Bayesian Networks
Basics

Example Mammography

- To promote the early detection of breast cancer from a certain age, women are recommended to participate regularly in screening (test for women without symptoms). They perform screening in a specific area of the country. In the region concerned, the following data are available for women aged between 40 and 50 who have no symptoms and are participating in mammography screenings.

- The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that her mammogram is positive. However, if a woman has no breast cancer, the probability is 7 percent that her mammogram is still positive. Suppose a woman's mammogram is positive. What is the probability that she actually has breast cancer?

Solution **(A) Probability is 0.09** **or** **(B) Probability is 0.9** **?**

Bayes Analysis

H health states

M observation

$P(H=ill) = 0.008$

$P(M=pos \mid H=ill) = 0.9$

$P(M=pos \mid H=\neg ill) = 0.07$

Bayes Theorem:

$$P(H=ill \mid M=pos) = \frac{7}{77}$$

↑

Probabilist-like

Frequencies

$H \rightarrow M$

$$\begin{array}{ccccccc} & & & 1000 & & & \\ & & \swarrow^{ill} & & \searrow^{\neg ill} & & \\ & 8 & & & & 992 & \\ \swarrow^{pos} & & \searrow^{neg} & & \swarrow^{pos} & & \searrow^{neg} \\ 7 & & 1 & & 70 & & 922 \end{array}$$

$$\frac{\#(ill \wedge pos)}{\# pos} = \frac{7}{7+70}$$

$$\#(ill \mid pos) = \frac{7}{77}$$

↑

Normal-Human-Like

State space

M \ H	ill	¬ill
pos	0.007	0.07
neg	0.001	0.922

$$P(ill \mid pos) = \frac{P(ill, pos)}{P(pos)} = \frac{0{,}007}{0{,}077}$$

↑

Computer-like

The Big Objective(s)

In a wide variety of application fields two main problems need to be addressed:

1. How can (expert) knowledge of complex domains be efficiently represented?
2. How can inferences be carried out within these representations?
3. How can such representations be (automatically) extracted from collected data?
4. How to revise this representation in the light of new knowledge?

We will give some answers to these questions during the lecture.

Example: Planning in car manufacturing

Available information

"Engine type e_1 can only be combined with transmission t_2 or t_5."

"Transmission t_5 requires crankshaft c_2."

"Convertibles have the same set of radio options as SUVs."

Possible questions/inferences:

"Can a station wagon with engine e_4 be equipped with tire set y_6?"

"Supplier S_8 failed to deliver on time. What production line has to be modified and how?"

"Are there any peculiarities within the set of cars that suffered an aircondition failure?"

Example: Medical reasoning

Available information:

“Malaria is much less likely than flu.”

“Flu causes cough and fever.”

“Nausea can indicate malaria as well as flu.”

“Nausea never indicated pneunomia before.”

Possible questions/inferences

“The patient has fever. How likely is he to have Covid-19?”

“How much more likely does flu become if we can exclude Covid-19?”

Common Problems

Both scenarios share some severe problems:

Large Data Space

It is intractable to store all value combinations, i. e. all car part combinations or inter-disease dependencies.

(Example: VW Bora has 10^{200} theoretical value combinations*)

Sparse Data Space

Even if we could handle such a space, it would be extremely sparse, i. e. it would be impossible to find good estimates for all the combinations.

(Example: with 100 diseases and 200 symptoms, there would be about 10^{62} dif- ferent scenarios for which we had to estimate the probability.*)

* The number of particles in the observable universe is estimated to be between 10^{78} and 10^{85}.

Decomposition of a high dimensional Probability Distribution

It is often possible to exploit local constraints (e.g. structural and expert knowledge-based) in a way that allows for a decomposition of the large (intractable) distribution $P(X_1, \ldots, X_n)$ into several sub-structures $\{C_1, \ldots, C_m\}$ such that:

- The collective size of those sub-structures is much smaller than that of the original distribution P.
- The original distribution P is recomposable (with no or at least as few as possible errors) from these sub-structures.

A ***Bayes Network*** (V, E, P) consists of a set $V = \{X_1, \ldots, X_n\}$ of random variables, a set E of directed edges between the variables and a probability.

Each variable has a finite set of mutual exclusive and collectively exhaustive states.The variables in combination with the edges are required to form a **directed, acyclic graph (DAG)**.

Each variable Y with parent nodes $X_1, \ldots, X_m$ is assigned the conditional probability distribution $P(Y \mid X_1, \ldots, X_m)$. These (local) probabilities between these nodes model their connections. They not necessarily express a causal relationship, often it is a stochastic dependency or an association.

