Session 3:
Building Bayesian Networks

Marek J. Drużdżel

Wydział Informatyki

Politechnika Białostocka

m.druzdzel@pb.edu.pl

http://aragorn.wi.pb.bialystok.pl/~druzdzel/

Session overview

- Elicitation of structure
- Elicitation of probabilities
- Canonical (a.k.a. ICI, independence of causal influences) models
- Do parameters matter?
- Does structure matter?
- Other relevant issues
  - Sensitivity analysis
  - Strength of influence
  - Value of information
  - Clarity test

What I want you to know after this session?

- Know how to start and how to proceed with building Bayesian networks
- Be somewhat familiar with the idea of obtaining subjective probabilities from experts
- Know Noisy-OR gates (and know of existence of other canonical models)
- Do not worry too much about precision of numerical probabilities, use sensitivity analysis
- Know how to simplify the structure
- Know sensitivity analysis, strength of influences, value of information, and clarity test

Elicitation of structure

- Create a network node for every variable in your problem
- Connect nodes that “directly impact each other” by means of direct arcs
- What does it mean for two variables to directly impact each other?
- The game is a correct factorization of the probability distribution
- How do you go from direct impact to correct factorization?

Bayesian network models are snapshots of the World

- An important thing to realize is that they represent snapshots of the world, a static situation, and do not model dynamic systems with feedback loops, etc.
- Is this a limitation? Not really ...

Subjectivity and usefulness of models

- **Another important point is that there are multiple representations of the same problem possible.**
- **Models are subjective!**
- ***“All models are wrong, but some are useful”* – Statistician George E P Box, in "Science and statistics", *Journal of the American Statistical Association,* 71:791-799, 1976**

Causality

- **One way of thinking about direct impact is through causality**
- **Following causal structure typically guarantees us that the resulting factorization will be correct**
- **Why is that ☺?**

Elicitation of probabilities

Elicitation of probabilities

Three fundamental methods:

- **Ask directly**
- **Symmetric bets**
- **Reference lottery**

Three additional issues:

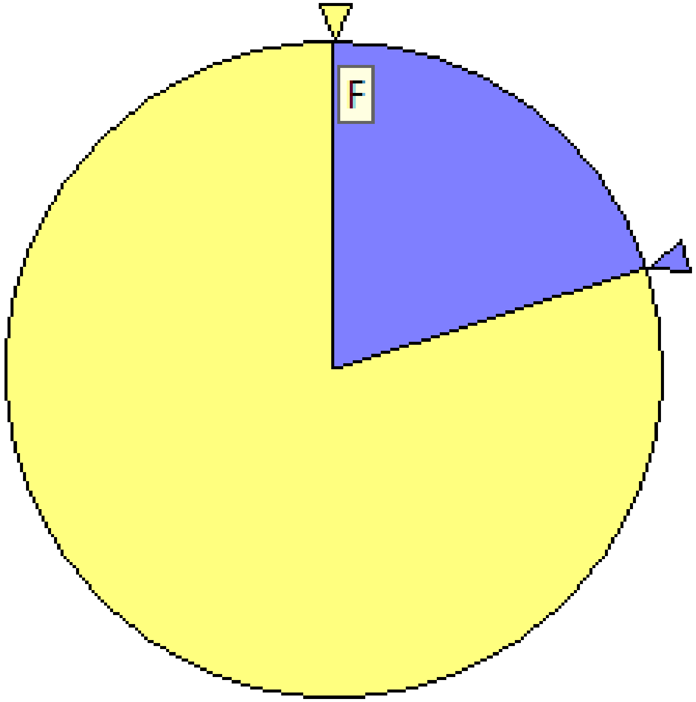
- **Assessing continuous distributions**
- **Discretization of continuous distributions**
- **Decomposition**

Elicitation of probabilities: Direct assessment

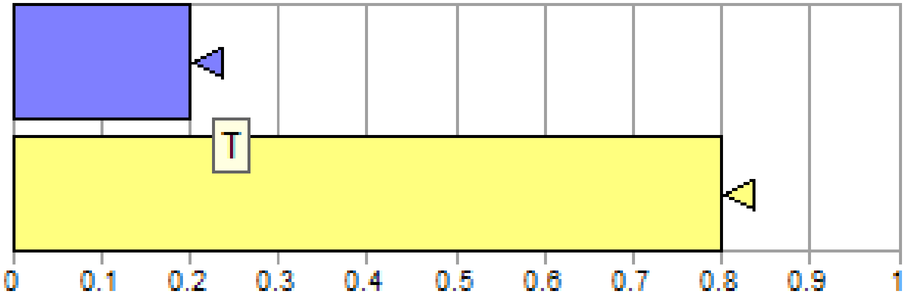
Example:

"What is your belief regarding the probability that event A will occur?"

Graphical aids that make it indirect: Probability wheel



Bar chart



Elicitation of probabilities: Symmetric bets

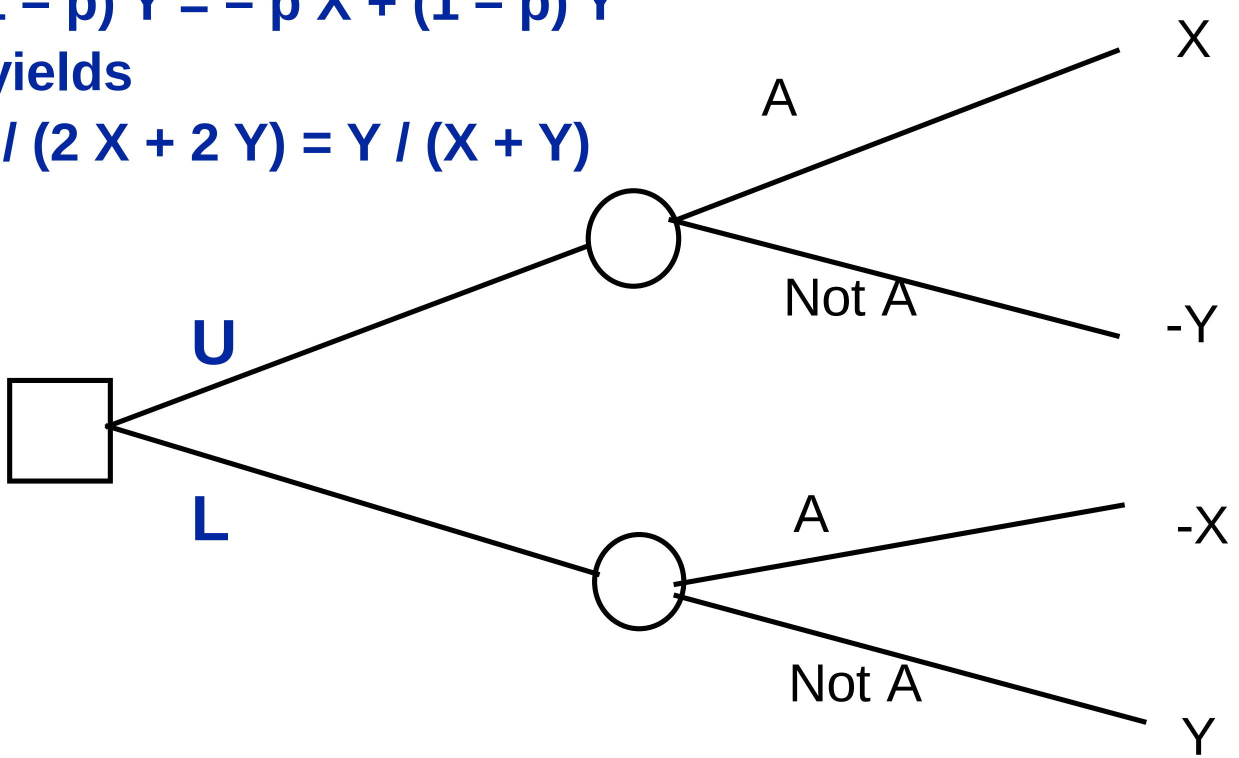
Offer choice between two lotteries, adjust values until the expert is indifferent between the two lotteries.

Then we have:

$p X - (1 - p) Y = - p X + (1 - p) Y$

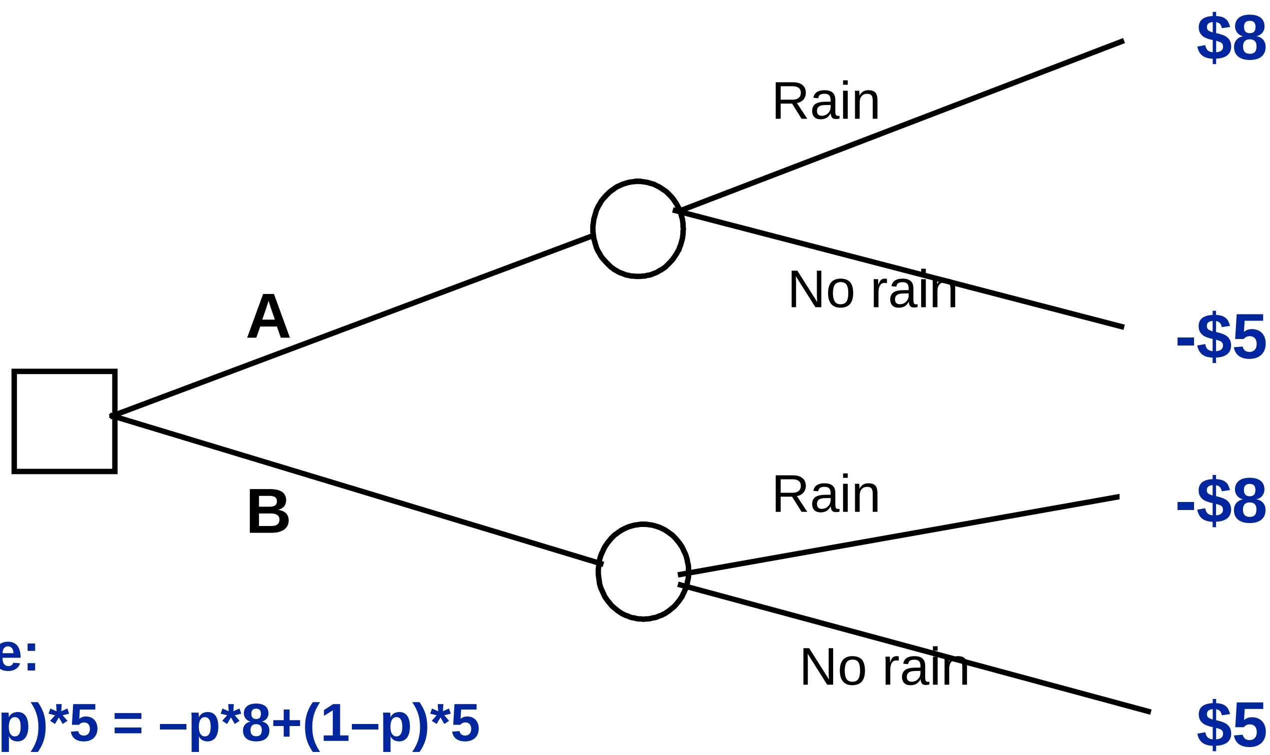
which yields

$p = 2 Y / (2 X + 2 Y) = Y / (X + Y)$



Elicitation of probabilities: Symmetric bets

What is the probability that it will rain tomorrow (in downtown Pittsburgh)?



