$$P(\textit{Fire} \mid \textit{Tampering}) = P(\textit{Fire})$$

$$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) = P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire})$$

$$P(\textit{Smoke} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}) = P(\textit{Smoke} \mid \textit{Fire})$$

$$P(\textit{Leaving} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}) = P(\textit{Leaving} \mid \textit{Alarm})$$

$$P(\textit{Report} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}, \textit{Leaving}) = P(\textit{Report} \mid \textit{Leaving})$$

Example continued

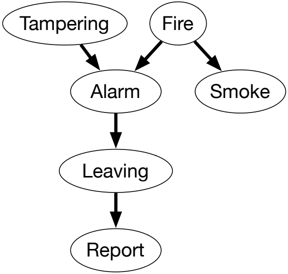
Conditional independence assumptions

$$P(\textit{Tampering}) = P(\textit{Tampering})$$
$$P(\textit{Fire} \mid \textit{Tampering}) = P(\textit{Fire})$$
$$P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) = P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire})$$
$$P(\textit{Smoke} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}) = P(\textit{Smoke} \mid \textit{Fire})$$
$$P(\textit{Leaving} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}) = P(\textit{Leaving} \mid \textit{Alarm})$$
$$P(\textit{Report} \mid \textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}, \textit{Leaving}) = P(\textit{Report} \mid \textit{Leaving})$$

Pluggin this into chain rule give simplified representation of joint distribution:

$$\begin{aligned} &P(\textit{Tampering}, \textit{Fire}, \textit{Alarm}, \textit{Smoke}, \textit{Leaving}, \textit{Report}) = \\ &\qquad P(\textit{Tampering}) \cdot P(\textit{Fire}) \cdot P(\textit{Alarm} \mid \textit{Tampering}, \textit{Fire}) \cdot P(\textit{Smoke} \mid \textit{Fire}) \cdot \\ &\qquad\qquad P(\textit{Leaving} \mid \textit{Alarm}) \cdot P(\textit{Report} \mid \textit{Leaving}) \end{aligned}$$

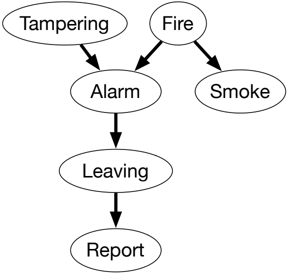
Representation of conditional dependencies in a graph:



The graph is **directed** and **acyclic**.

Graphical Representation

Representation of conditional dependencies in a graph:



The graph is **directed** and **acyclic**.

Bayesian Network

A **Bayesian Network** for variables $A_1, \dots, A_k$ consists of

- a directed acyclic graph with nodes $A_1, \dots, A_k$
- for each node a **conditional probability table** specifying the conditional distribution $P(A_i \mid \textit{parents}(A_i))$ (*parents*(A_i) denotes the **parents** of A_i in the graph)

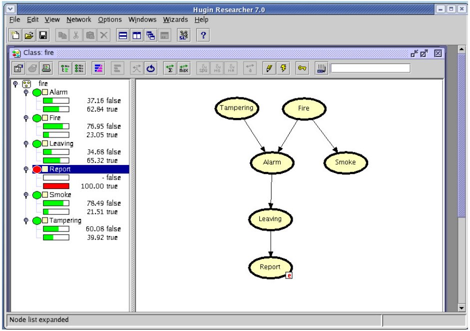
and through the chain rule provides a compact representation of a joint probability distribution.

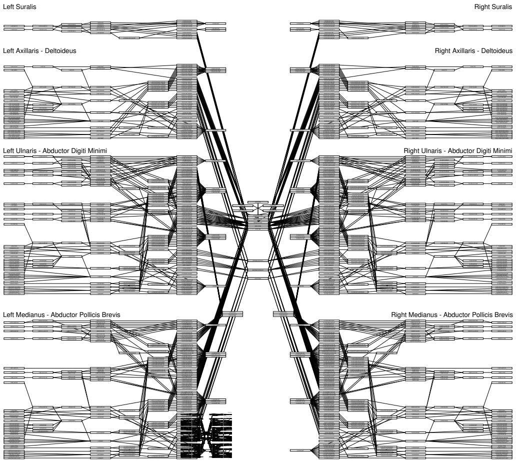
A Bayesian network specifies a joint distribution from which arbitrary conditional probabilities can be derived.