The (global) probability is defined by

$$P(V) = \prod_{v \in V : P(c(v)) > 0} P(v \mid c(v))$$

with $c(v)$ being the parent nodes of v.

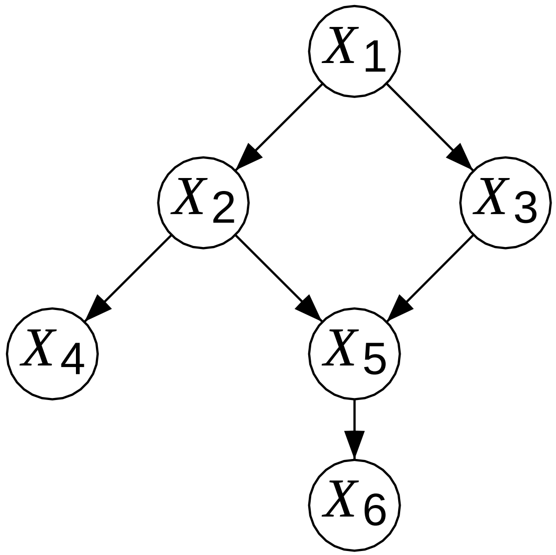
Example 1

For arbitrary random variables $X_1,\ldots,X_n$ the so called „chain rule" holds, e.g.

$$\begin{aligned} P(X_1,\ldots,X_6) = \; & P(X_6 \mid X_5,\ldots,X_1)\cdot \\ & P(X_5 \mid X_4,\ldots,X_1)\cdot \\ & P(X_4 \mid X_3, X_2, X_1)\cdot \\ & P(X_3 \mid X_2, X_1)\cdot \\ & P(X_2 \mid X_1)\cdot \\ & P(X_1) \end{aligned}$$

Example 1

Given a DAG, we define the probability according to the (in)dependency structure



$$
\begin{aligned}
P(X_1, \ldots, X_6) = \; & P(X_6 \mid X_5) \cdot \\
& P(X_5 \mid X_2, X_3) \cdot \\
& P(X_4 \mid X_2) \cdot \\
& P(X_3 \mid X_1) \cdot \\
& P(X_2 \mid X_1) \cdot \\
& P(X_1)
\end{aligned}
$$

Bayes Networks are directed acyclic graphs (DAGs) where the nodes represent random variables and the directed edges model a direct dependence between the connected nodes. The strength of the dependence is defined by conditional probabilities.

Constructing a DAG

For a given probability we can find a suitable DAG

input $P(X_1, \ldots, X_n)$
output a DAG G

1: Set the nodes of G to $\{X_1, \ldots, X_n\}$.

2: Choose a total ordering on the set of variables (e.g. $X_1 \prec X_2 \prec \cdots \prec X_n$)

3: For X_i find the smallest (uniquely determinable) set $S_i \subseteq \{X_1, \ldots, X_n\}$ such that $P(X_i \mid S_i) = P(X_i \mid X_1 \ldots, X_{i-1})$.

4: Connect all nodes in S_i with X_i and store $P(X_i \mid S_i)$ as quantization of the dependencies for that node X_i (given its parents).

5: return G

Example 2

Let a_1, a_2, a_3 be three blood groups and b_1, b_2, b_3 three indications of a blood group test.

Variables:	A (blood group)	B (indication)
Possible values:	$\{a_1, a_2, a_3\}$	$\{b_1, b_2, b_3\}$

Result of a data analysis

$P(\{(a_i, b_j)\})$	b_1	b_2	b_3	$\sum$
a_1	0.64	0.08	0.08	0.8
a_2	0.01	0.08	0.01	0.1
a_3	0.01	0.01	0.08	0.1
$\sum$	0.66	0.17	0.17	1

Model, that explains the situation

$$P(A, B) = P(B \mid A) \cdot P(A)$$

Example 3

Expert Knowledge (cancer clinic)

Metastatic cancer is a possible cause of brain cancer, and an explanation for elevated levels of calcium in the blood. Both phenomena together can explain that a patient falls into a coma. Severe headaches are possibly associated with a brain tumor.

Special Case

A patient has severe headaches.

Question

Will this patient go into a coma?