We have:

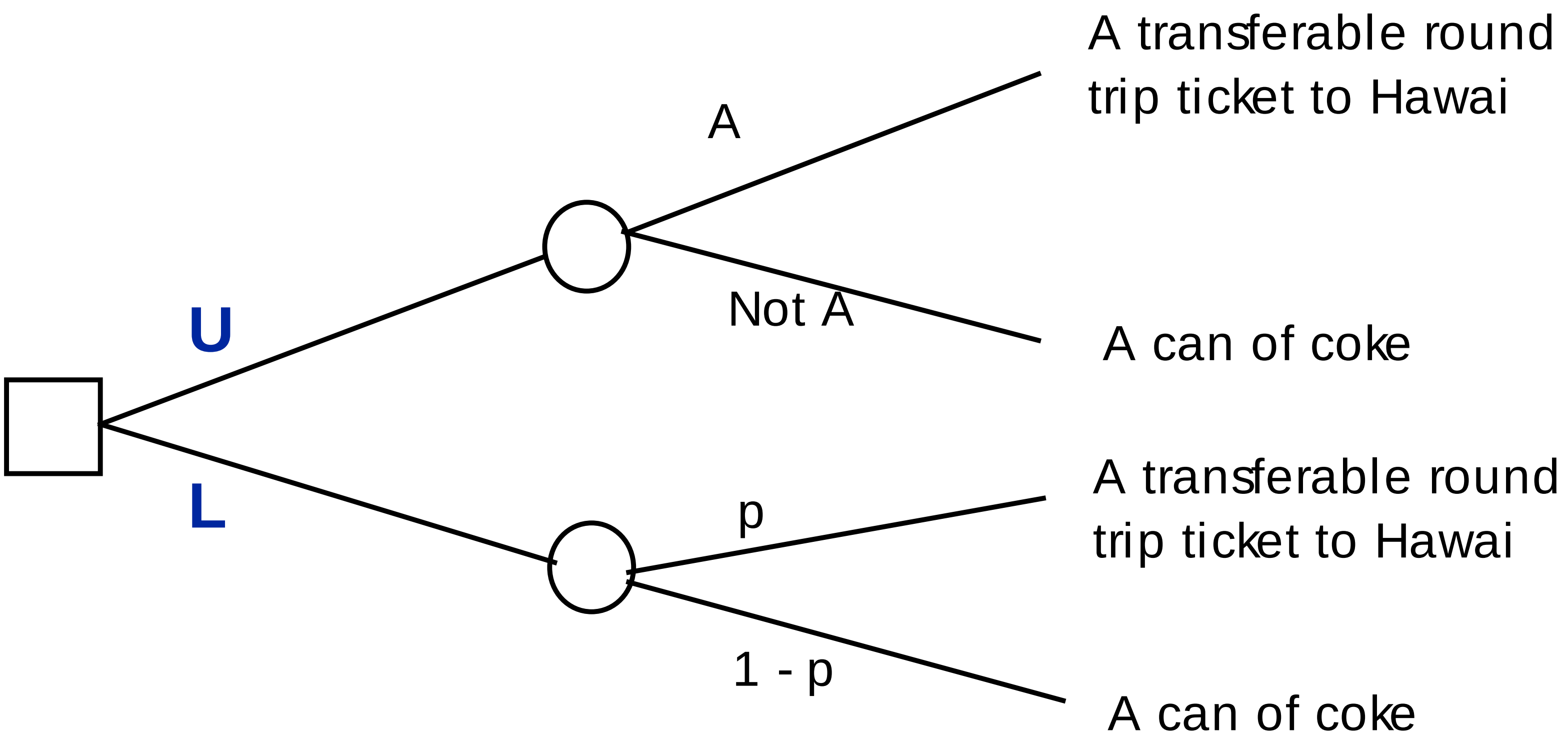
p*8–(1–p)*5 = –p*8+(1–p)*5

which yields

p=2*5/(2*8+2*5)=5/(8+5)=5/13=0.38

Expert choice: indifferent

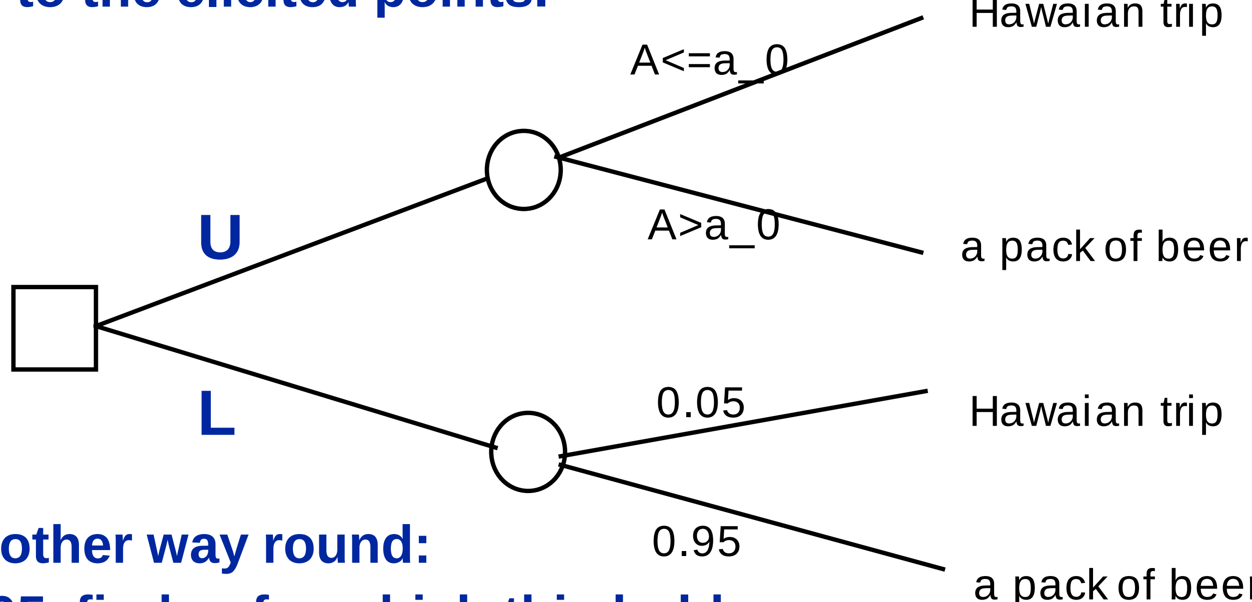
Elicitation of probabilities: Reference lottery



Use a tool like probability wheel (to hide the numbers).

Elicitation of probabilities: Continuous distributions

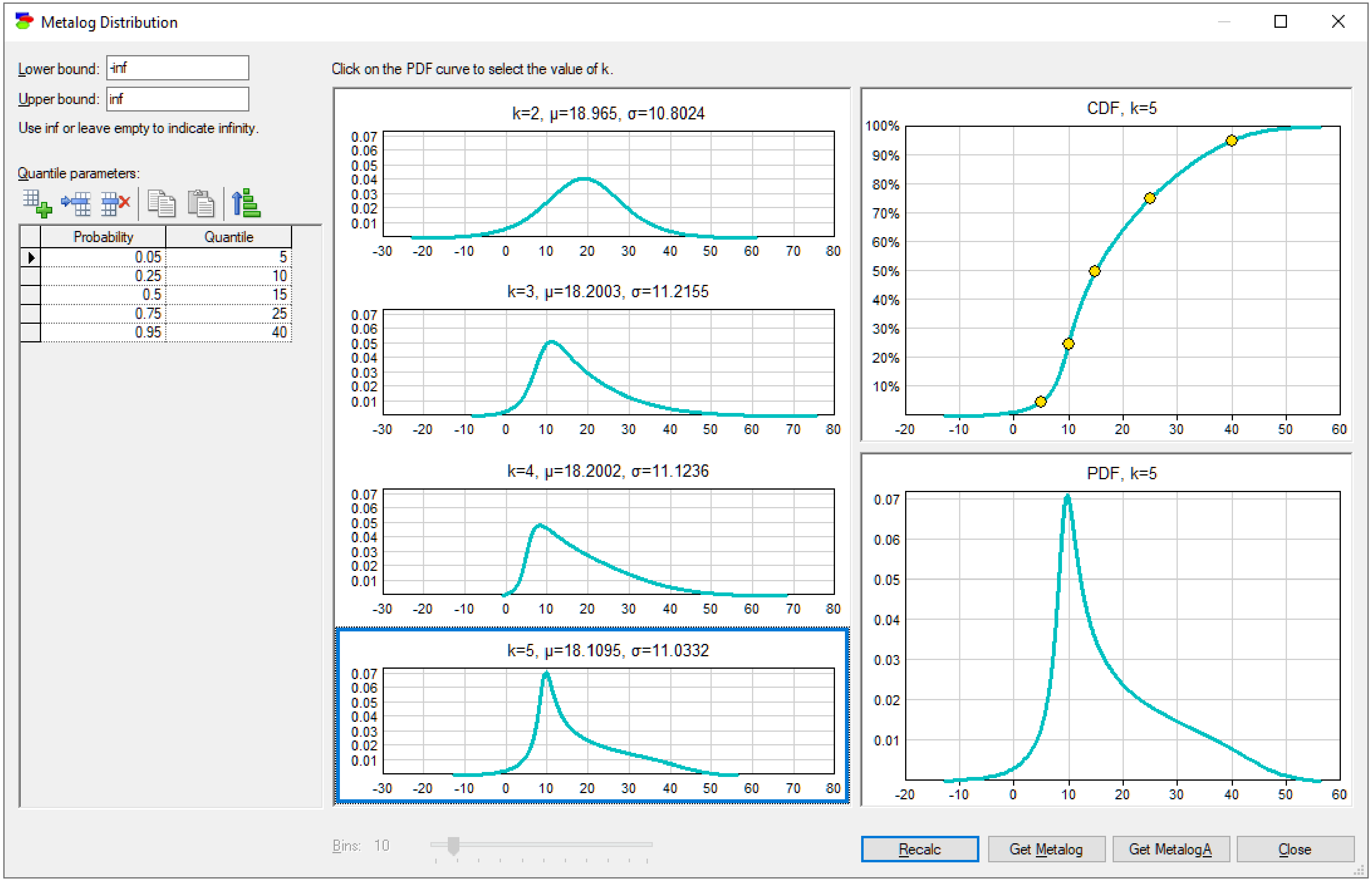
- Use methods for elicitation of discrete probabilities but conduct a series of elicitations.
- Reduces each step of elicitation to $P(A \leq a_0)$, where a_0 varies.
- Fit the CDF curve to the elicited points.



Possible to do the other way round:

- Given $P(A \leq a_0)=0.05$, find a_0 for which this holds.
- Manipulate a_0 until the expert is indifferent between the two options.
- Use the following fractiles: 0.05, 0.95, 0.25, 0.75. 0.5.

Elicitation of probabilities: Metalog Distribution



Elicitation of probabilities: Discretization of continuous distributions

Two methods of discretizing continuous distributions:

(1) Extended Pearson and Tuckey:

3 point approximation: 0.05, 0.5, 0.95

Assign them p=0.185, 0.63, 0.185

(2) Bracket medians:

Split the range into intervals, assess the value that corresponds to probability that is median of each interval. Usually borders of intervals are 0.0, 0.2, 0.4, 0.6, 0.8, 1.0.

Elicitation of probabilities: Decomposition

Breaking the assessment into manageable chunks.

The goal is to make the assessment easier (and more reliable!).

Sometimes it is easier to introduce another variable.

For example, instead of assessing P(quadriplegic), i.e., probability that the decision maker becomes quadriplegic, we assess P(quadriplegic|*) P(*), where * are various ways of becoming quadriplegic, e.g., a car accident.

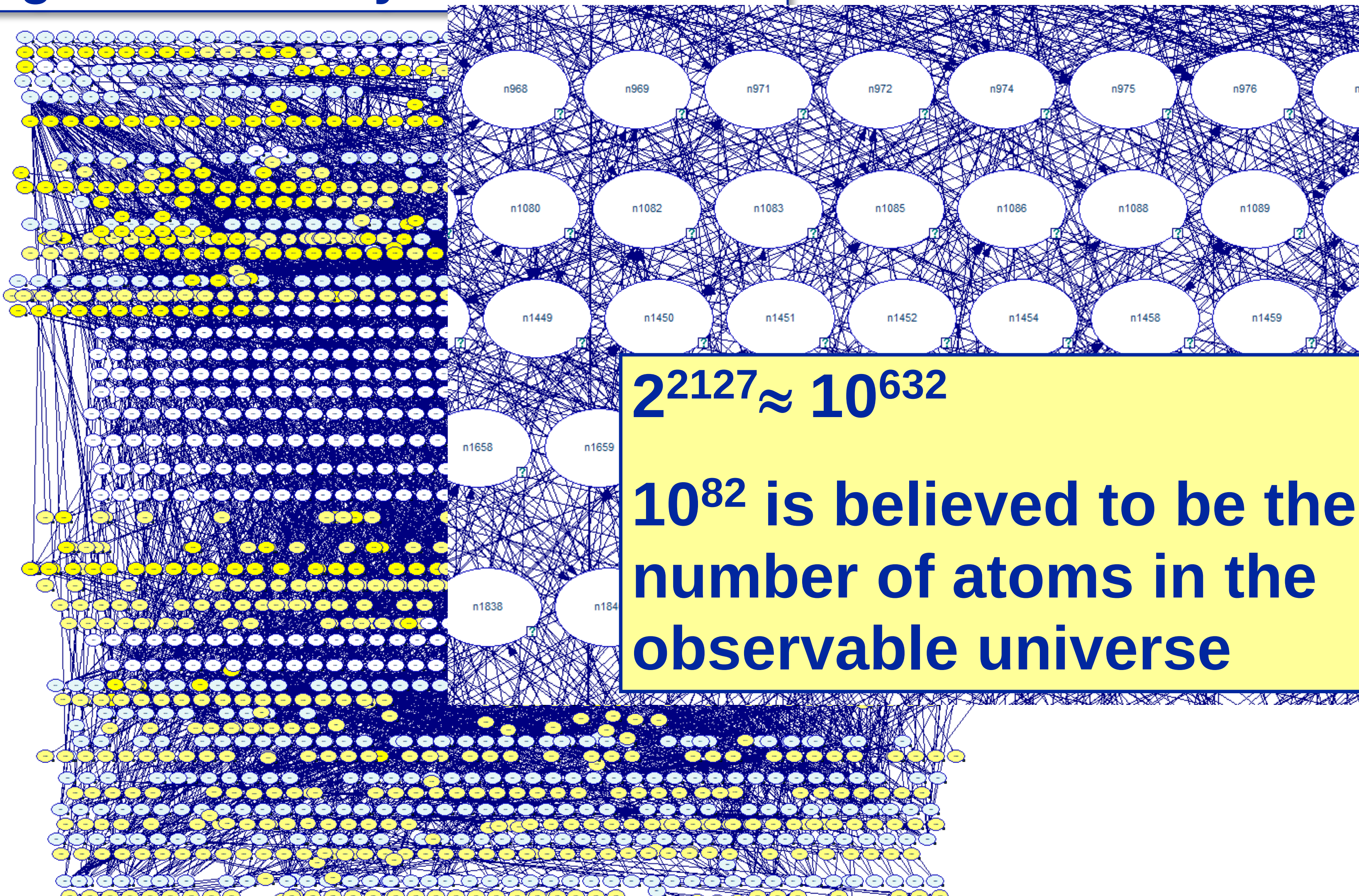
(1) Think how the event in question is related to other events (e.g., P(stock price up | market up))

(2) Think what kinds of alternative uncertain events could eventually lead to the event in question

(3) Think through all of different events that must happen before the event in question occurs.

