Representations of the Joint Probability Distribution

Marek J. Druzdzel

University of Pittsburgh

School of Information Sciences
and Intelligent Systems Program

marek@sis.pitt.edu

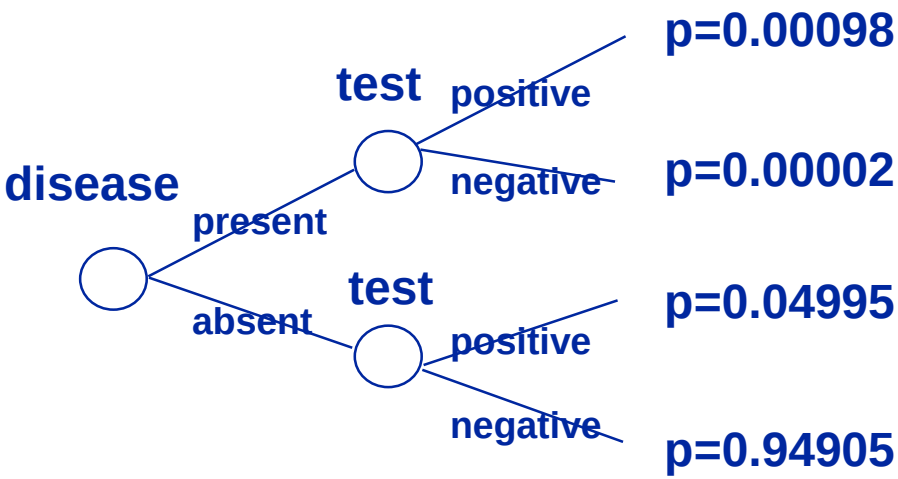
http://www.pitt.edu/~druzdzel

Probabilistic knowledge representations

- **A probabilistic (Bayesian) model encodes the *joint probability distribution* over its variables.**
- **Knowledge of the joint probability distribution is sufficient to derive any marginal and conditional probability over the model's variables.**

Probability trees

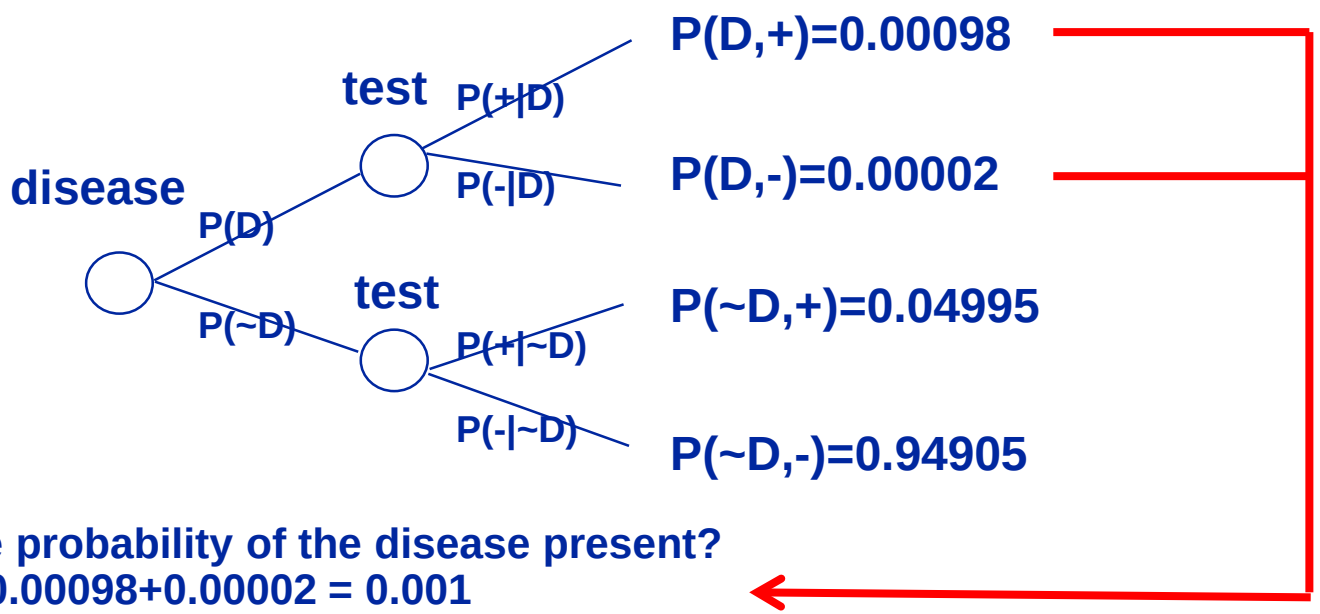
The simplest and quite natural graphical representation of a joint probability distribution over discrete variables



P(disease present ∧ test positive) = P(D ∩ +) = 0.00098
P(disease present ∧ test negative) = P(D ∩ +) = 0.00002
P(disease absent ∧ test positive) = P(~D ∩ +) = 0.04995
P(disease absent ∧ test negative) = P(~D ∩ +) = 0.94905

Computation in probability trees

We can calculate any marginal or conditional probability distribution from the joint probability distribution encoded in the tree.