Example 3: Choice of universe of discourse

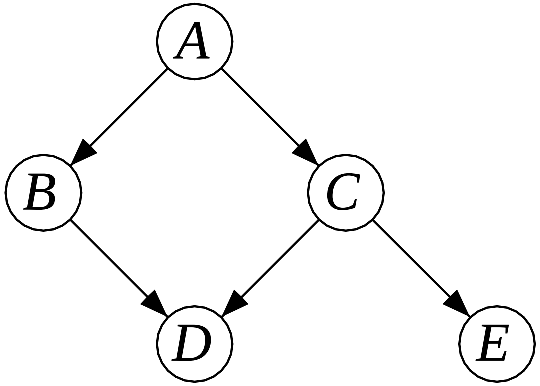
	Variable	Values
A	metastatic cancer	$\{a_1, a_2\}$
B	increased serum calcium	$\{b_1, b_2\}$
C	brain tumor	$\{c_1, c_2\}$
D	coma	$\{d_1, d_2\}$
E	headache	$\{e_1, e_2\}$

(Index 1 means: present, 2 means: absent)

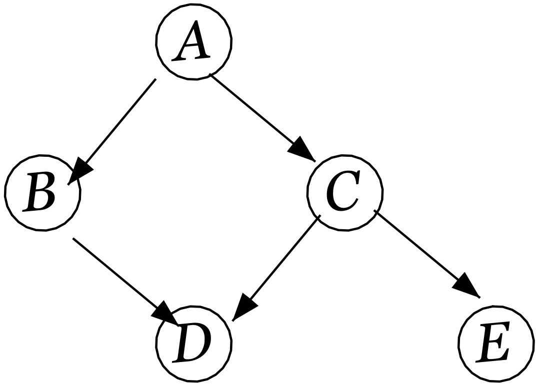
Universe is $\{a_1, a_2\} \times \cdots \times \{e_1, e_2\}$

32 possible values

Analysis of dependencies



Example 3: Definition of the probability space



$$P(a,b,c,d,e) \overset{\text{abbr.}}{=} P(A=a,B=b,C=c,D=d,E=e)$$
$$= P(e \mid c)P(d \mid b,c)P(c \mid a)P(b \mid a)P(a)$$

Shorthand notation

11 values to store instead of 31

Consult experts, textbooks, case studies, surveys, etc.

Example 3: Choice of the conditional probabilities

$$\left.\begin{aligned} P(e_1 \mid c_1) &= 0.8 \\ P(e_1 \mid c_2) &= 0.6 \end{aligned}\right\} \text{headaches common, but more common if tumor present}$$

$$\left.\begin{aligned} P(d_1 \mid b_1, c_1) &= 0.8 \\ P(d_1 \mid b_1, c_2) &= 0.8 \\ P(d_1 \mid b_2, c_1) &= 0.8 \\ P(d_1 \mid b_2, c_2) &= 0.05 \end{aligned}\right\} \text{coma rare but common, if either cause is present}$$