Canonical models (Noisy OR/MAX)

Practical BN models can be very large and densely connected



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

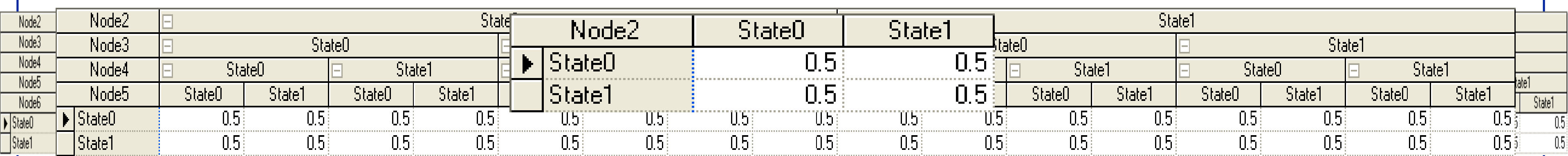
10^{82} is believed to be the number of atoms in the observable universe

[Przytula et al.] 2,127 variables, 3,595 arcs, 2,261,001 independences, 12,351 numerical parameters (instead of $2^{2,127} \approx 10^{632}$!)

Fundamental problem: (too) many parameters

- Size of conditional probability tables (CPTs) grows exponentially in the number of parents
- This can become quickly unmanageable

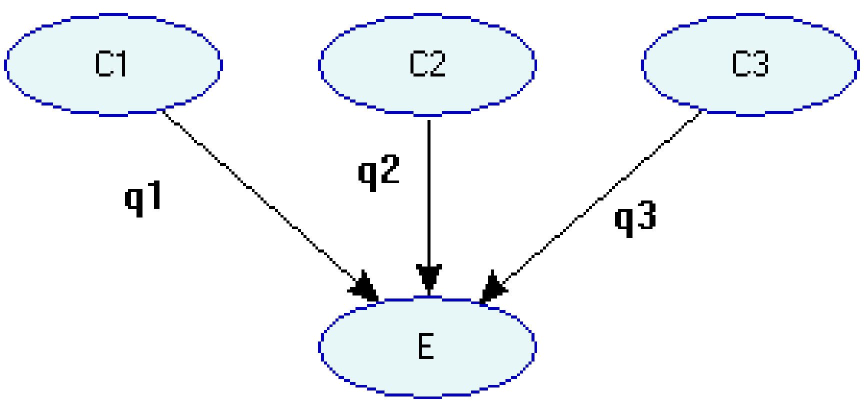
4 parents



- Not uncommon to see 10-15 parents (would need between 1,024 and 32,768 parameters).
- A lot of work for experts or a lot of data needed.

Solution: Canonical gates

- Various solutions were proposed, but one of them seems to be most popular and useful: Noisy-OR
- We assume that all nodes are binary {present, absent}
- We specify the interaction between the parents and the child by means of one numerical parameter q_i per parent



Solution: Canonical gates

Conditions that have to be fulfileld in practice for Noisy-OR to be applicable:

- **There should be a causal mechanism for each parent such that the parent is able to impact the child variable in the absence of the other parents.**
- **The causal mechanisms through which each parent influences the child should be independent?**
- **If there are other, unmodeled causes, they should be independent of the modeled causes.**

Noisy-OR: The meaning of q_i?

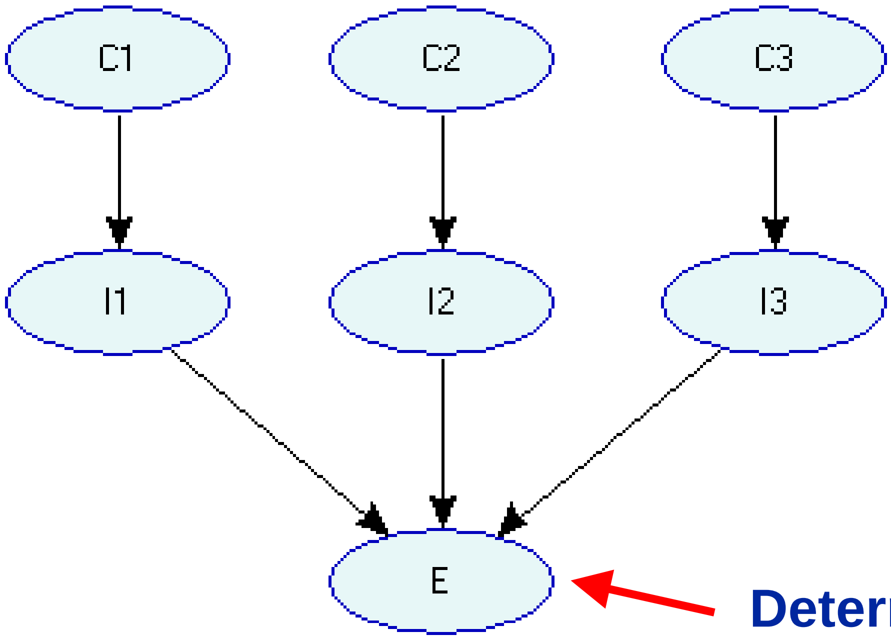
q_i is the probability that E=*present* given C_i=*present* and all other parents $C_{j \neq i}$=*absent*

q_i=P(E=*present* | C_1=*absent*, …, C_i=*present*, …, C_n=*absent*)

Why is it called Noisy-OR?

If all parameters q_i=1, noisy-OR becomes logical OR

Here is an alternative representation of Noisy-OR



Node0	present	absent
present	q	0
absent	1-q	1

Deterministic OR

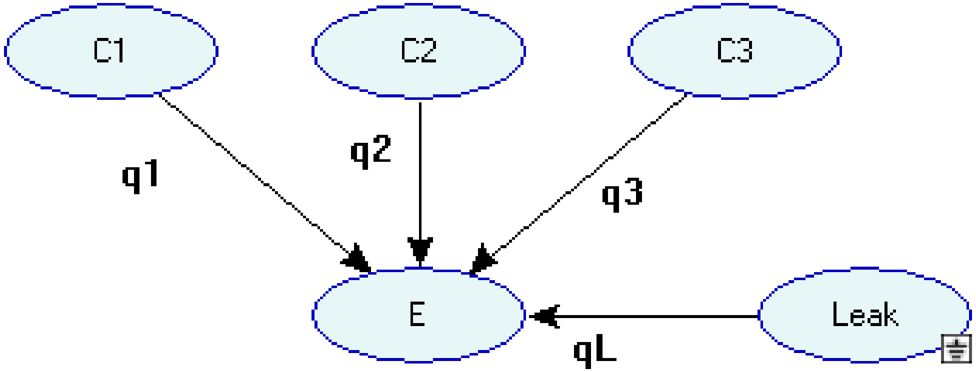
Noisy-OR vs. CPT

Noisy-OR always defines a unique CPT (i.e., you can always calculate the CPT that is defined by a noisy-OR gate)

$$P(E = absent \mid C1, \ldots, Cn) = \prod_{C_i = present} (1 - q_i)$$

Leaky Noisy-OR

- Noisy-OR assumes that the effect will be absent with probability 1 if all the causes are absent. This is not very realistic
- Leak is a special dummy node, that represents the influence of all unmodeled causes on the effect node
- Leak is always present



Leaky Noisy-OR: Parameters

- Leaky Noisy-OR is an extension of the Noisy-OR
- Two parameterizations of leaky Noisy-OR: due to Henrion and Diez (*compound* and *net* parameters)
- They are mathematically equivalent, however they imply different questions in knowledge elicitation

Leaky Noisy-OR: Diez

Leak probability q_L:

$$q_L = P(E = present \mid C1 = absent, .., CN = absent)$$

Link probability q_i:

$$q_i = P(E = present \mid C1 = absent, ..., Ci = present,$$
$$CN = absent, L = absent)$$

How to calculate the CPT:

$$P(E = absent \mid C1, ..., Cn) = (1 - q_L) \prod_{C_i = present} (1 - q_i)$$

Leaky Noisy-OR: Henrion

- **Leak probability p_L:** **(same as Diez)**

$$p_L = P(E = present \mid C1 = absent, .., CN = absent)$$

- **Link probability p_i:** **(no leak term)**

$$p_i = P(E = present \mid C1 = absent, ..., Ci = present, CN = absent)$$

- **How to calculate CPT:**

$$P(E = absent \mid C1, ..., Cn) = (1 - p_L) \prod_{C_i = present} \frac{1 - p_i}{1 - p_L}$$

Henrion vs. Diez

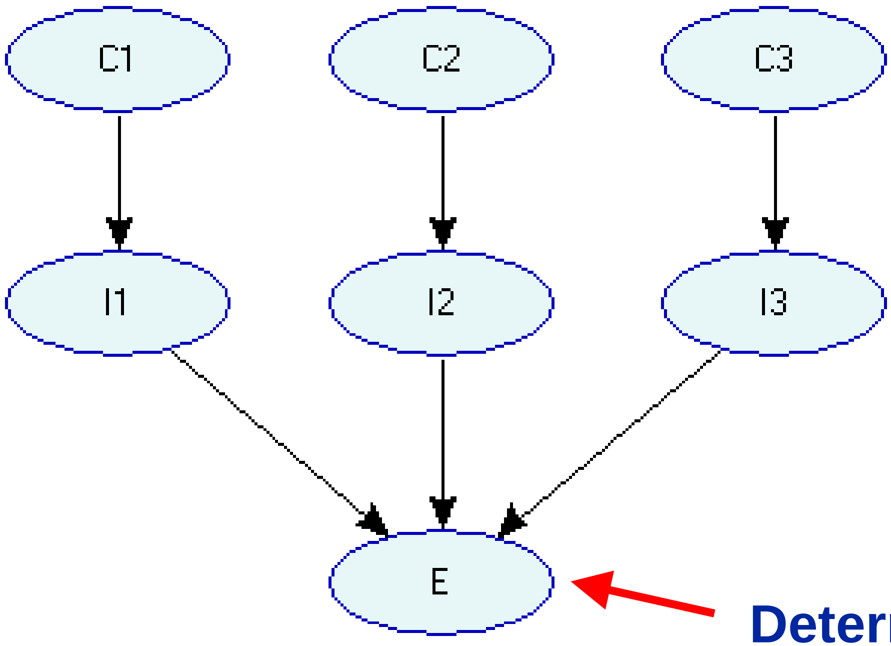
- **They imply different questions to ask of experts:**
- **Henrion:**

  "What is the probability that E is present given that C_i is present and all other modeled causes are absent?"
- **Diez:**

  "What is the probability that E is present given that C_i is present and all other modeled and unmodeled causes are absent?"

Noisy-MAX

Noisy-MAX is a version of Noisy-OR for multi-valued nodes.