Inference

The most common task is to compute the posterior distribution over a query variable A given the observed values of some evidence nodes $E_i = e_i$, for $i = 1, \ldots, l$:

$$P(A \mid E_1 = e_1, \ldots, E_l = e_l).$$





Characteristics:

- Approximately 1100 variables.
- Each variable has between 2 and 20 states.
- 10^{600} possible state configurations!

A system for diagnosing neuro-muscular diseases.

Constructing a Bayesian Network

Construction via chain rule

1. put the random variables in some order
2. write the joint distribution using chain rule
3. simplify conditional probability factors by conditional independence assumptions. That determines the *parents* of each node, i.e. the graph structure
4. specify the conditional probability tables

Note: the structure of the resulting network strongly depends on the chosen order of the variables.

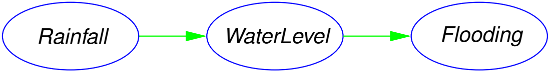
Construction via causality

- Draw and edge from variable A to variable B if A has a direct causal influence on A.

Note: this may not always be possible:

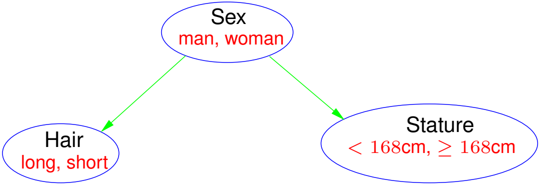
- *Inflation* $\rightarrow$ *salaries* or *salaries* $\rightarrow$ *Inflation* ?
- *Rain* doesn't cause *Sun*, and *Sun* doesn't cause *Rain*, but they are not independent either!

Transmission of evidence

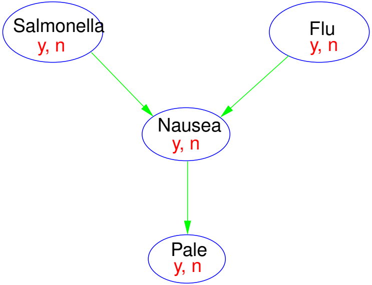


- If there has been a flooding does that tell me something about the amount of rain that has fallen?
- The water level is high: If there has been a flooding does that tell me anything new about the amount of rain that has fallen?

Reasoning under uncertainty 2

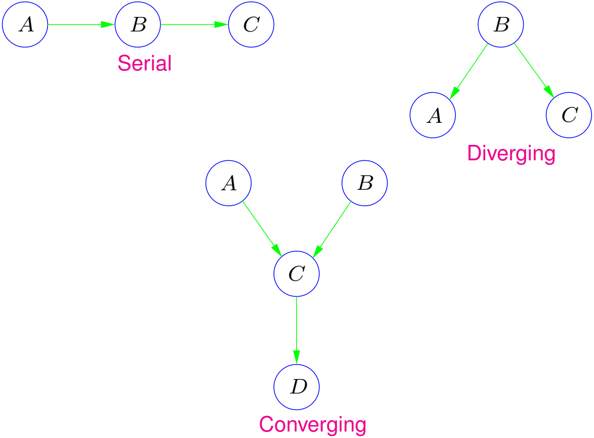


- If a person has long hair does that say something about his/her stature?
- It is a woman: If she has long hair does that say something about her stature?