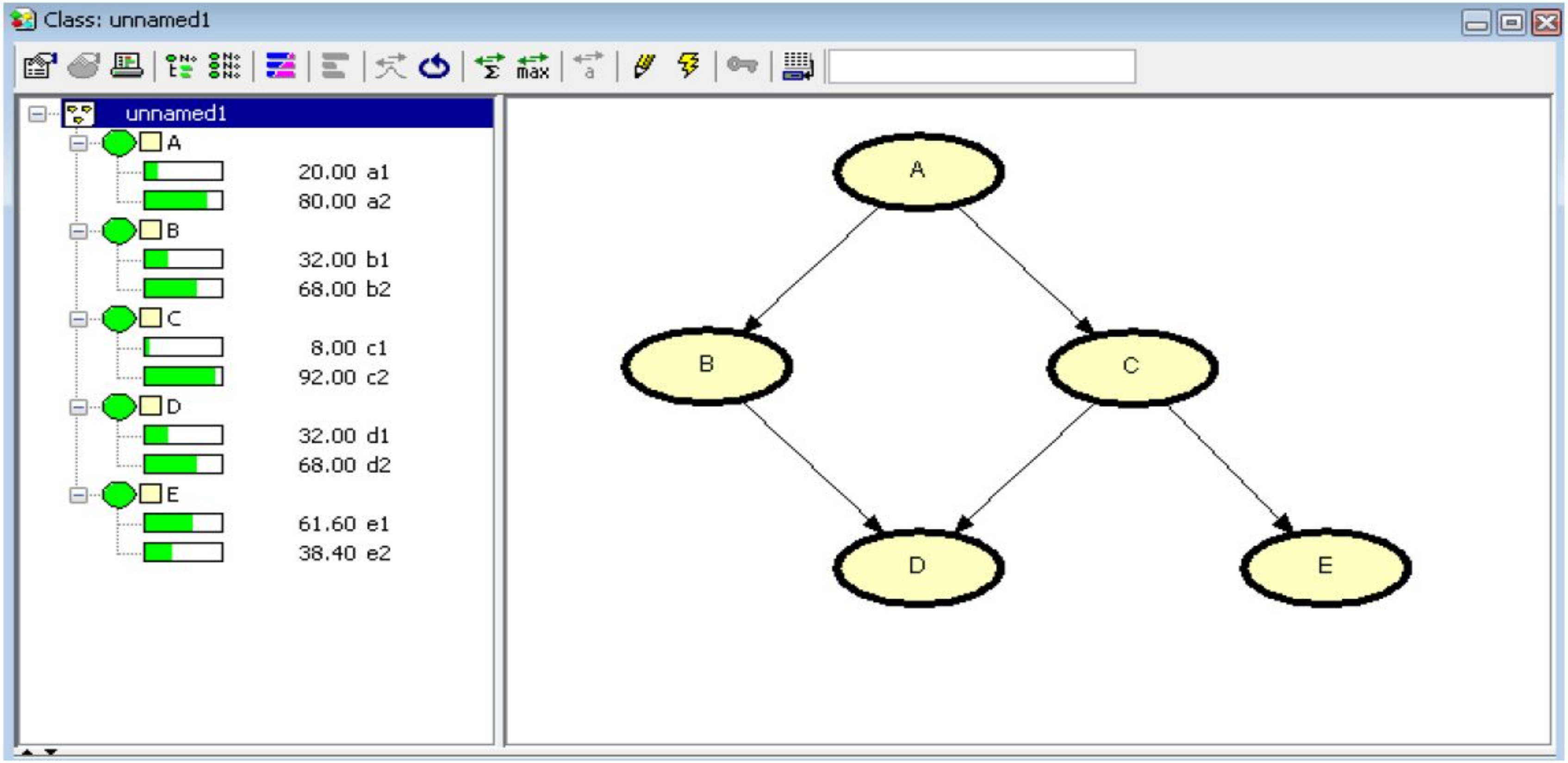
$$\left.\begin{aligned} P(b_1 \mid a_1) &= 0.8 \\ P(b_1 \mid a_2) &= 0.2 \end{aligned}\right\} \begin{array}{l} \text{increased calcium uncommon,} \\ \text{but common consequence of metastases} \end{array}$$

$$\left.\begin{aligned} P(c_1 \mid a_1) &= 0.2 \\ P(c_1 \mid a_2) &= 0.05 \end{aligned}\right\} \text{brain tumor rare, and uncommon consequence of metastases}$$

$$\left.P(a_1) = 0.2 \right\} \text{incidence of metastatic cancer in relevant clinic}$$