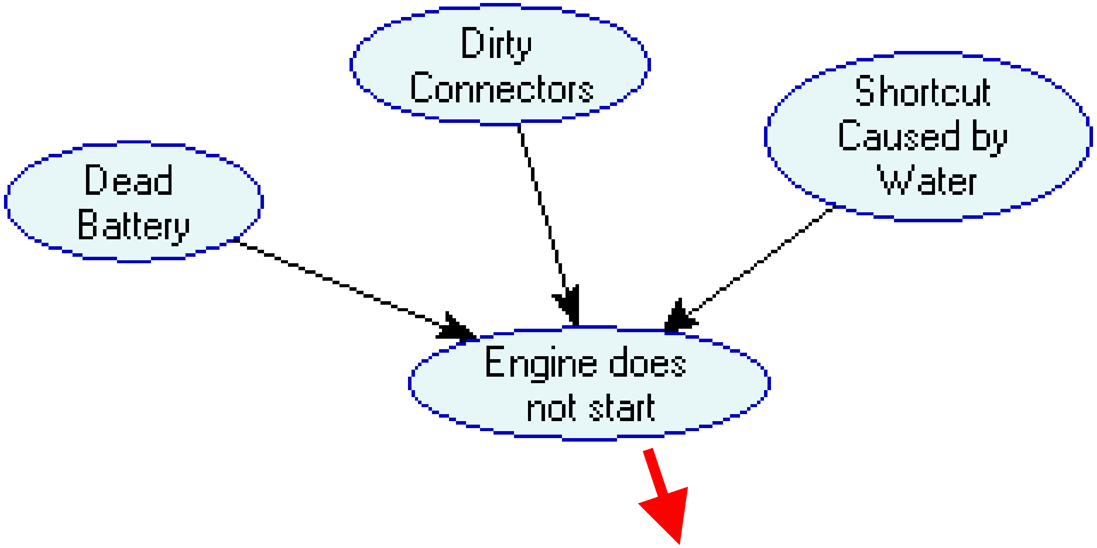


Node2	high	med	low
high	0.7	0.5	0
medium	0.2	0.3	0
low	0.1	0.2	1

Deterministic MAX

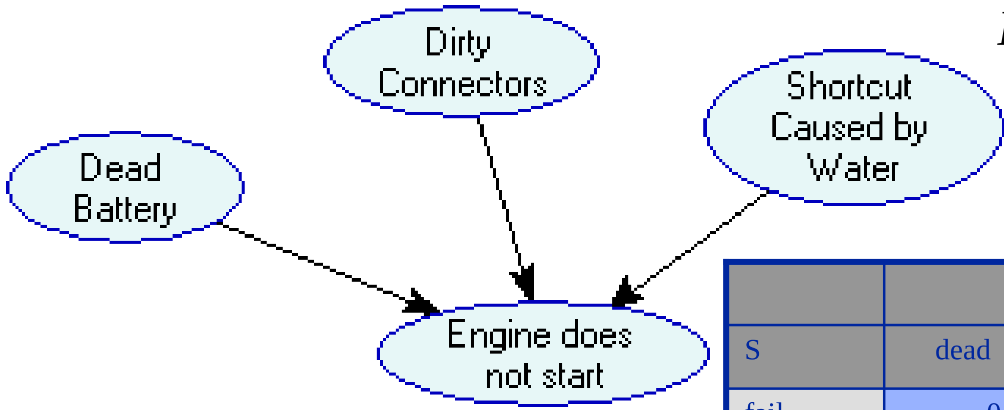
Example

Deterministic OR



DB	ok				dead			
DC	clean		dirty		clean		dirty	
S	ok	short	ok	short	ok	short	ok	short
fail	0	1	1	1	1	1	1	1
start	1	0	0	0	0	0	0	0

Noisy-OR

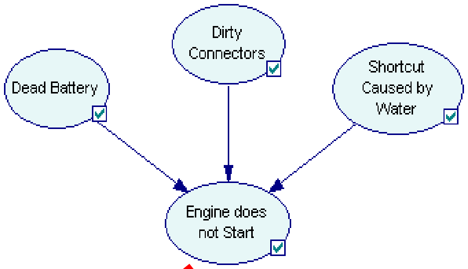


$$P(E = absent \mid C1,\ldots,Cn) = \prod_{C_i = present} (1 - q_i)$$

	DB		DC		S	
S	dead	ok	dirty	clean	short	ok
fail	0.9	0	0.8	0	0.5	0
stat	0.1	1	0.2	1	0.5	1

DB	ok				dead			
DC	clean		dirty		clean		dirty	
S	ok	short	ok	short	ok	short	ok	short
fail	0	0.5	0.8	0.9	0.9	0.95	0.98	0.99
stat	1	0.5	0.2	0.1	0.1	0.05	0.02	0.01

Leaky Noisy-OR



We use a "leak" or "background" probability to model all unmodeled causes

	DB		DC		S		leak
S	dead	ok	dirty	clean	short	ok	
fail	0.9	0	0.8	0	0.5	0	0.1
stat	0.1	1	0.2	1	0.5	1	0.9

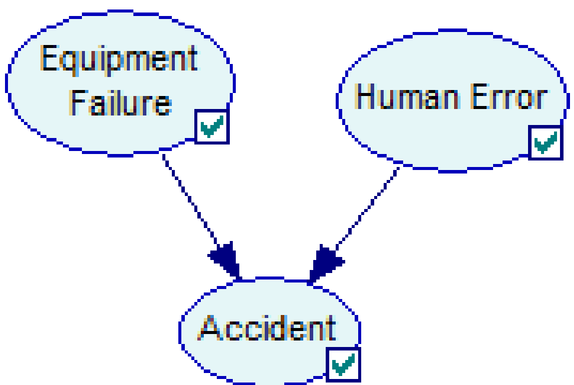
DB	ok				dead			
DC	clean		dirty		clean		dirty	
S	ok	short	ok	short	ok	short	ok	short
fail	0.1	0.5	0.8	0.888	0.9	0.944	0.977	0.987
stat	0.9	0.5	0.2	0.112	0.1	0.056	0.023	0.013

$$P(E = absent \mid C1,\ldots,Cn) = (1-q_L)\prod_{C_i = present}\frac{1-q_i}{1-q_L}$$

Noisy-AND/MIN

Based on the DeMorgan's law:

$$X \wedge Y = \neg(\neg X \vee \neg Y))$$

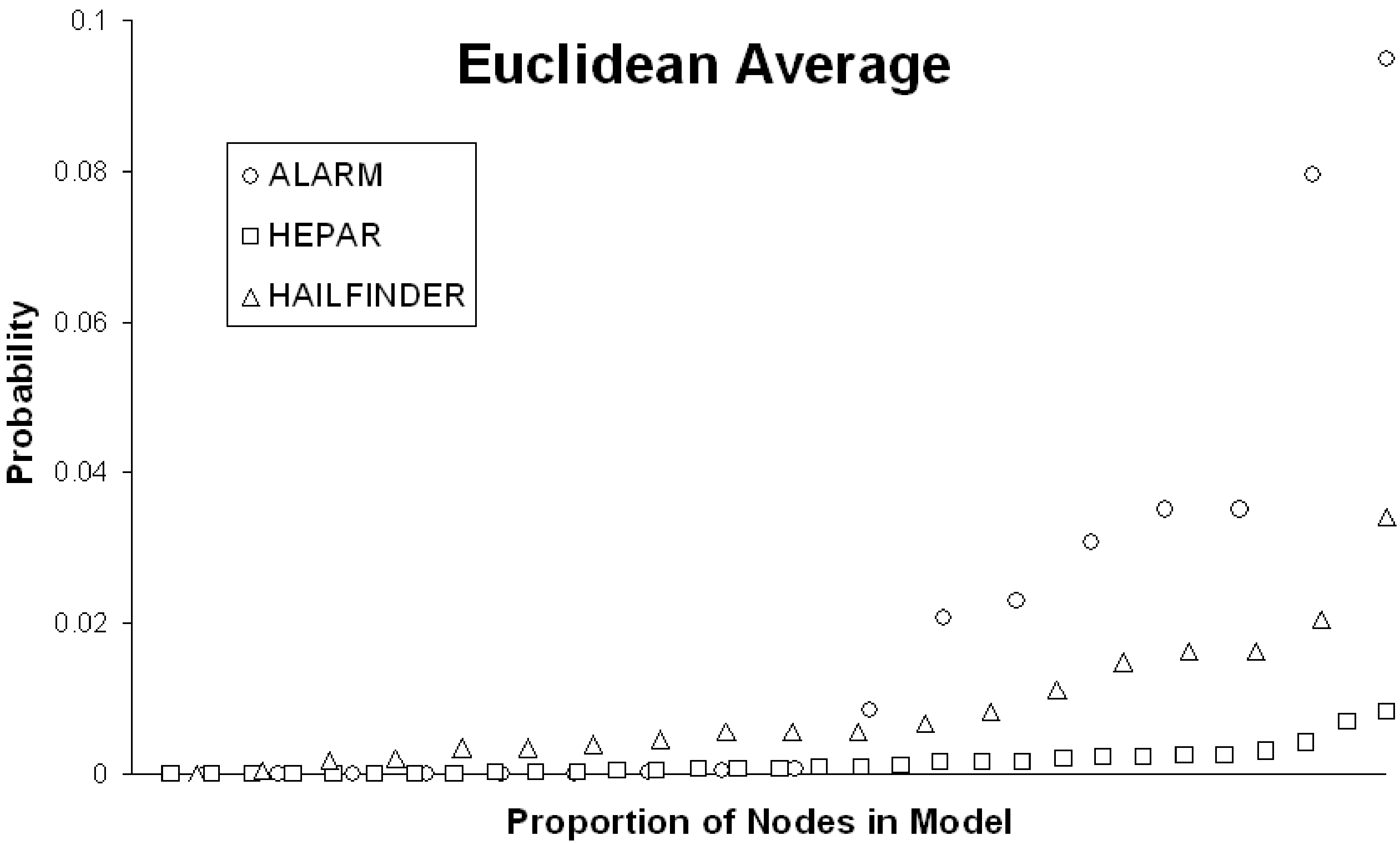


We negate all nodes by reversing the order of their states.

Equipment Failure	NoFailure		Failure	
Human Error	NoError	Error	NoError	Error
NoAccident	0.981	0.81	0.905	0.05
Accident	0.019	0.19	0.095	0.95

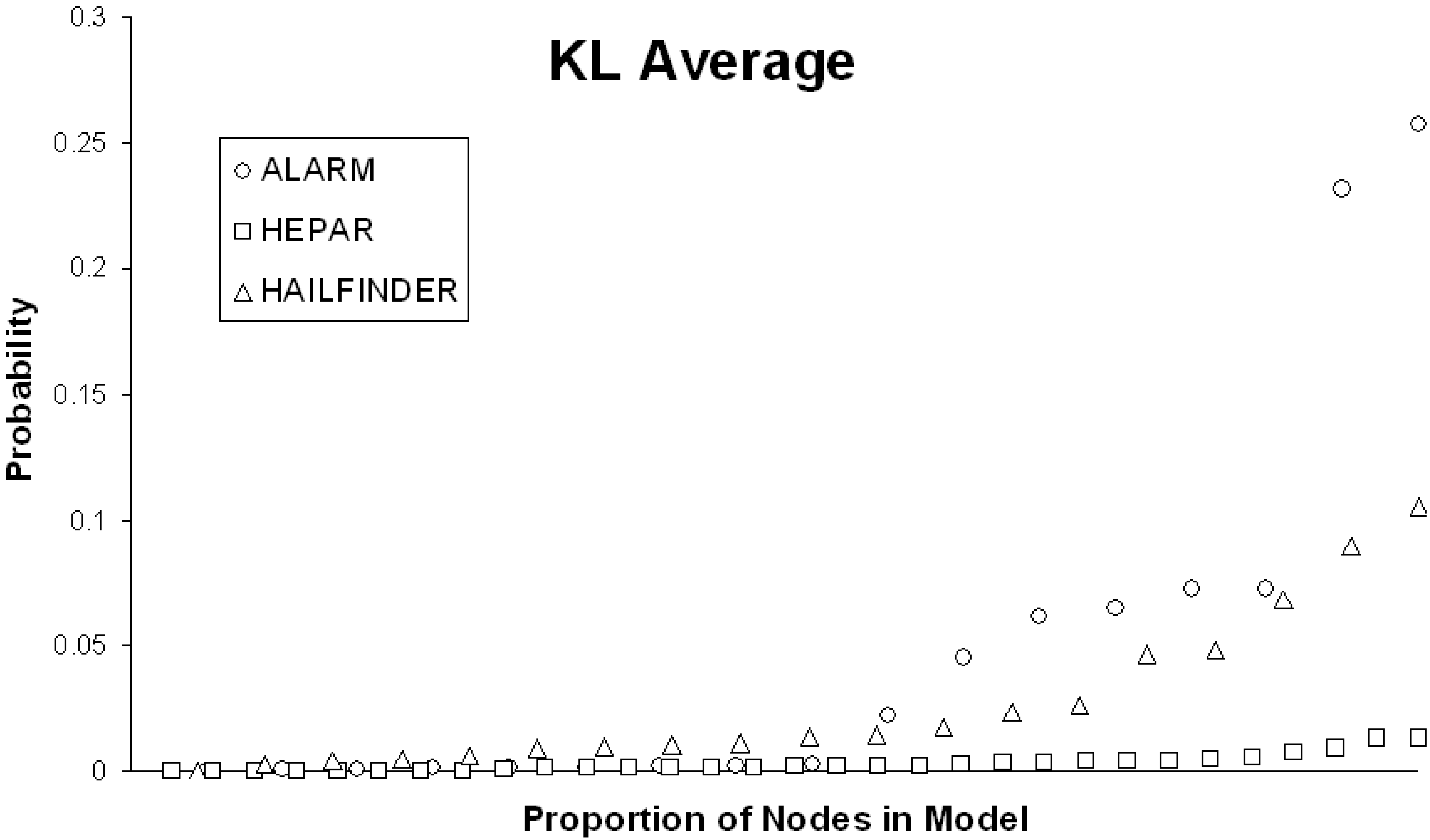
Canonical Gates in Practical Models

Noisy MAX in practical models



[Zagorecki & Druzdzel 2011]