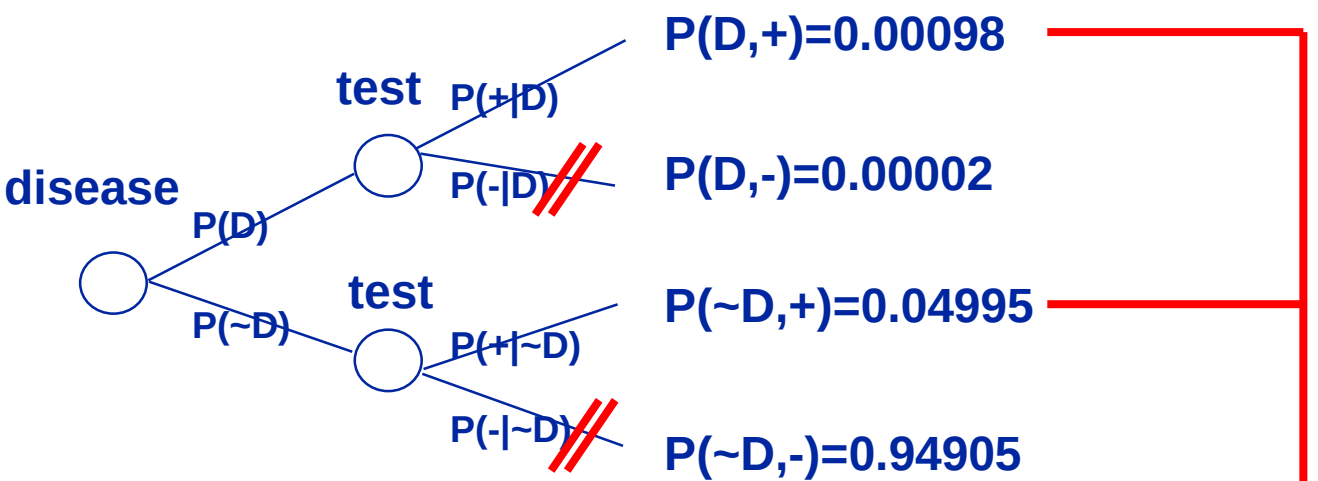


What is the probability of the disease present?

P(D) = 0.00098+0.00002 = 0.001

Computation in probability trees

We can calculate any marginal or conditional probability distribution from the joint probability distribution encoded in the tree.

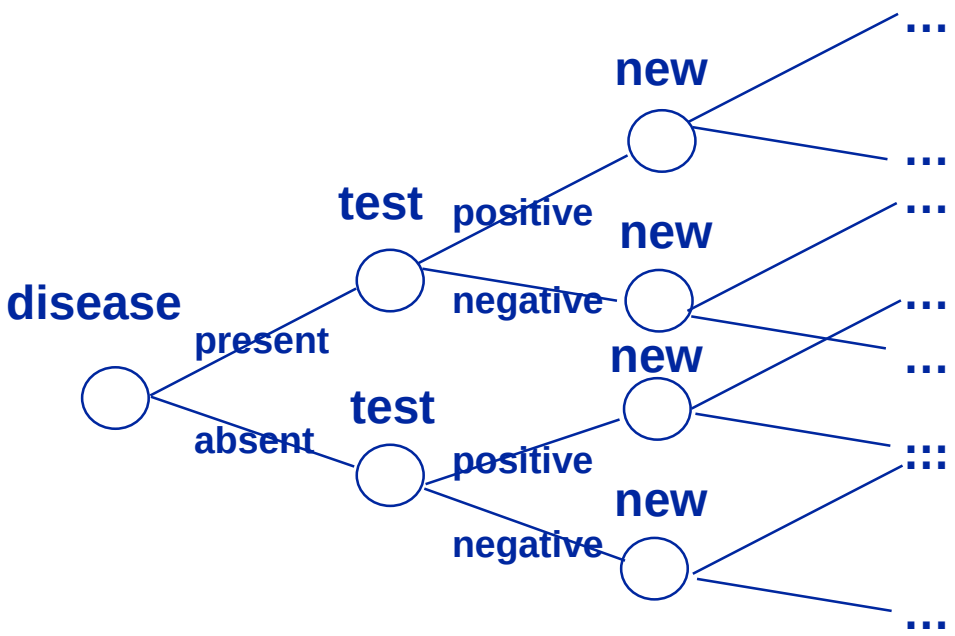


What is the probability of the disease present given a positive test result? Observation of a positive test result makes some of the branches of the tree impossible. What we need to do is just renormalize the remaining, possible (i.e., those that are compatible with the evidence) branches!

$P(D|+) = 0.00098/(0.00098+0.04995) \approx 0.01924$

What is wrong with probability trees?

Trees grow exponentially with the number of variables



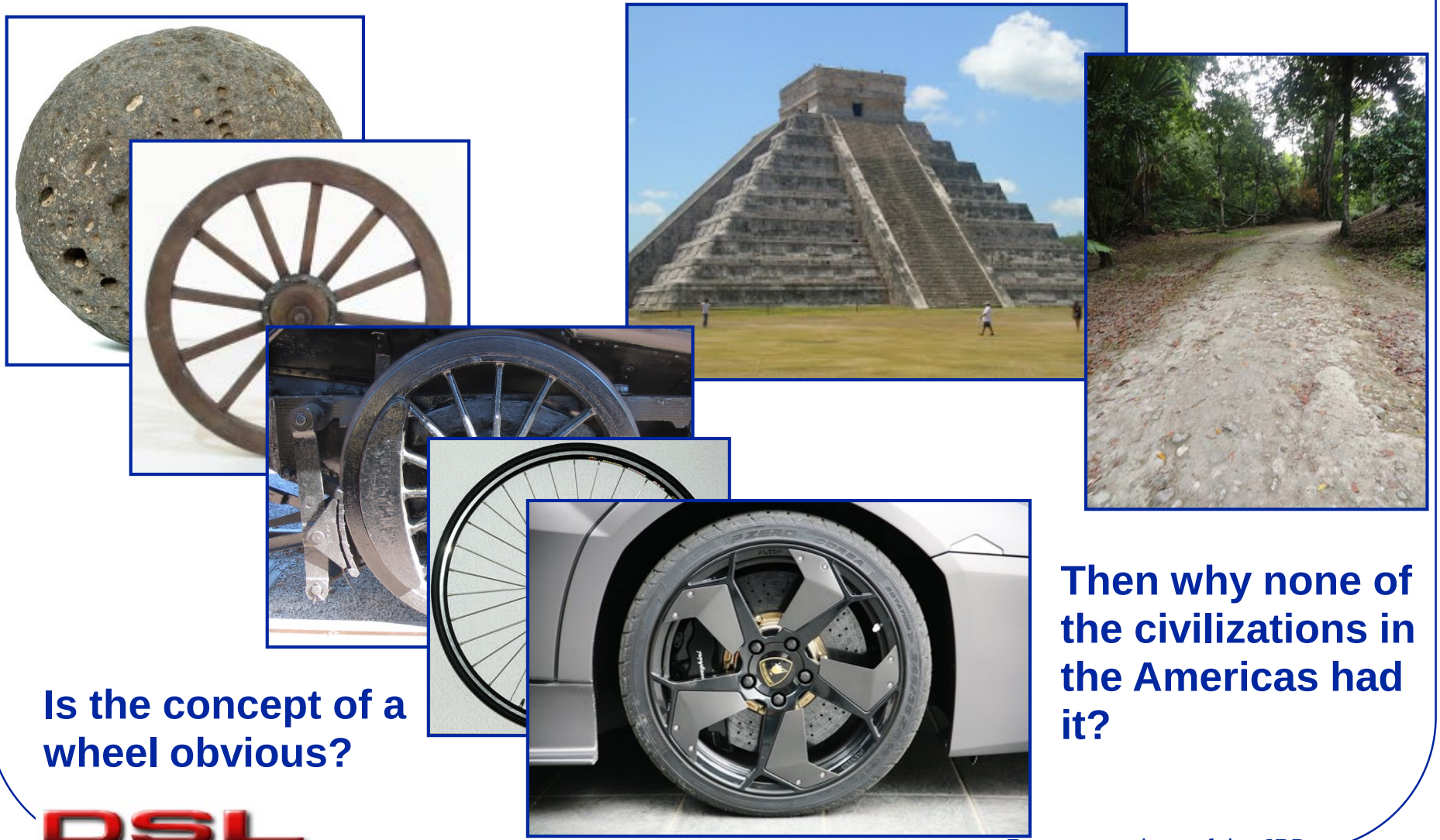
For n binary variables, we will have 2^n branches.
When n=10, the total number of branches is 1,024
When n=11, it is 2,048
…
When n=20, it is 1,048,576 (which is a lot ☺)