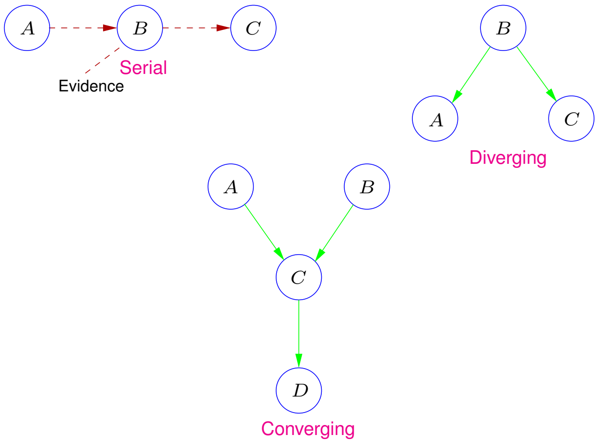


- Does salmonella have an impact on Flu?
- If a person is Pale, does salmonella then have an impact on Flu?

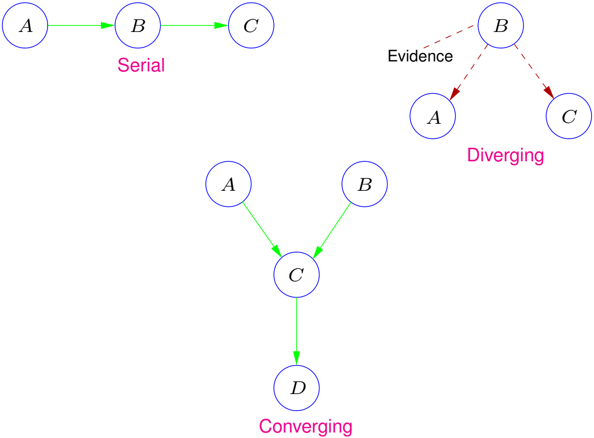
Relevance changes with evidence



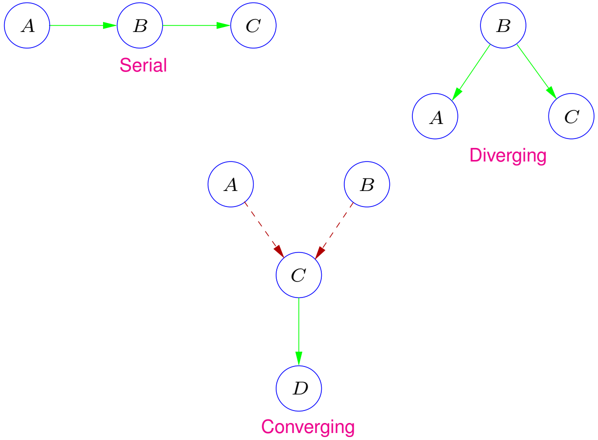
Relevance changes with evidence



Relevance changes with evidence

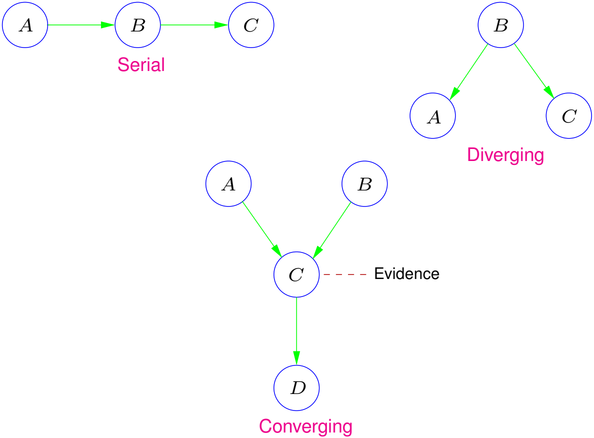


Relevance changes with evidence

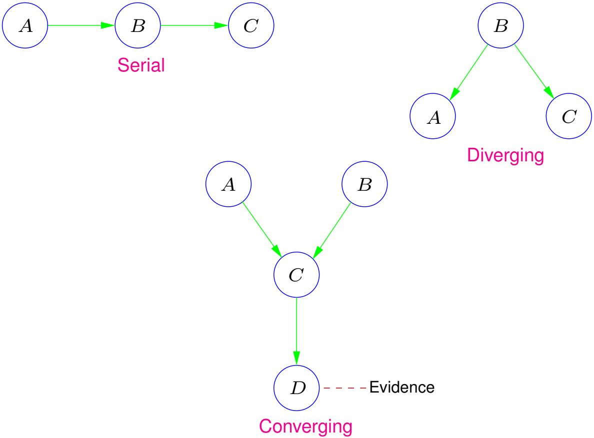


Transmission of evidence 1

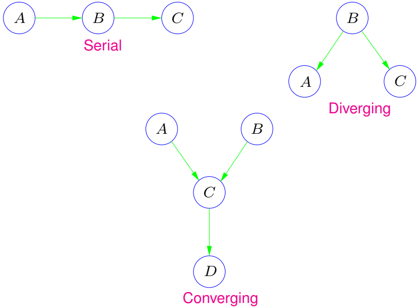
Relevance changes with evidence



Relevance changes with evidence