Noisy MAX in practical models



[Zagorecki & Druzdzel 2011]

Concluding remarks

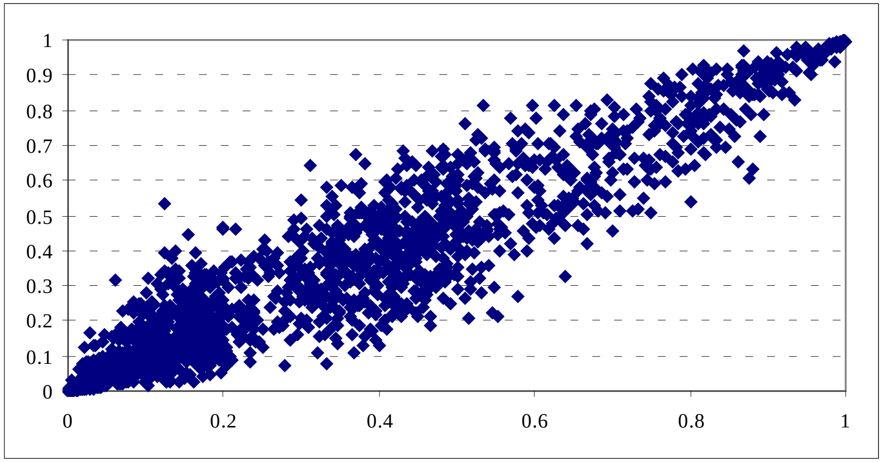
- In practical models, canonical gates are the only way to go
- There are significant computational advantages that stem from canonical gates

Do parameters matter?

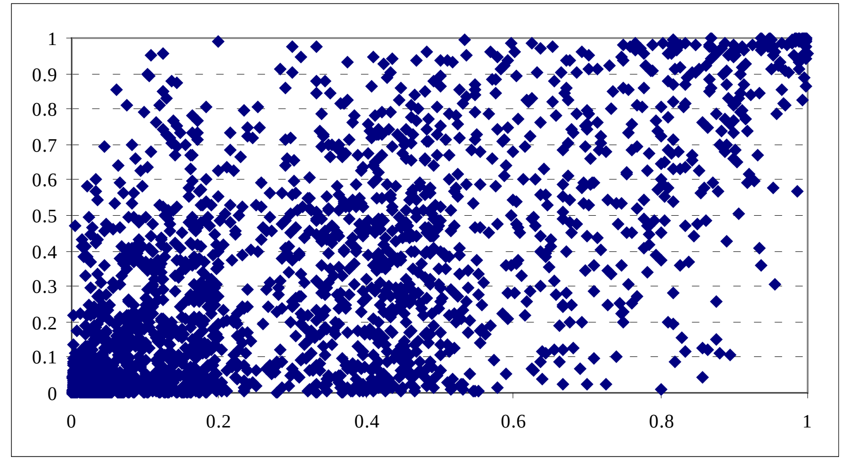
Random noise, Normal(0,σ)

transformed parameters

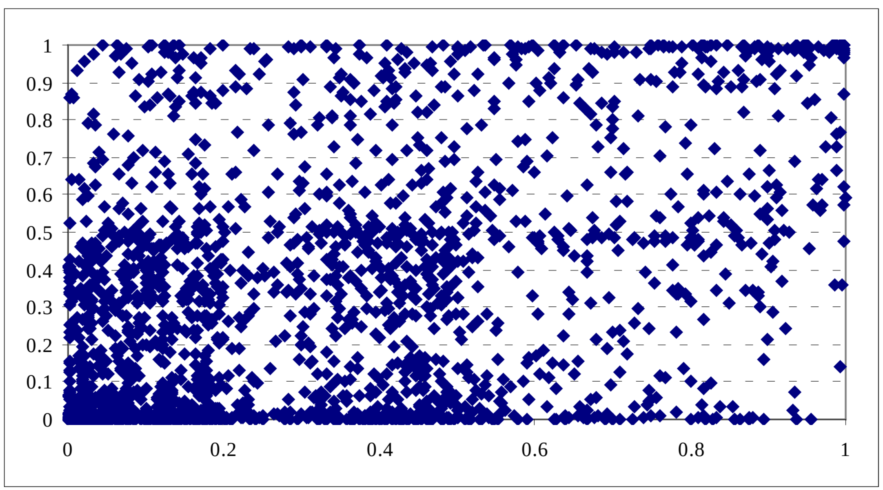
σ = 0.3



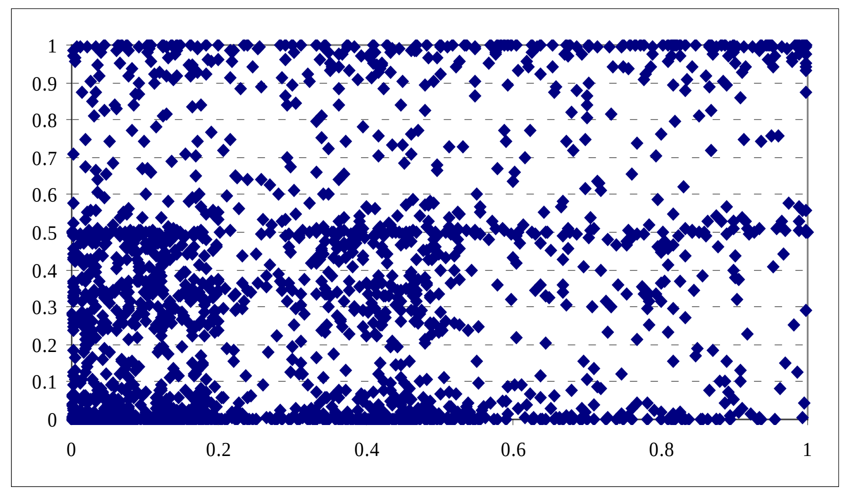
σ = 1.0



σ = 2.0



σ = 3.0

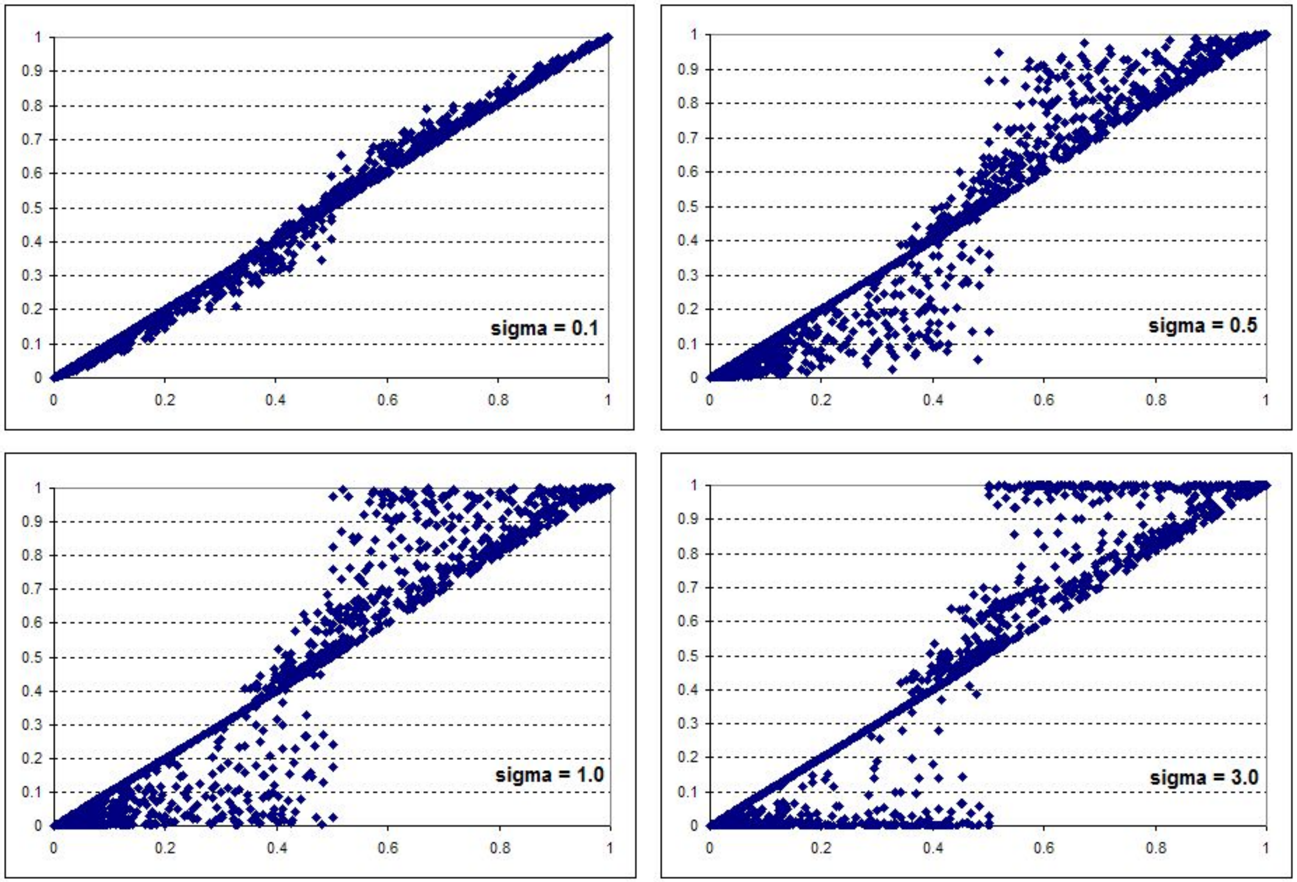


original parameters

[Onisko & Druzdzel 2011]

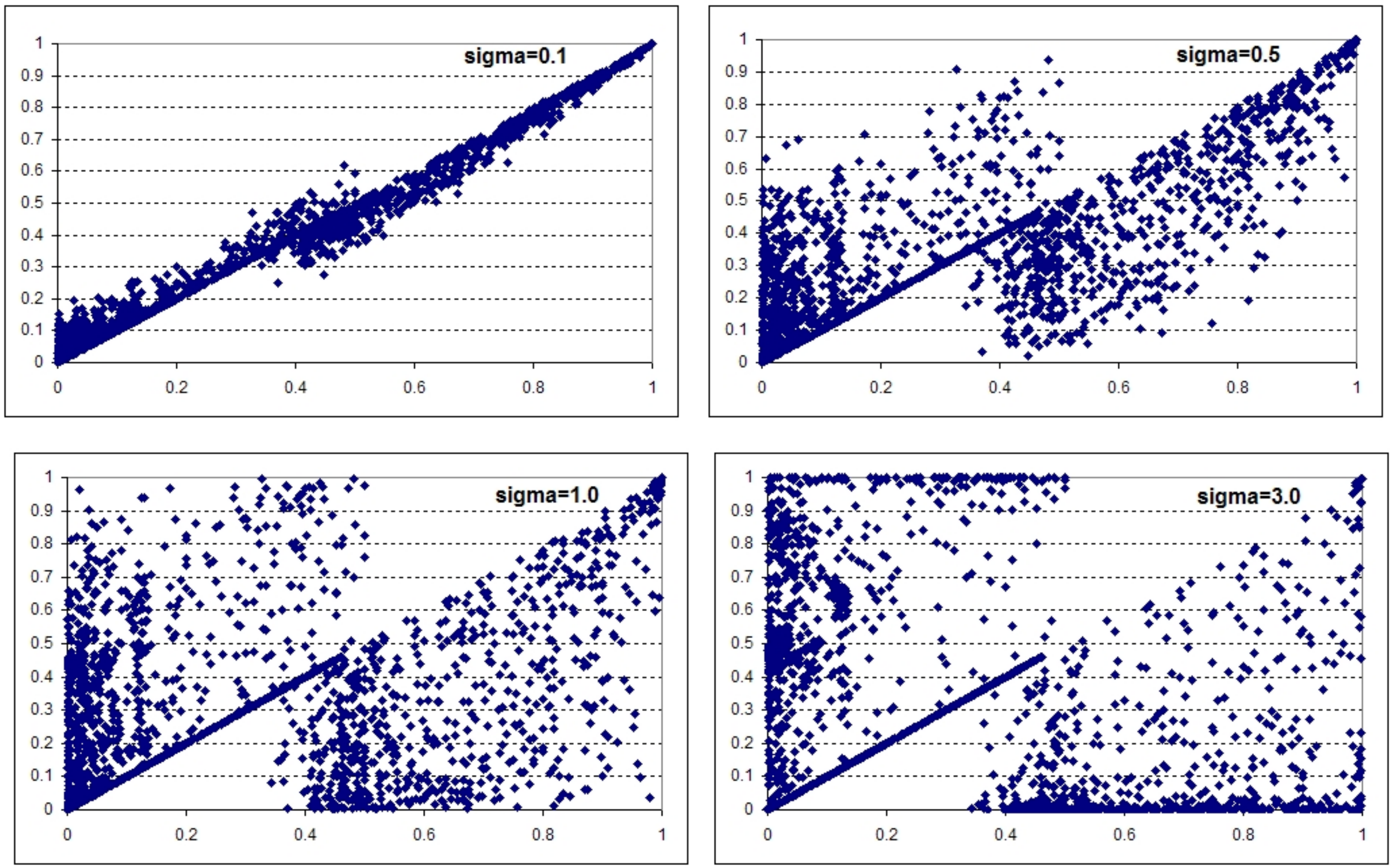
Biased noise (overconfidence), Normal(0,σ) added to the largest probability in a distribution