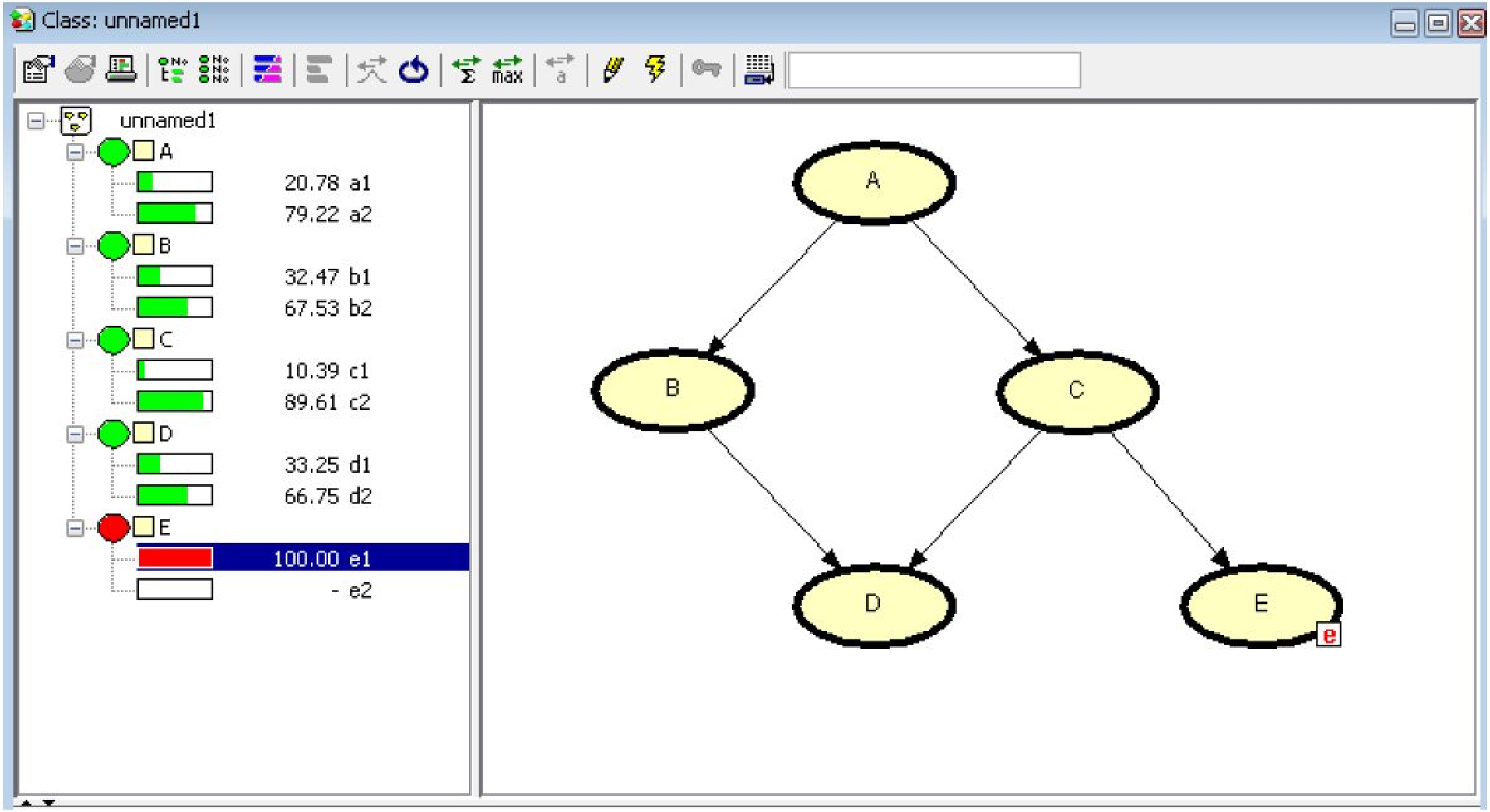
There are many tools for handling BN. Most of them have a graphical user interface.
Free Download of the BN-System Hugin Lite 8.9

https://www.hugin.com/index.php/hugin-lite/



A priori Knowledge about D, Marginal Probabilities: $P(d_1) = 0.3200$,...

Example 3



New Evidence e_1 , Belief Update for D via Conditioning : $P(d_1 \mid e_1) = 0.3325, \ldots$

Crux of the Matter

Knowledge acquisition: Where do the numbers come from?
→ learning methods

Computational complexities: How to handle real problem with 200 attributes?
→ exploit independencies

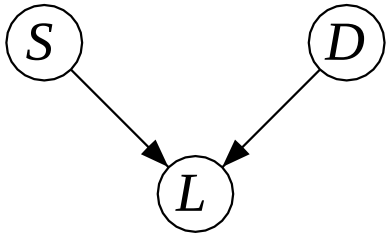
When does an independency of X and Y given Z hold in a Bayes network (V, E, P)?

How to determine a decomposition that fits of the graph structure?

→ study separation in the DAG

Example 4

For each Bayes Net with probability P and the DAG on the right holds:



$$P(S, D, L) \;=\; P(L \mid S, D) \cdot P(S) \cdot P(D)$$

It is easy to prove that S and D are independent:

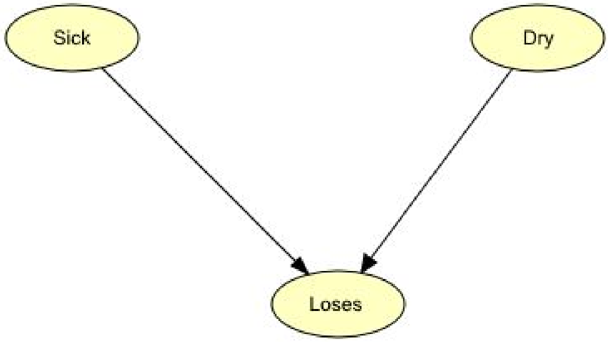
$$P(S, D, L) \;=\; \frac{P(S, D, L)}{P(S, D)} \cdot P(S) \cdot P(D)$$

$$P(S, D) \;=\; P(S) \cdot P(D)$$

On the other hand, it is not possible to prove that S and D are conditionally independent from L

Example 4

A farmer discovers that his finest apple tree is losing its leaves. Now, he wants to know why this is happening. He knows that if the tree is dry (caused by a drought). There is no mystery - it is very common for trees to lose their leaves during a drought. On the other hand the losing of leaves can be an indication of a disease.