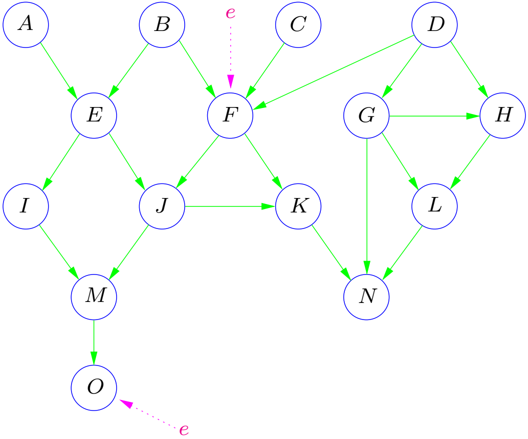


Summary of transmission rules (d-separation rules)

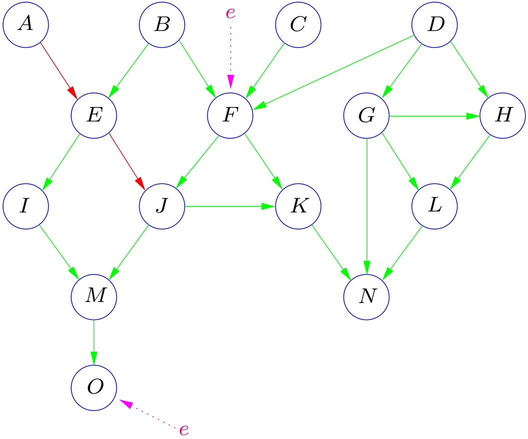


Rules for transmission of evidence

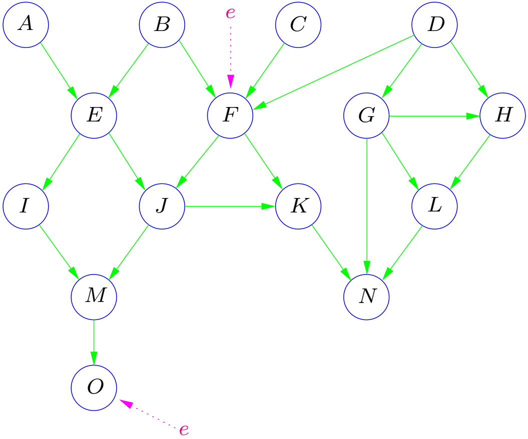
- Evidence may be transmitted through a serial or diverging connection unless it is instantiated.
- Evidence may be transmitted through a converging connection only if either the variable in the connection or one of its descendants has received evidence.



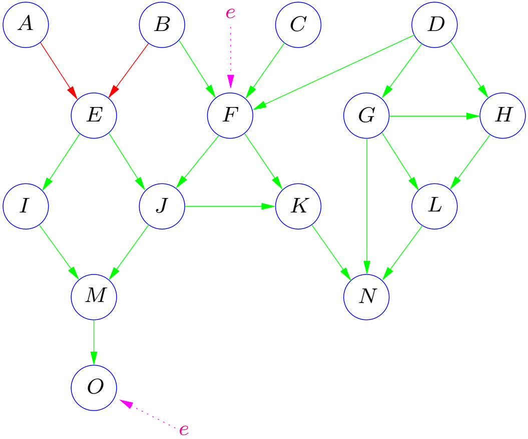
Can knowledge of A have an impact on our knowledge of J?