Great idea (only 30-40 years old)

Use independences among variables in the joint probability distribution to reduce the number of parameters in its representation!

Due to seminal work on probabilistic independence by A. Philip Dawid and Judea Pearl

All brilliant ideas are obvious (once we have them ☺)



Is the concept of a wheel obvious?

Then why none of the civilizations in the Americas had it?

Factorability of the joint probability distribution

Every joint probability distribution can be factorized, i.e., rewritten as a product of prior and conditional probability distributions of each of the model's variables

$$f(X_1, X_2, ..., X_n) = f(X_1 \mid X_2, X_3, ..., X_n)\, f(X_2 \mid X_3, ..., X_n) \ldots f(X_{n-2} \mid X_{n-1}, X_n)\; f(X_{n-1} \mid X_n)\, f(X_n)$$

e.g., four variables (a, b, c, d), we have:

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

There are n! different directed graphs corresponding to various ways of factorizing a joint probability distribution over n variables.

For n=4, we have 4!=24 different factorizations.

Factorability of the joint probability distribution

- Any factorization can be simplified if we consider independencies among variables.
- Those factorizations that become the simplest are better than others in terms of efficiency of representation.

e.g., suppose we know that $B \perp D|C$, $D \perp A|C$, and $A \perp C$

We can simplify

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

into

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

Bayesian networks

- This underlies the very idea of Bayesian networks.
- We draw a directed graph with arc from the conditioning variables to the variables in the factorization.

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

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

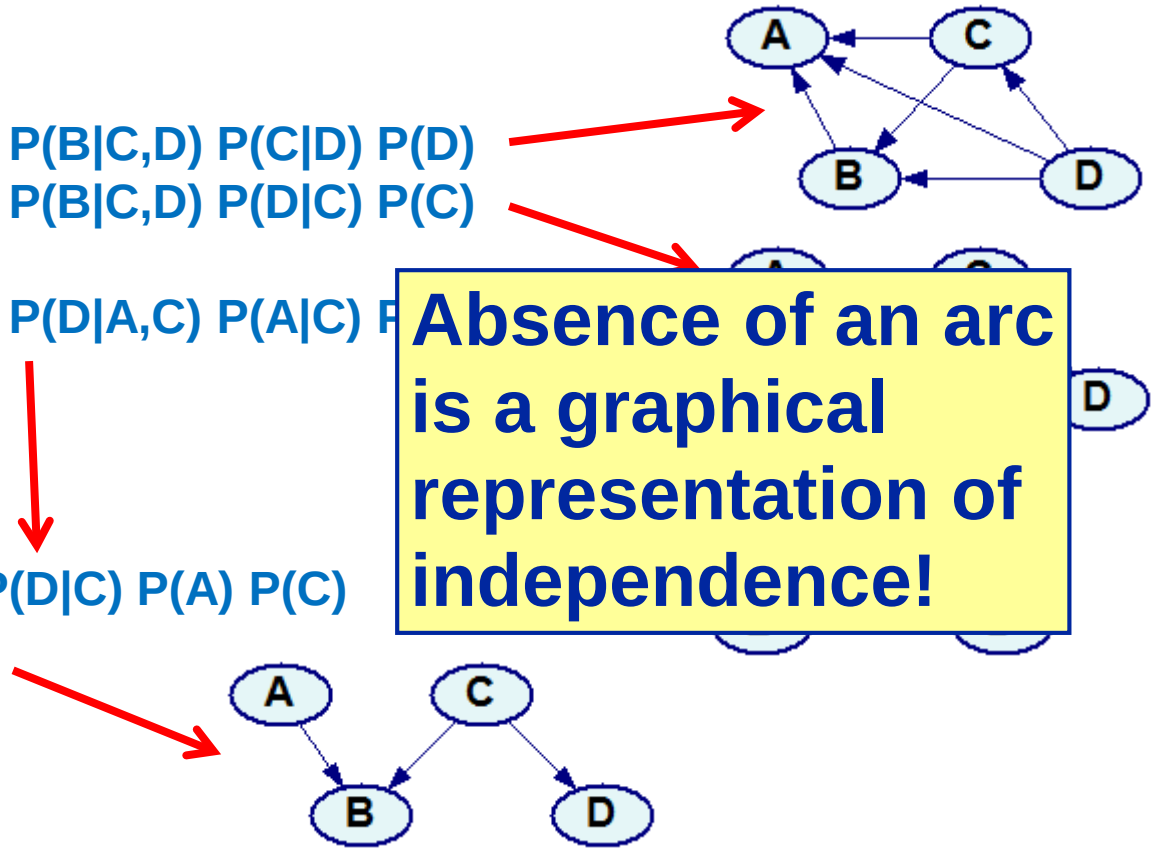
...