Sick = "sick"	**Sick** = "not"
0.1	0.9

Tabel 1: P(Sick)

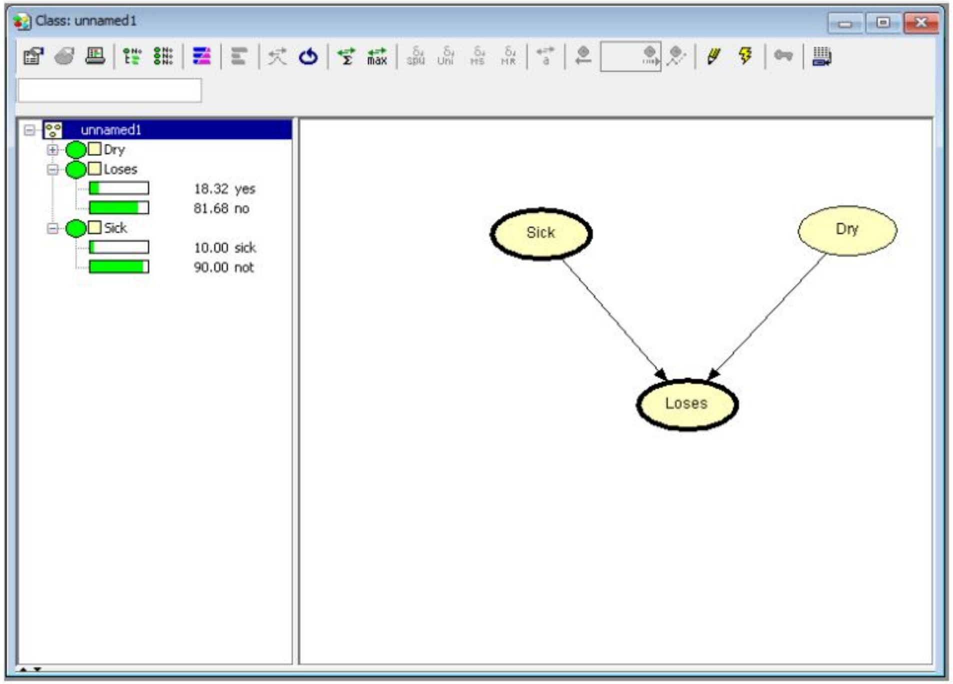
Dry = "dry"	**Dry** = "not"
0.1	0.9

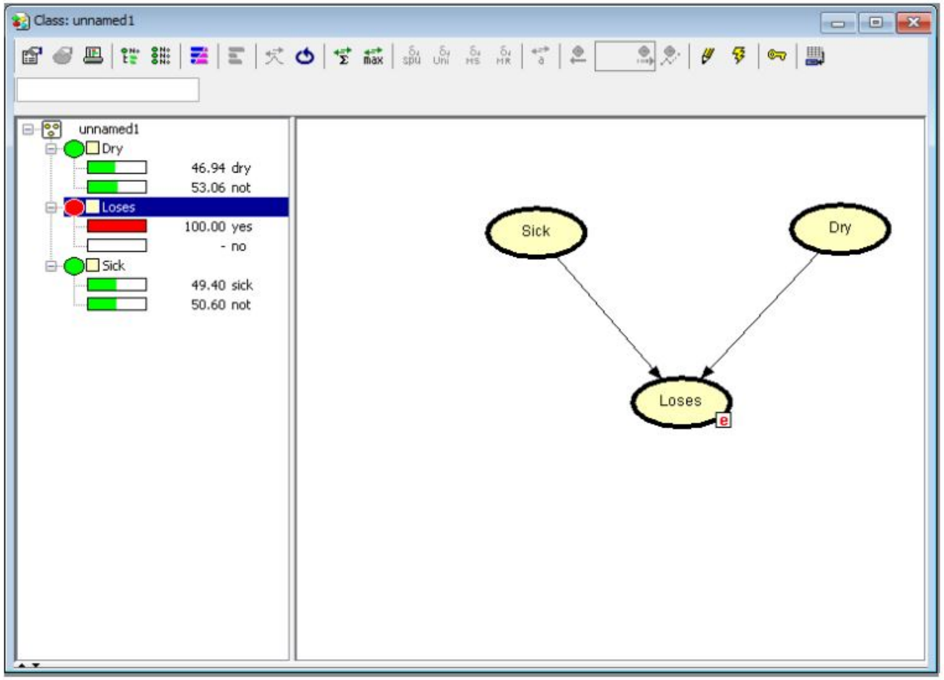
Tabel 2: P(Dry)