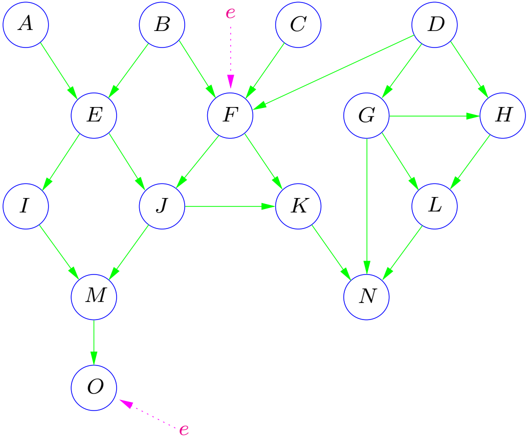
Can knowledge of A have an impact on our knowledge of J? yes!



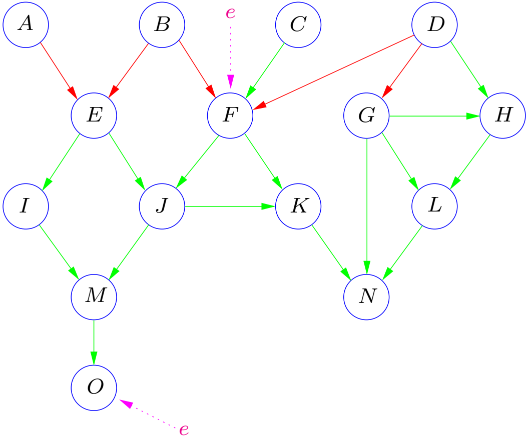
Can knowledge of A have an impact on our knowledge of B?



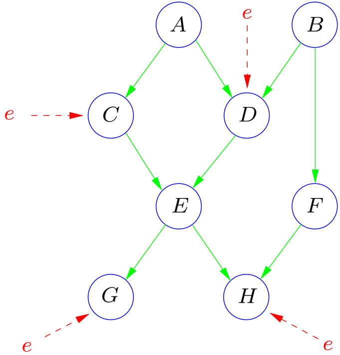
Can knowledge of A have an impact on our knowledge of B? yes!



Can knowledge of A have an impact on our knowledge of G?



Can knowledge of A have an impact on our knowledge of G? yes!



Is E d-separated from A?

The d-separation theorem

Theorem

For all pairwise disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$ of nodes in a Bayesian network:

If $\mathbf{C}$ *d-separates* $\mathbf{A}$ *from* $\mathbf{B}$*, then* $P(\mathbf{A} \mid \mathbf{B}, \mathbf{C}) = P(\mathbf{A} \mid \mathbf{C})$.

There are no more general graphical conditions than d-separation for which such a result holds.

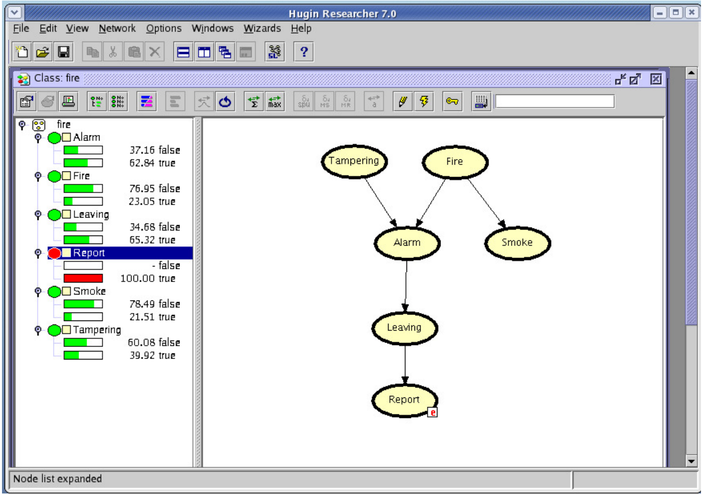
Use of d-Separation

Why is d-separation important?