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

...

$B \perp D|C,\ D \perp A|C,\ A \perp C$

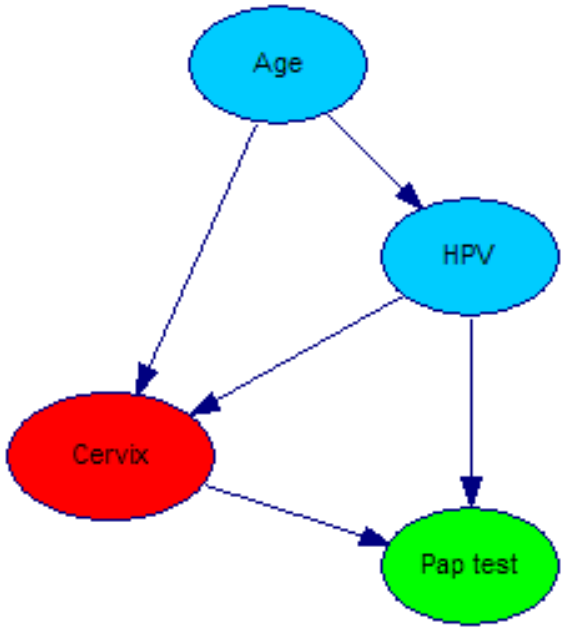
$P(A,B,C,D)=P(B|A,C)\,P(D|C)\,P(A)\,P(C)$



Absence of an arc is a graphical representation of independence!

Bayesian networks

A Bayesian network [Pearl 1988] is an acyclic directed graph consisting of:

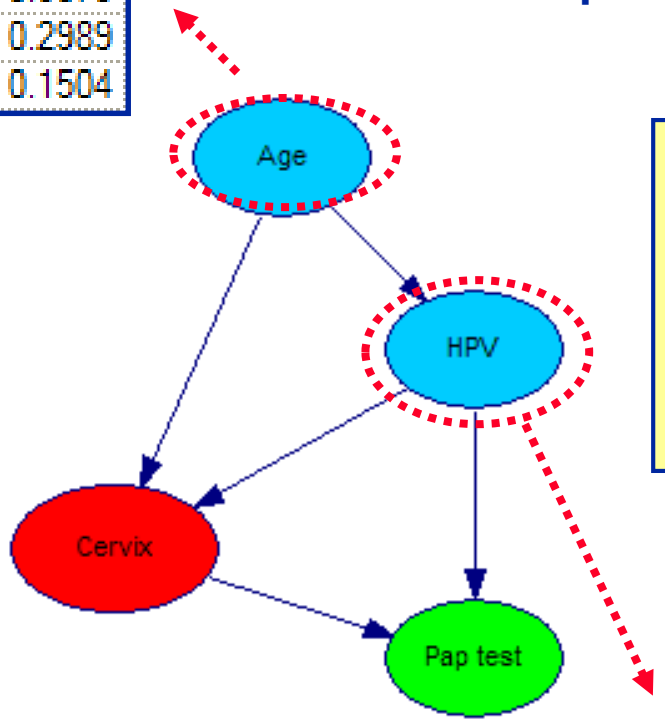


- **The qualitative part, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).**
- **The quantitative part, encoding the joint probability distribution over these variables.**

Bayesian networks: Numerical parameters

a1_below_20	0.0416
a2_20_29	0.2012
a3_29_45	0.3079
a4_45_60	0.2989
a5_60_up	0.1504

Prior probability distribution tables for nodes without predecessors (Age)



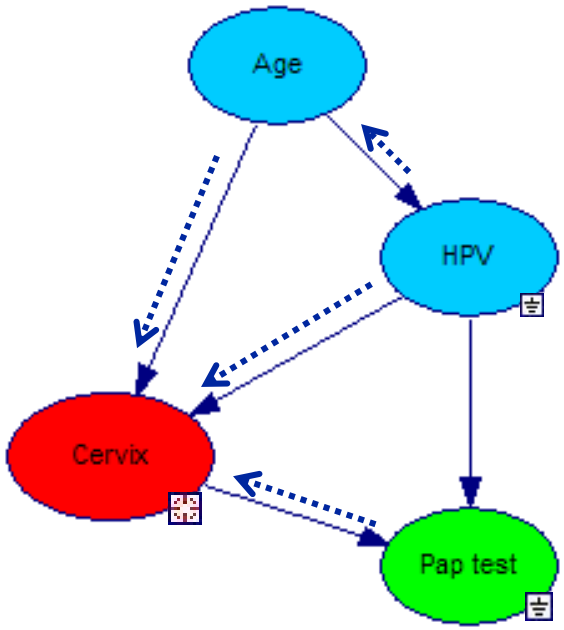
Please note that each absence of an arc (i.e., each independence modeled) is means one less dimension in the corresponding conditional probability table!

Conditional probability distributions tables for nodes with predecessors (HPV, Pap test, Cervix)

Age	a1_below_20	a2_20_29	a3_29_45	a4_45_60	a5_60_up
NA	0.8652	0.8387	0.7904	0.8055	0.8851
Negative	0.069	0.0901	0.1782	0.1765	0.1012
Positive	0.0613	0.0667	0.0282	0.0142	0.0082
Qns	0.0045	0.0045	0.0032	0.0038	0.0055

Reasoning in Bayesian networks

The most important type of reasoning in Bayesian networks is updating the probability of a hypothesis (e.g., a diagnosis) given new evidence (e.g., medical findings, test results).