transformed parameters



original parameters

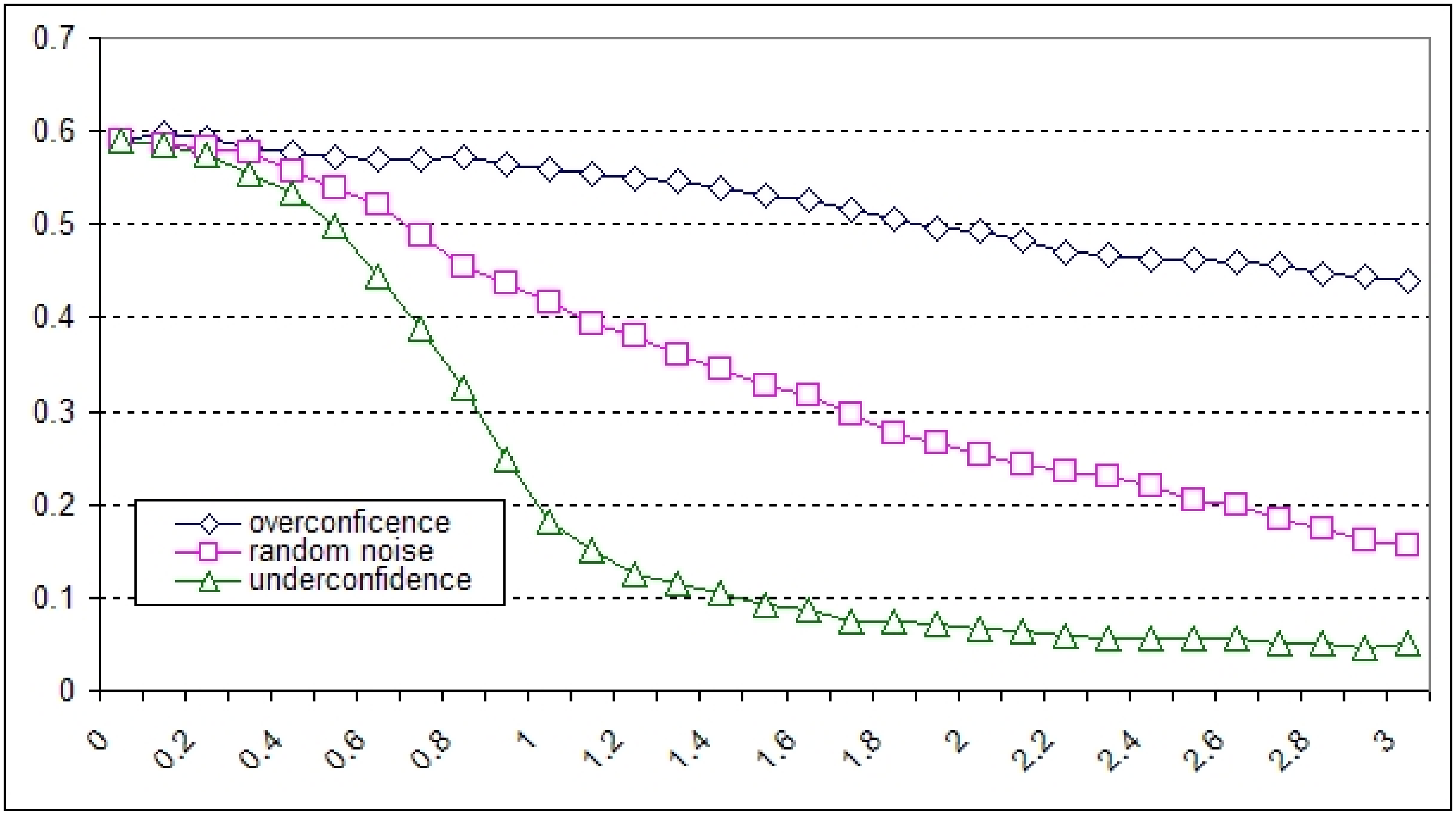
Biased noise (underconfidence), Normal(0,σ) subtracted from the largest probability in a distribution



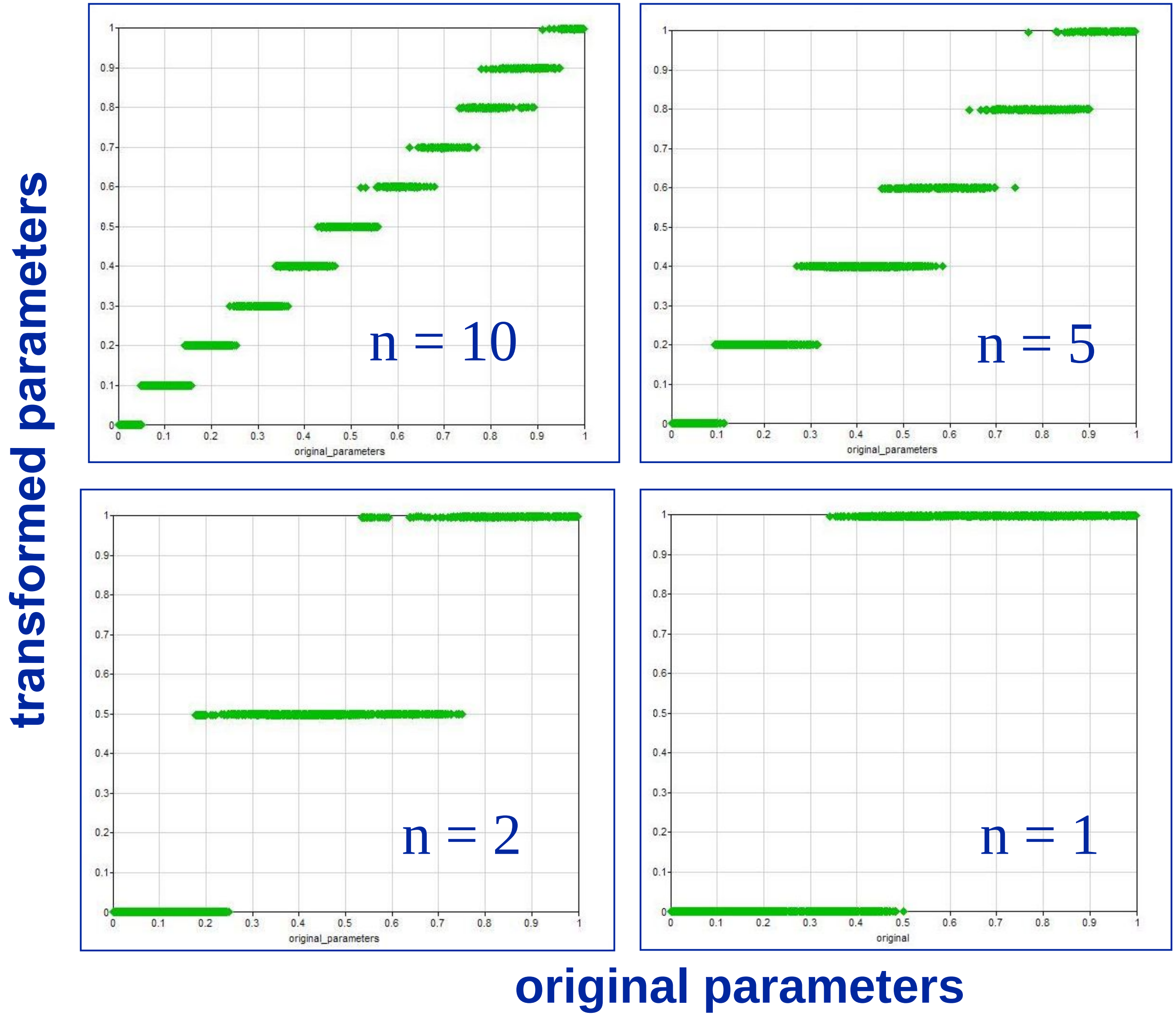
transformed parameters

original parameters

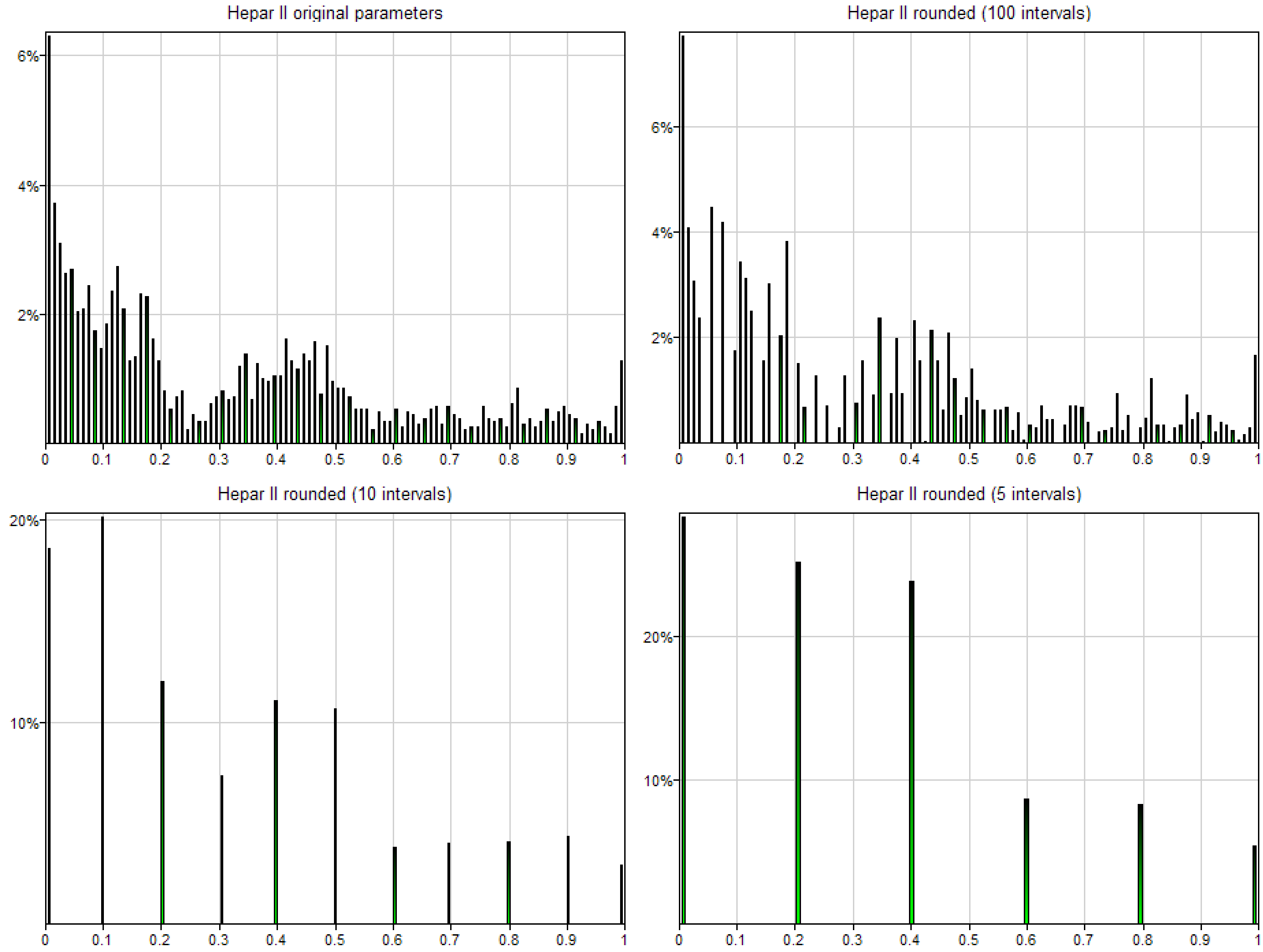
Diagnostic performance as a function of parameter accuracy