- Gaining insight: given a (correct) Bayesian network model, can derive insight into the dependencies among the variables
- Debugging a model: given a Bayesian network model, check whether entailed independence relations are plausible
- Correctness of algorithms: certain computational procedures depend on validity of special independence relations

Fire Example

The Fire example in the HUGIN Bayesian Network system (`http://www.hugin.com/`)

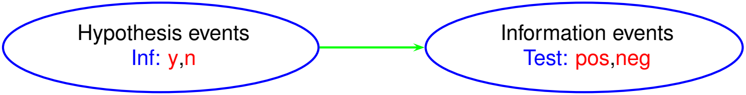


Specifying the structure of a Bayesian network

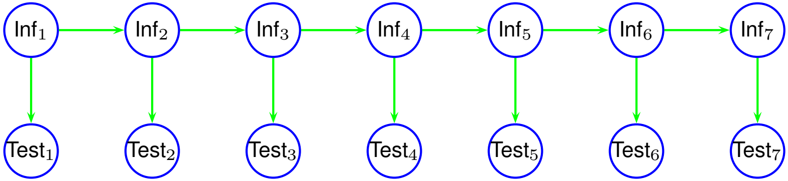
Milk from a cow may be infected. To detect whether or not the milk is infected, you can apply a test which may either give a positive or a negative test result. The test is not perfect: It may give false positives as well as false negatives.

Building models

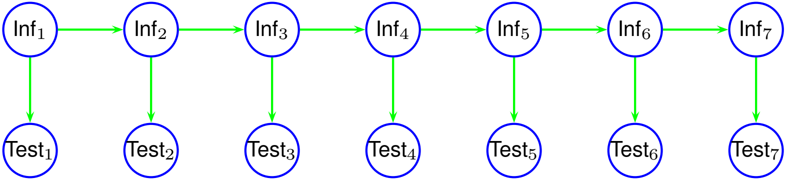
Milk from a cow may be infected. To detect whether or not the milk is infected, you can apply a test which may either give a positive or a negative test result. The test is not perfect: It may give false positives as well as false negatives.



Infections develop over time:



Infections develop over time:



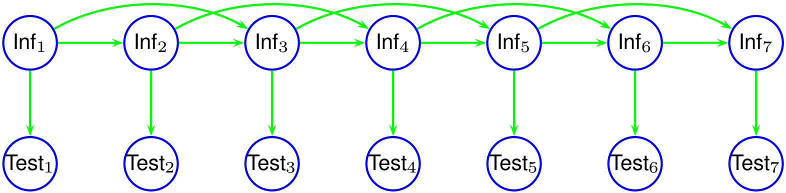
Assumption

The Markov property: If I know the present, then the past has no influence on the future:

Inf_{i-1} is d-separated from Inf_{i+1} given Inf_i.

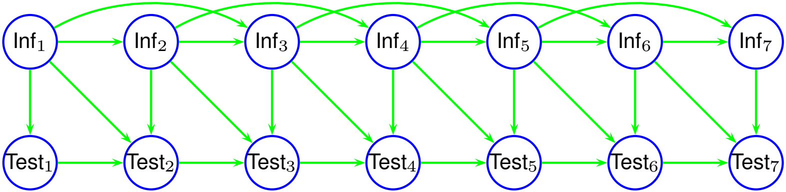
But what if yesterday's Inf-state has an impact on tomorrow's Inf-state?

Non-Markov relations



Yesterday's Inf-state has an impact on tomorrow's Inf-state.

Relations between observations



The test-failure is dependent on whether or not the test failed yesterday.