P(CxCa | HPV=positive, HSIL=yes)

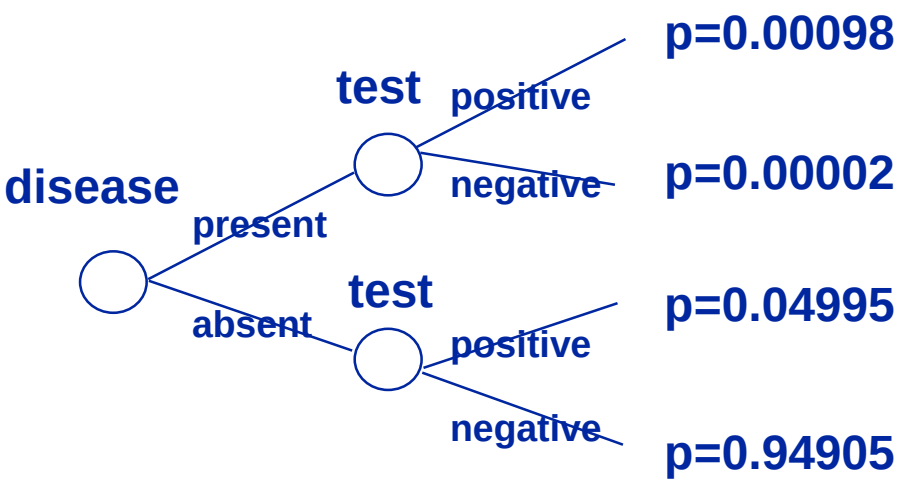
Example:

What is the probability of invasive cervical cancer in a (female) patient with high grade dysplasia with a history of HPV infection?

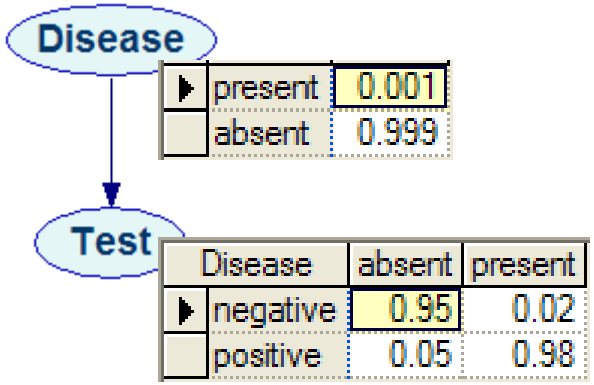
Generally, the more sparse the structure of your network, the fewer parameters, the faster inference in the Bayesian network.