	Dry = "dry"		**Dry** = "not"	
	Sick = "sick"	**Sick** = "not"	**Sick** = "sick"	**Sick** = "not"
Loses = "yes"	0.95	0.85	0.90	0.02
Loses = "no"	0.05	0.15	0.10	0.98

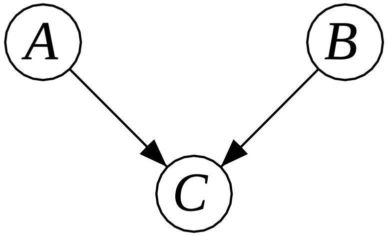
Example 4

P(Sick=yes)= 0.1)
P(Sick=yes I Looses = Yes) = 0,494





Example 5



Meal quality

A quality of ingredients

B cook's skill

C meal quality

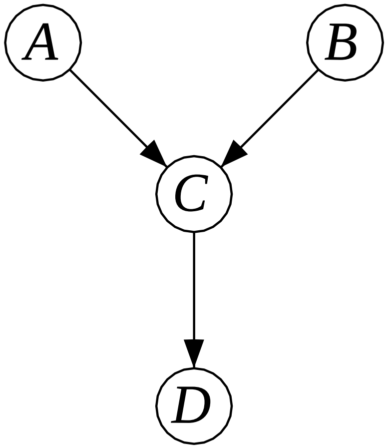
Note that A,B,C are variables!

Intuition:

If C is not known, then A and B ***should*** be independent.

If C is known, then A and B ***should*** become (conditionally) dependent given C.

Example 5 (cont.)



Meal quality	
A	quality of ingredients
B	cook’s skill
C	meal quality
D	restaurant success

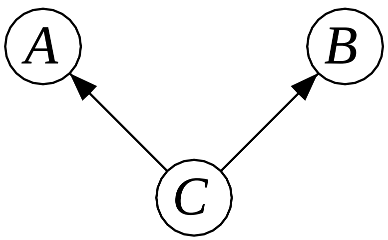
If nothing is known about the restaurant success or meal quality or both, the cook’s skills and quality of the ingredients are unrelated, that is, *independent*.

However, if we observe that the restaurant has no success, we can infer that the meal quality might be bad.

If we further learn that the ingredients quality is high, we will conclude that the cook’s skills must be low, thus rendering both variables *dependent*.

$$P(A,B,C,D) = P(A)P(B)P(C|A,B)P(D|C)$$

Example 6

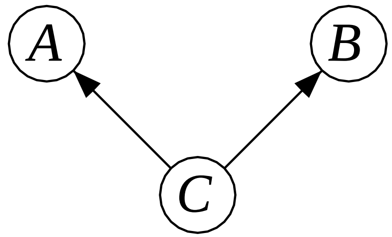


Diagnosis
A body temperature
B cough
C disease

If C is unknown, knowledge about A is relevant for B and vice versa, i.e. A and B are marginally dependent.

However, if C is observed, A and B become conditionally independent given C.

A influences B via C. If C is known it in a way blocks the information from flowing from A to B, thus rendering A and B (conditionally) independent.



Analysis of the corresponding Bayes networks

Decomposition according to the directed acyclic graph:

$$P(A, B, C) = P(A \mid C) \cdot P(B \mid C) \cdot P(C)$$

Embedded Independence:

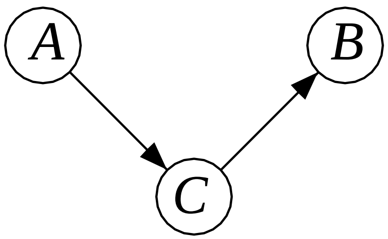
$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

Alternative derivation:

$$P(A, B, C) = P(A \mid C) \cdot P(B, C)$$

$$P(A \mid B, C) = P(A \mid C)$$

Example 7

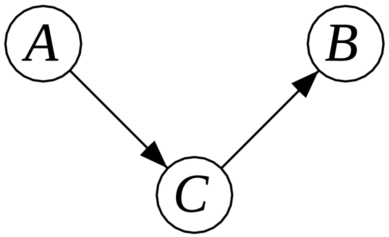


Accidents
A rain
B accident risk
C road conditions

Analog scenario to case 2

A influences C and C influences B. Thus, A influences B. If C is known, it blocks the path between A and B.