Sore throat

I wake up one morning with a sore throat. It may be the beginning of a cold or I may suffer from angina. If it is a severe angina, then I will not go to work. To gain more insight, I can take my temperature and look down my throat for yellow spots.

Sore throat

I wake up one morning with a sore throat. It may be the beginning of a cold or I may suffer from angina. If it is a severe angina, then I will not go to work. To gain more insight, I can take my temperature and look down my throat for yellow spots.

Hypothesis variables

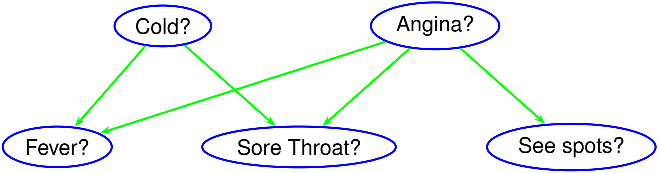
Cold? - {n, y}
Angina? - {no, mild, severe}

Information variables

Sore throat? - {n, y}
See spots? - {n, y}
Fever? - {no, low, high}

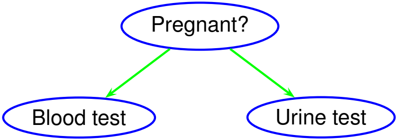
Model for sore throat

Model for sore throat



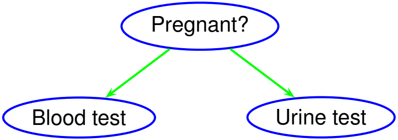
Insemination of a cow

Six weeks after the insemination of a cow, there are two tests: a Blood test and a Urine test.



Insemination of a cow

Six weeks after the insemination of a cow, there are two tests: a Blood test and a Urine test.

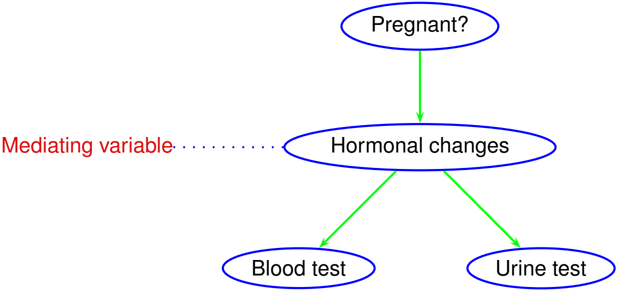


Check the conditional independences

If we know that the cow is pregnant, will a negative blood test then change our expectation for the urine test?

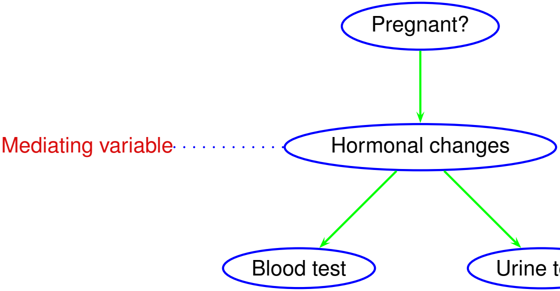
If it will, then the model does not reflect reality!

Insemination of a cow: A more correct model



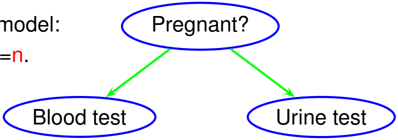
But does this actually make a difference?

Insemination of a cow: A more correct model



But does this actually make a difference?

Assume that both tests are negative in the *incorrect* model:
This will overestimate the probability for Pregnant?=n.

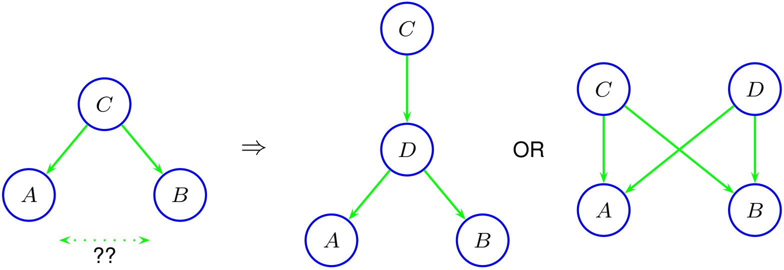


Why mediating variables?