Probability trees and Bayesian networks

probability tree

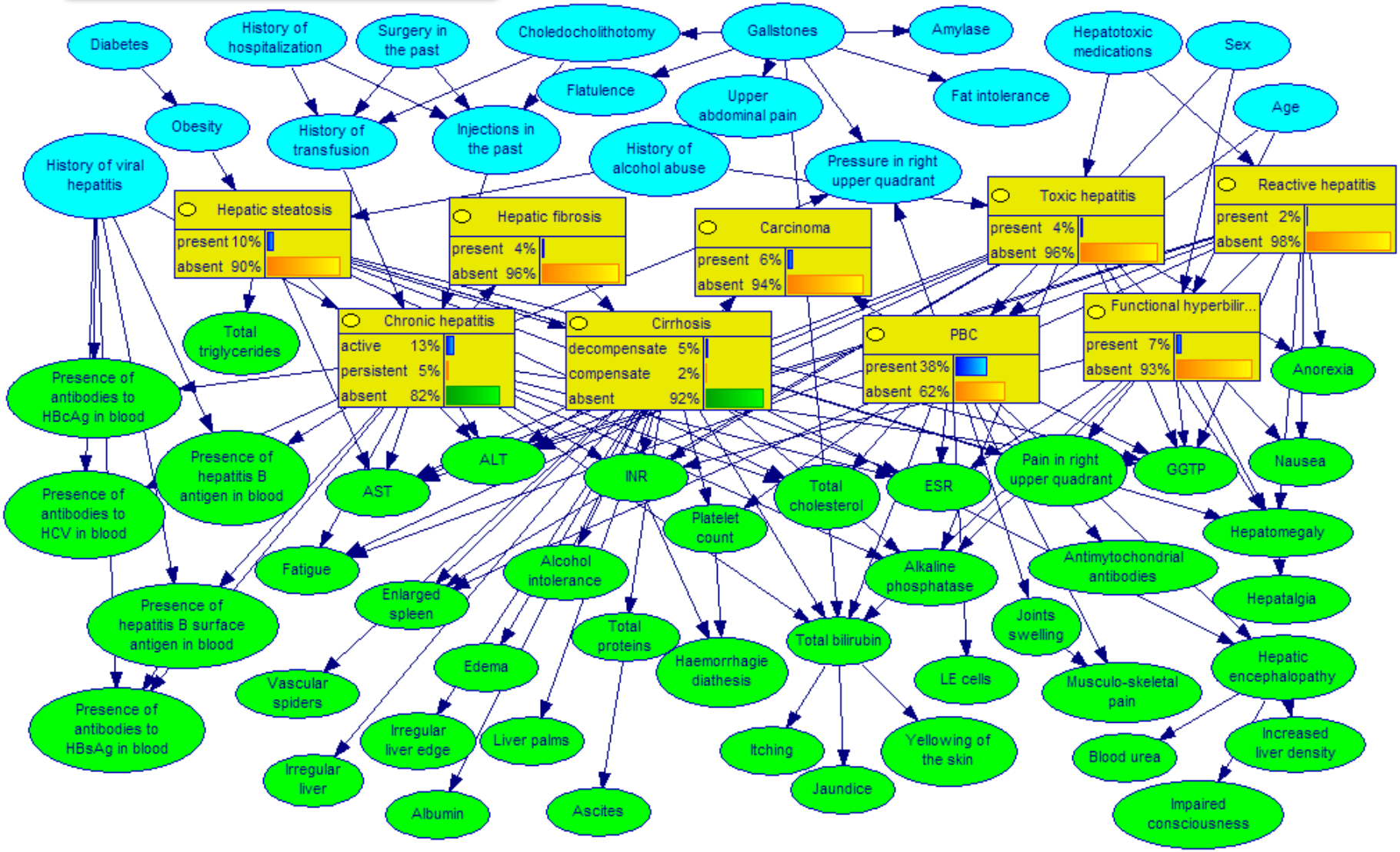


Bayesian network



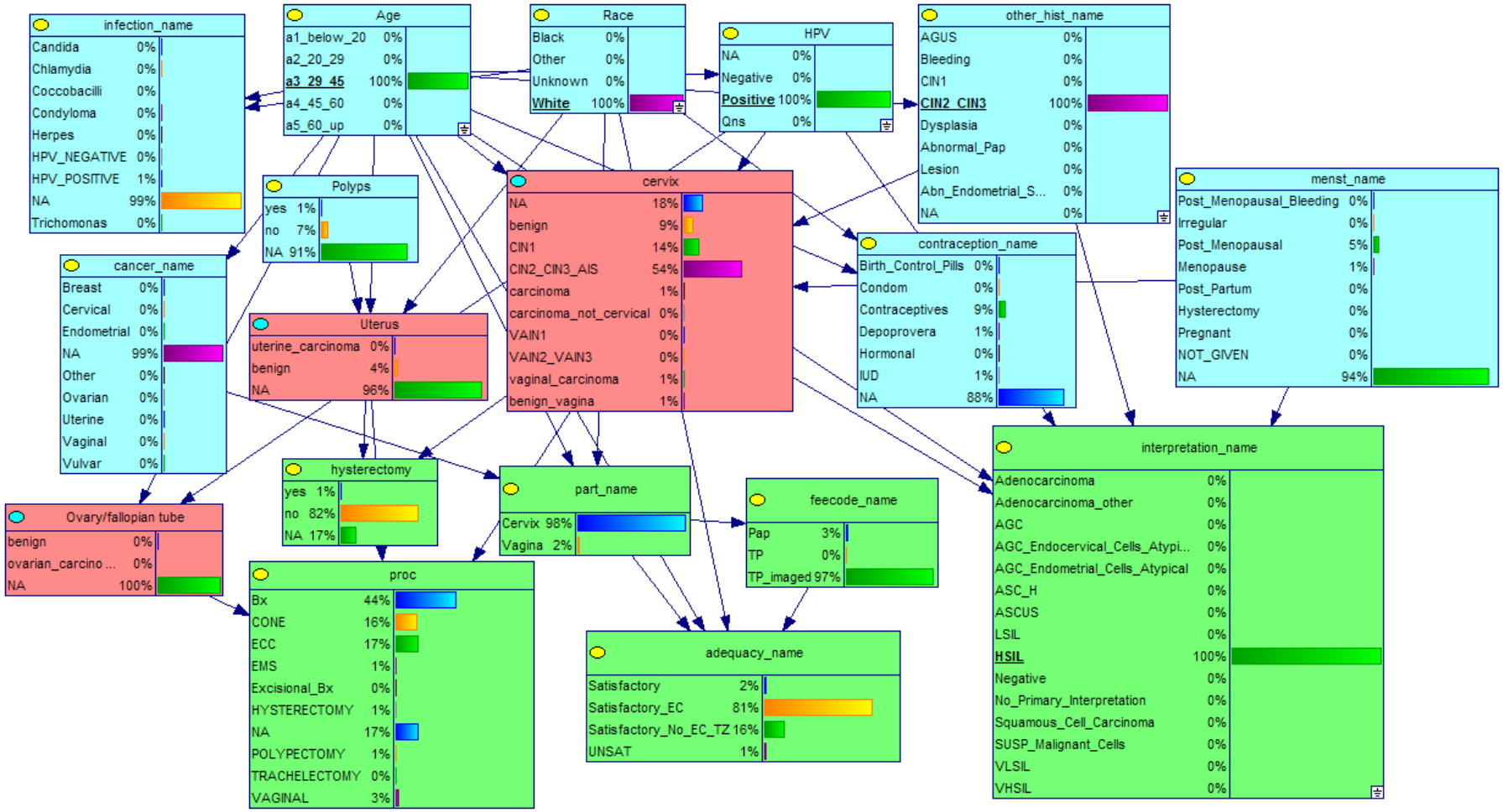
The two representations are equivalent
But, when there are independences in the domain,
Bayesian networks are much, much more efficient!

HEPAR II Model



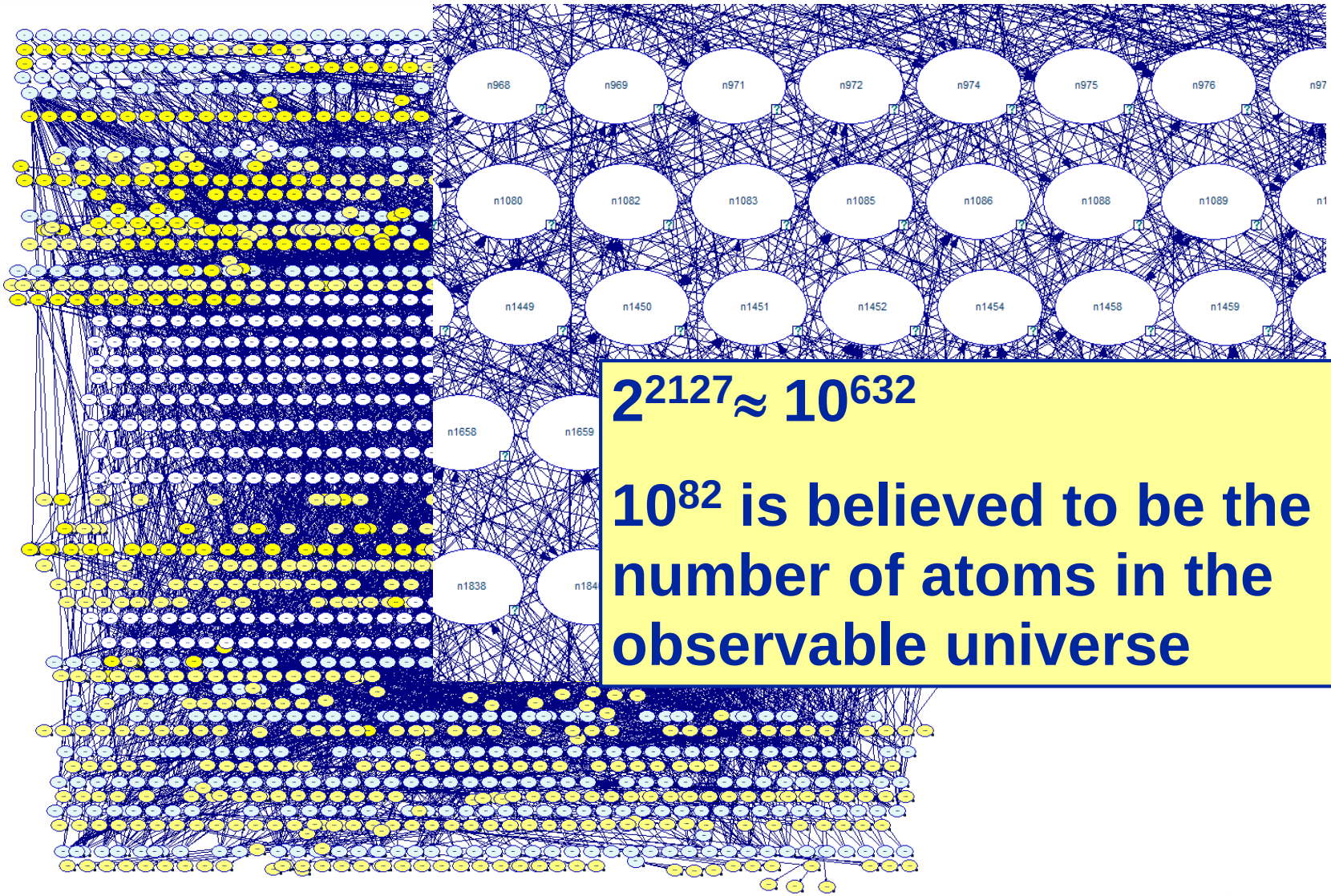
70 variables; 2,139 numerical parameters (instead of over $2^{70} \approx 10^{21}$!)