Rounded vs. original probabilities for various levels of rounding accuracy

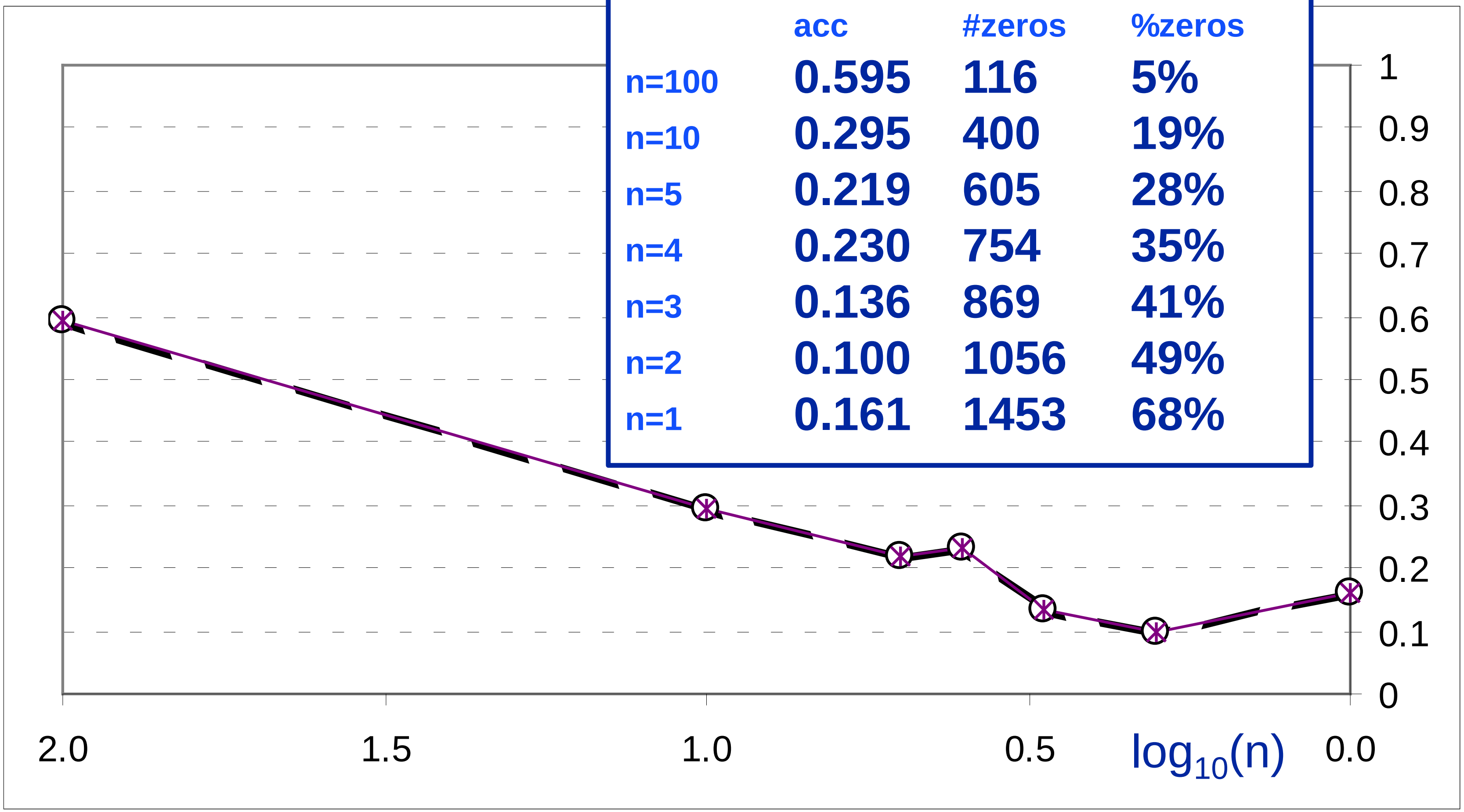


Histograms of original and rounded probabilities for various levels of rounding accuracy



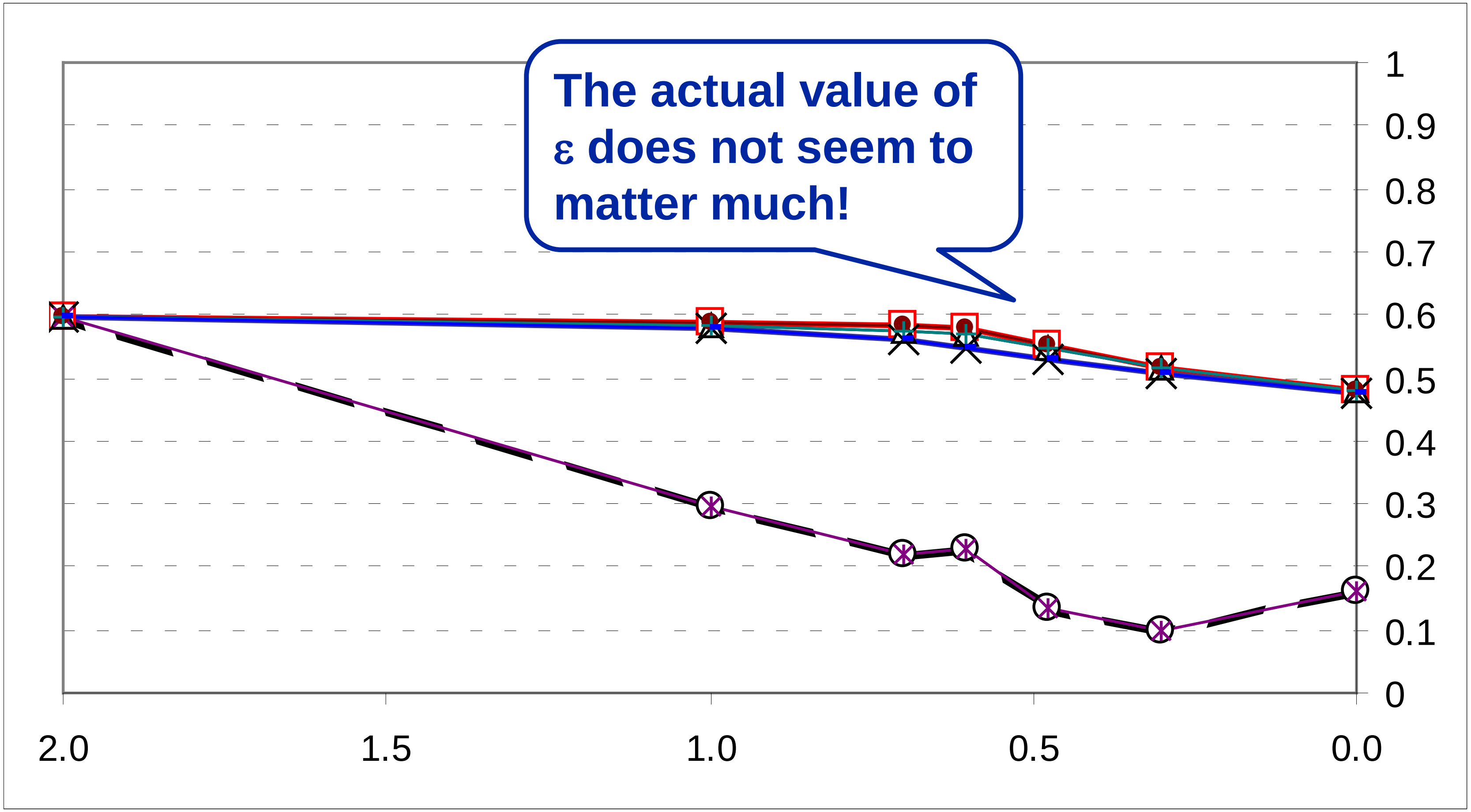
Diagnostic performance as a function of parameter accuracy (w=1)

	acc	#zeros	%zeros
n=100	0.595	116	5%
n=10	0.295	400	19%
n=5	0.219	605	28%
n=4	0.230	754	35%
n=3	0.136	869	41%
n=2	0.100	1056	49%
n=1	0.161	1453	68%



Diagnostic performance as a function of parameter accuracy and ε (w=1)

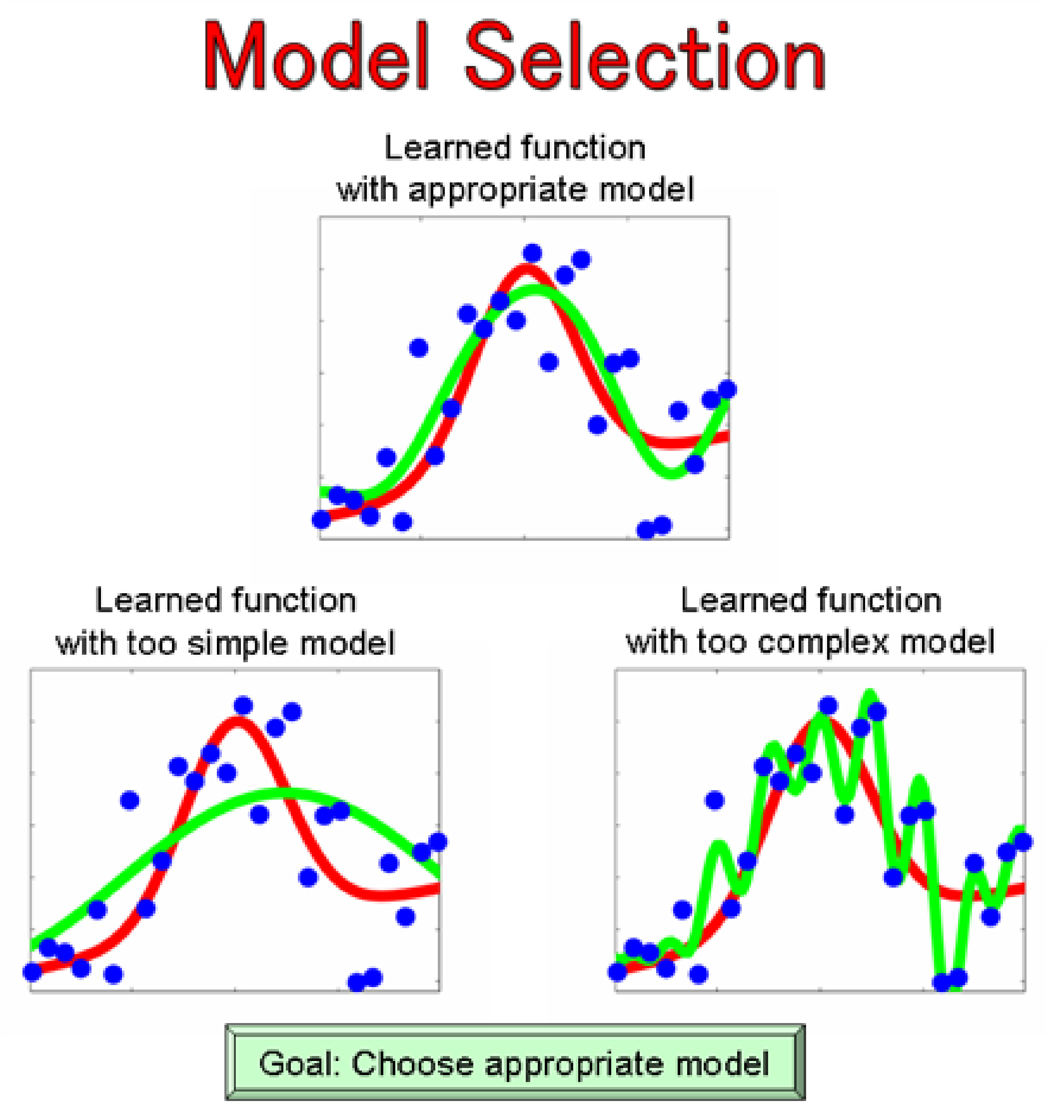
What if we replace all zeros by some small number ε ?