Why do we introduce mediating variables:

- Necessary to catch the correct conditional independences.
- Can ease the specification of the probabilities in the model.

For example: If you find that there is a dependence between two variables A and B, but cannot determine a causal relation: Try with a mediating variable!



A simplified poker game

The game consists of:

- Two players.
- Three cards to each player.
- Two rounds of changing cards (max two cards in the second round)

What kind of hand does my opponent have?

A simplified poker game

The game consists of:

- Two players.
- Three cards to each player.
- Two rounds of changing cards (max two cards in the second round)

What kind of hand does my opponent have?

Hypothesis variable:

OH - {no, 1a, 2v, fl, st, 3v, sf}

Information variables:

FC - {0, 1, 2, 3} and SC - {0, 1, 2}

The game consists of:

- Two players.
- Three cards to each player.
- Two rounds of changing cards (max two cards in the second round)

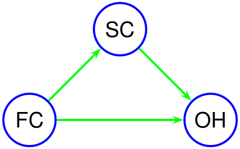
What kind of hand does my opponent have?

Hypothesis variable:

OH - {no, 1a, 2v, fl, st, 3v, sf}

Information variables:

FC - {0, 1, 2, 3} and SC - {0, 1, 2}

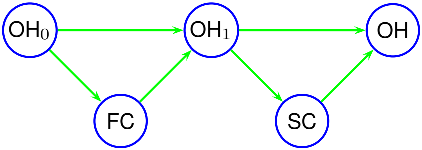


But how do we find:

$P(\text{FC})$, $P(\text{SC}|\text{FC})$ and $P(\text{OH}|\text{SC}, \text{FC})$??

Introduce mediating variables:

- The opponent's initial hand, OH_0.
- The opponent's hand after the first change of cards, OH_1.



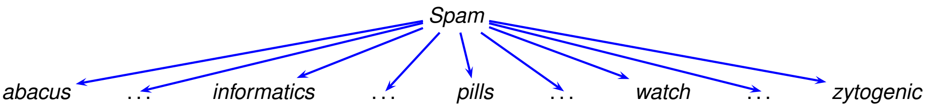
Note

The states of OH_0 and OH_1 are different from OH.

Example: Spam filter

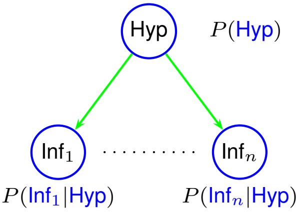
- A single *query variable*: *Spam*
- Many observable features (e.g. words appearing in the body of the message): *abacus*,...,*informatics*, *pills*, ..., *watch*,..., *zytogenic*

Network Structure:



- Inference with large number of variables possible
- Essentially how *Thunderbird* spam filter works

Naïve Bayes models

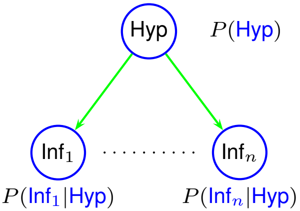


We want the posterior probability of the hypothesis variable Hyp given the observations $\{\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n\}$:

$$P(\text{Hyp}|\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n) = \frac{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})}{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n)}$$

Note: The model assumes that the information variables are independent given the hypothesis variable.

Naïve Bayes models



We want the posterior probability of the hypothesis variable Hyp given the observations $\{\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n\}$:

$$
\begin{aligned}
P(\text{Hyp}|\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n) &= \frac{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})}{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n)} \\
&= \mu \cdot P(\text{Inf}_1 = e_1|\text{Hyp}) \cdot \ldots \cdot P(\text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})
\end{aligned}
$$

Note: The model assumes that the information variables are independent given the hypothesis variable.