Pittsburgh Cervical Cancer Screening Model



[Oniśko et al.] 18 variables; 295,163 numerical parameters (instead of over 10^{13}!)

Diagnosis of Diesel locomotives



$2^{2127} \approx 10^{632}$

10^{82} is believed to be the number of atoms in the observable universe

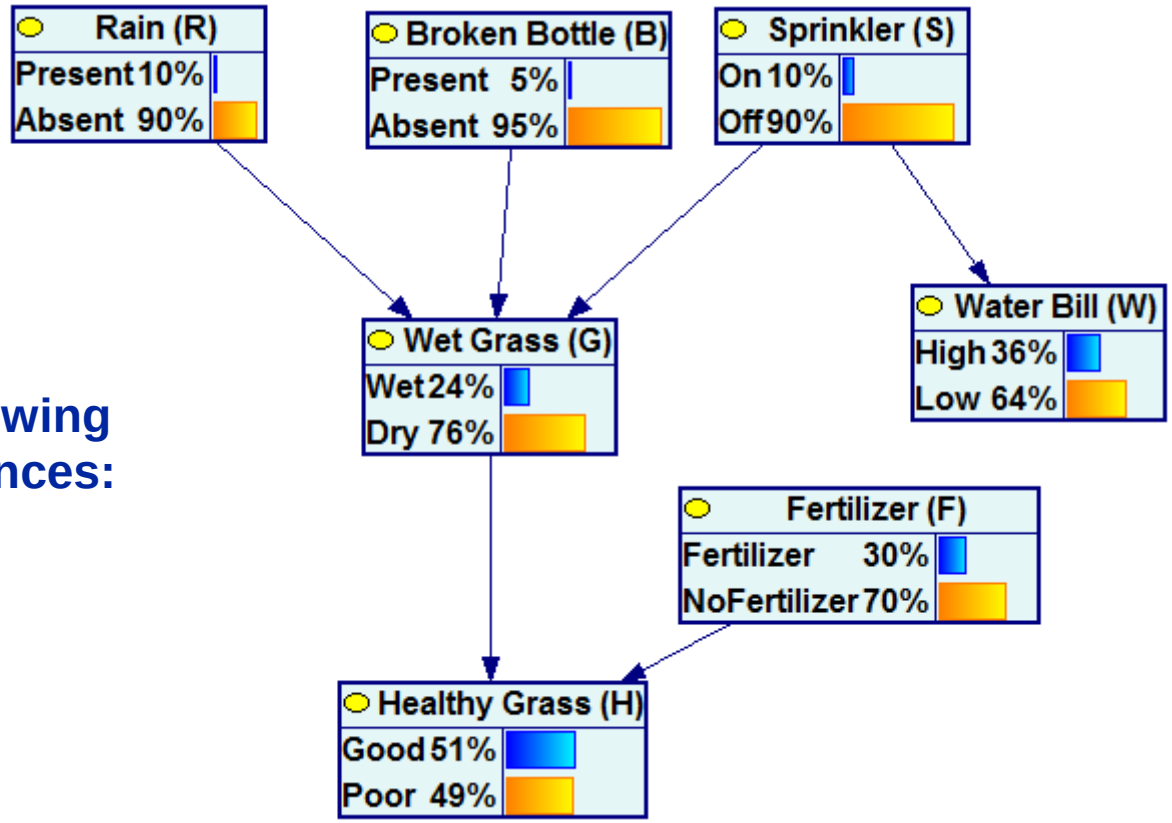
[Przytula et al.] 2,127 variables; 12,351 numerical parameters (instead of 2^{2127}!)

Independences: Markov condition

- Allows to read back dependences and independences from the graph.
- Informally speaking, it is an assumption that ties directed probabilistic graphs with probability, specifying how a directed graphs represents independence.
- A node is independent of its non-descendants given its predecessors.

Markov condition: Example

$$P(H,G,W,R,B,S,F)=P(H|G,F)\,P(G|R,B,S)\,P(W|S)\,P(R)\,P(B)\,P(S)\,P(F)$$



This graph implies the following (conditional) independences:

$R \perp B$, $R \perp S$, $B \perp S$, $R \perp F$, $B \perp F$, $S \perp F$

$R \perp W$, $B \perp W$, $W \perp F$, $G \perp F$

$R \perp H|G$, $B \perp H|G$, $S \perp H|G$, $W \perp H|G$

$W \perp *|S$

$R \perp W|G,S$, $B \perp W|G,S$

Equation-based systems and graphical models

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

← Core equations

cload = 15

tload = 6

nstud = 22102

nfac = 3006

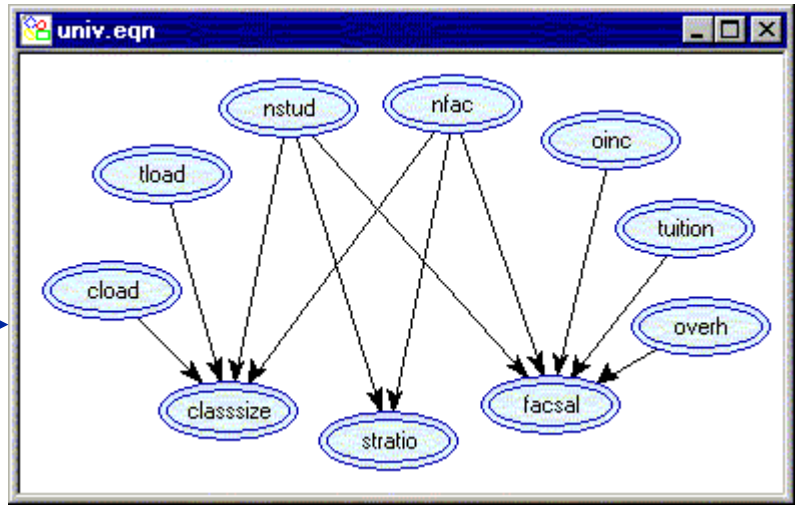
oinc = 30000000

tuition = 12000

overh = 0.48

← Equations for exogenous variables

Together they determine the structure of the model →



Equation-based systems: Reversibility of causal ordering

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

cload = 15

tload = 6

nstud = 22102

~~*nfac* = 3006~~

oinc = 30000000

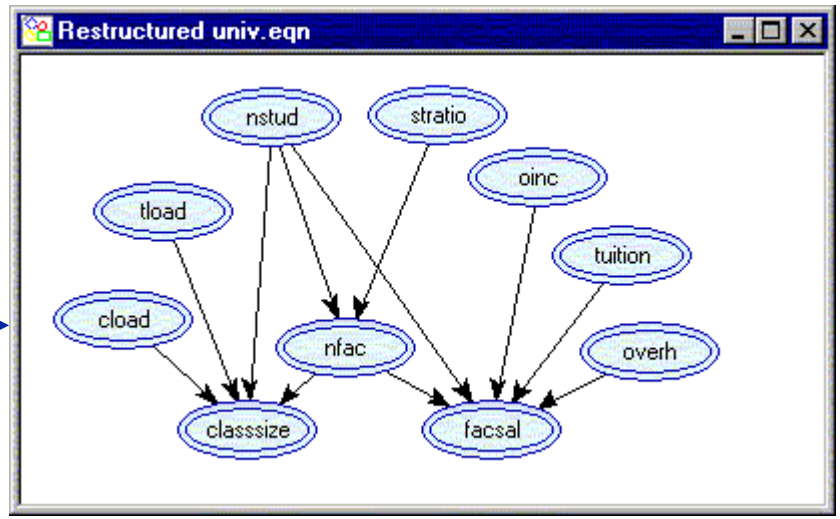
tuition = 12000

overh = 0.48

Setting *stratio* to be exogenous at the expense of *nfac*

stratio = 10

The new model structure



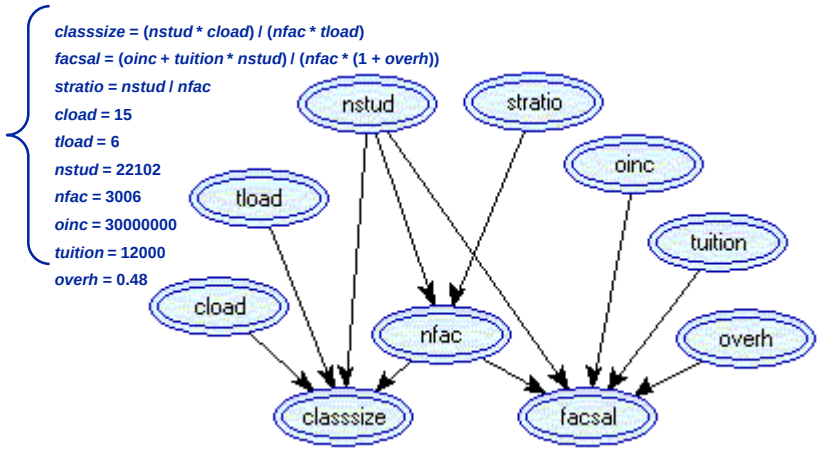
Explication of the asymmetries due to Herb Simon (early 1950s)

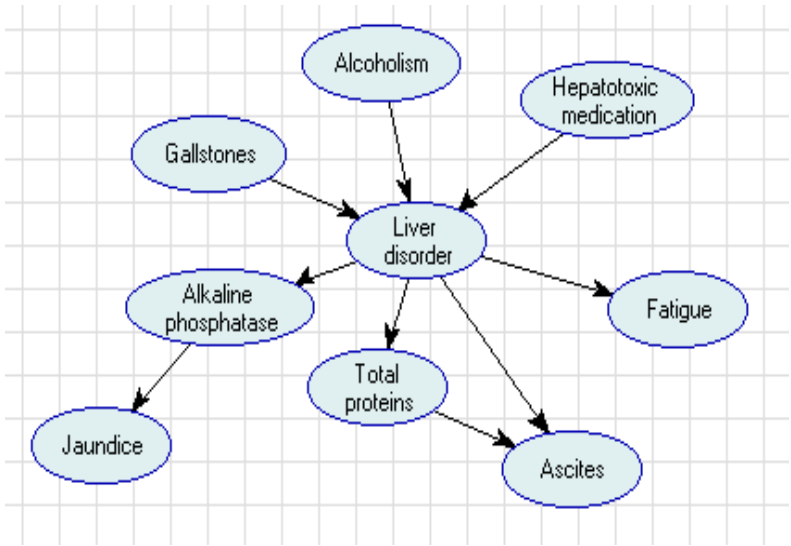
Advantages of directed graphs

- May be built to reflect the causal structure of a model (helps with obtaining insight into the problem)
- Can accommodate representation of uncertainty
- Can be reconfigured as needed
- Have sound theoretical foundations: We are dealing here with probability theory and decision theory
- We can talk (almost) the same language with statisticians, philosophers, and scientists

Family of directed graphs (a bigger picture)

(a.k.a. “influence nets,” “causal diagrams,” etc.)





Both, systems of equations and joint probability distributions can be pictured by acyclic directed graphs.