Example 7



Decomposition according to graph:

$$P(A, B, C) = P(B \mid C) \cdot P(C \mid A) \cdot P(A)$$

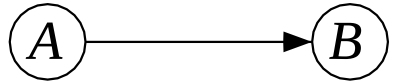
Embedded Independence:

$$P(A, B, C) = P(B \mid C) \cdot P\,(A, C)\; P\,(C)\; P\,(\mathrm{A})/\mathrm{P}(\mathrm{A})$$ (use Bayes Theorem)

$$\mathrm{P}(\mathrm{A},\mathrm{B} \mid \mathrm{C}) = \mathrm{P}(\mathrm{A}|\mathrm{C})\ \mathrm{P}(\mathrm{B}|\mathrm{C})$$

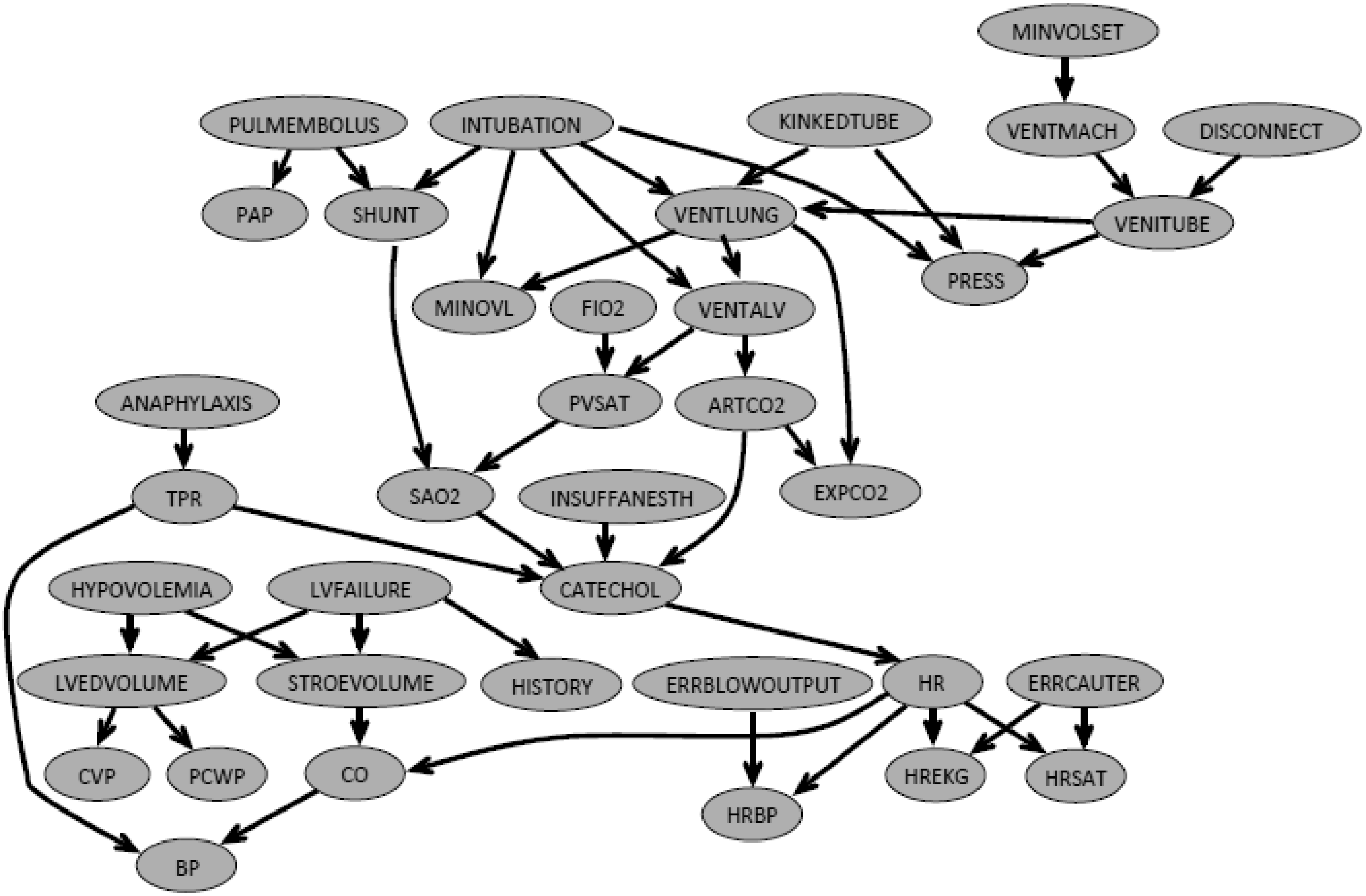
Example 8

$$P(A, B) = P(A) \cdot P(B)$$



$$P(A, B) = P(B \mid A) \cdot P(A)$$

Original joint distribution: 2^{37} parameters Depicted network: 509 par



Graph Theory in necessary to handle such big networks.