Model selection

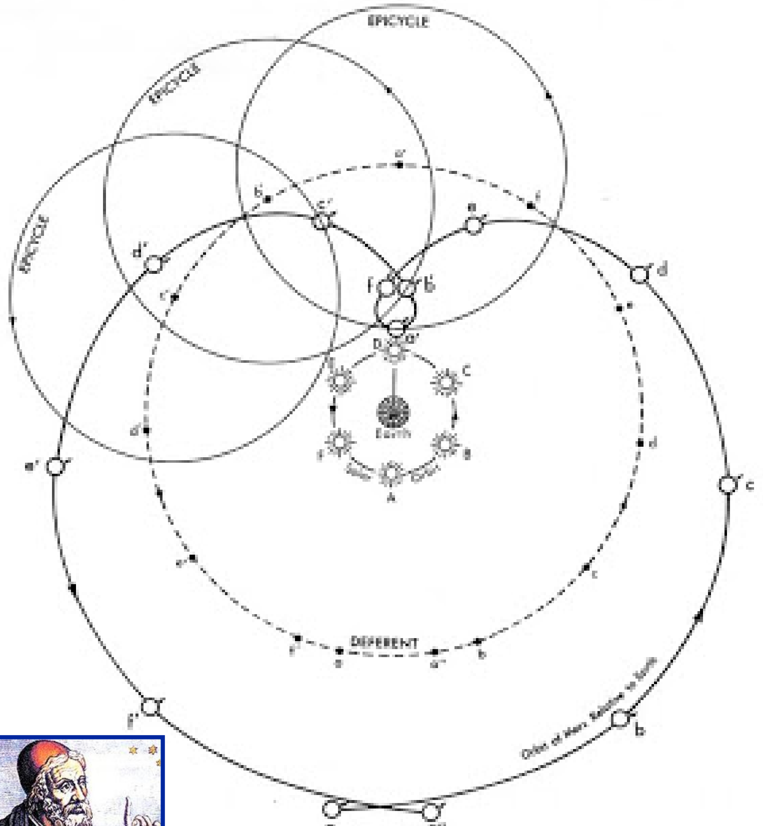
Definition

Model selection is the task of selecting a statistical model from a set of candidate models given data.

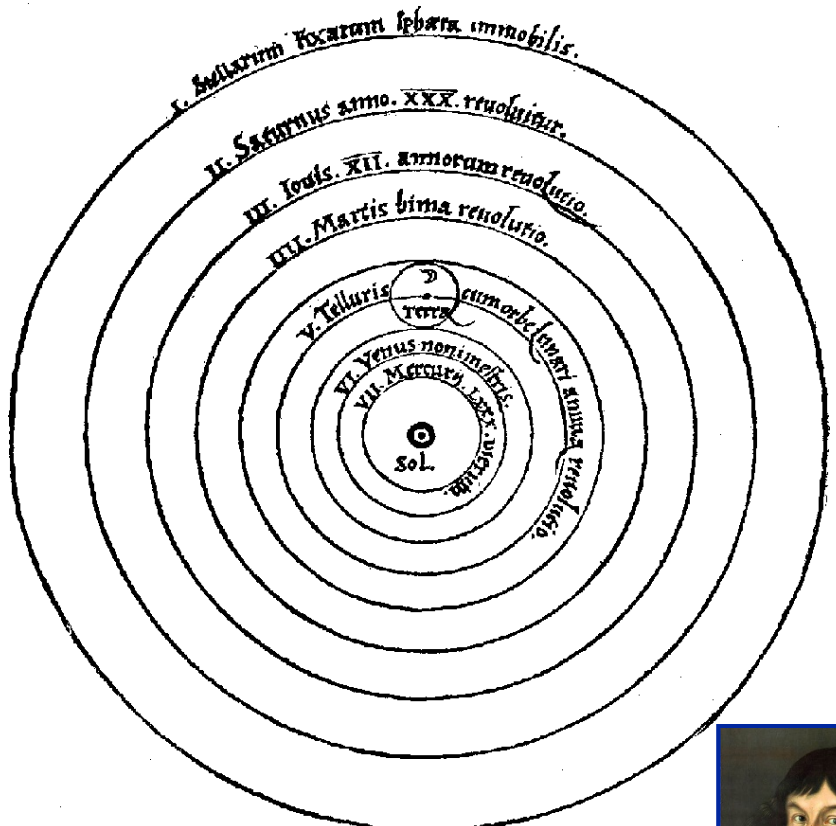


http://sugiyama-www.cs.titech.ac.jp/~sugi/research.html

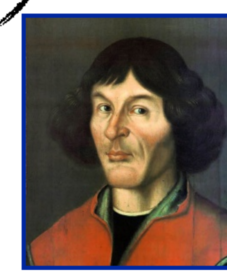
Model selection: Example



Ptolemy's model



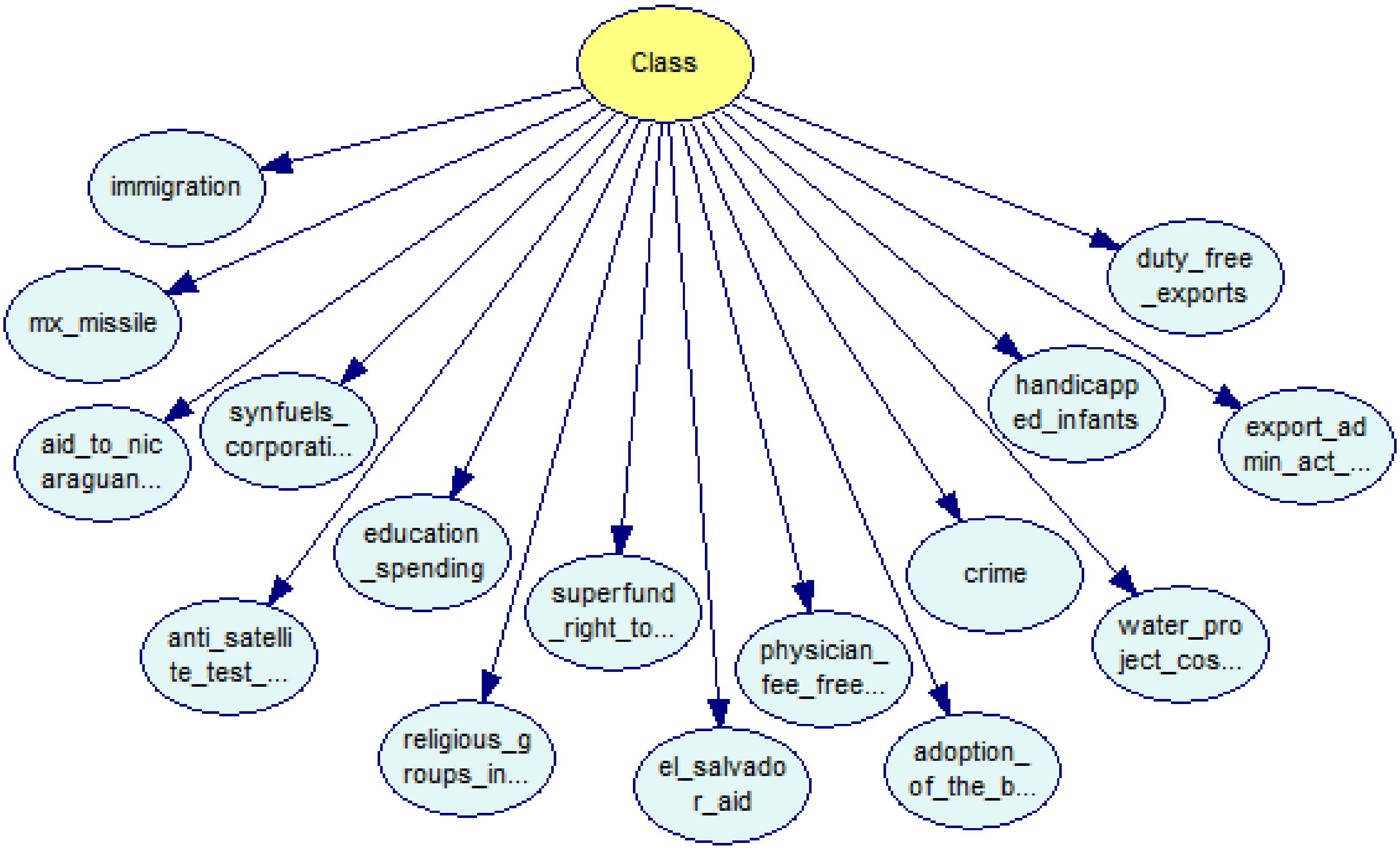
Copernicus' model



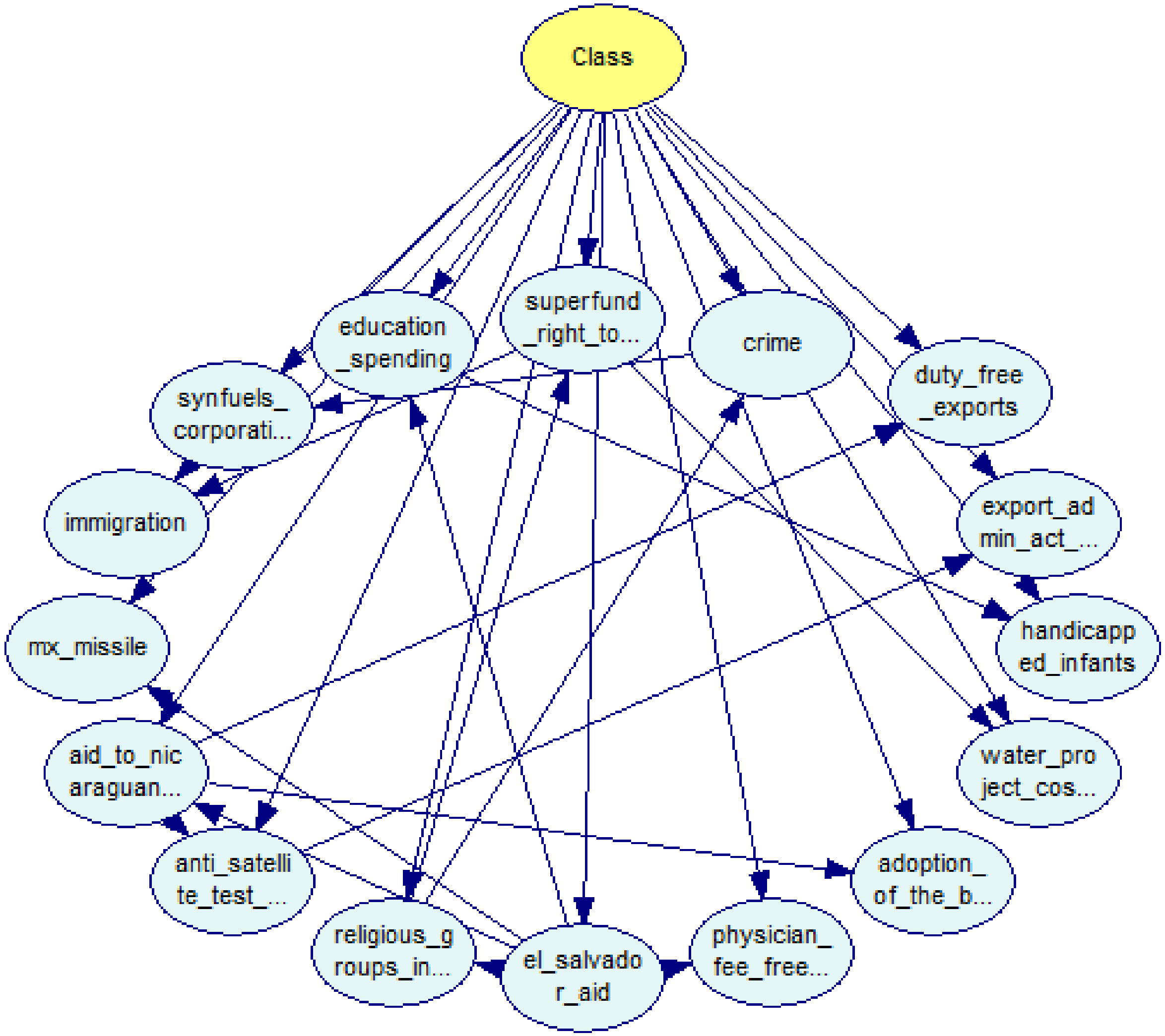
What criteria should be used to select between competing models?

Does structure matter?
Simple vs. complex models

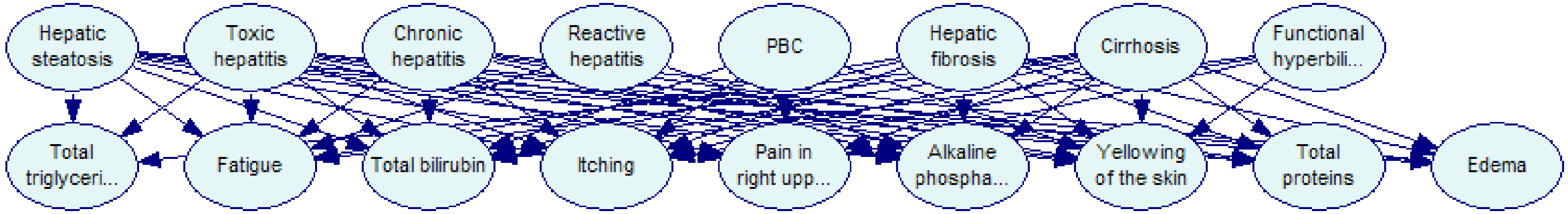
Naïve Bayes models



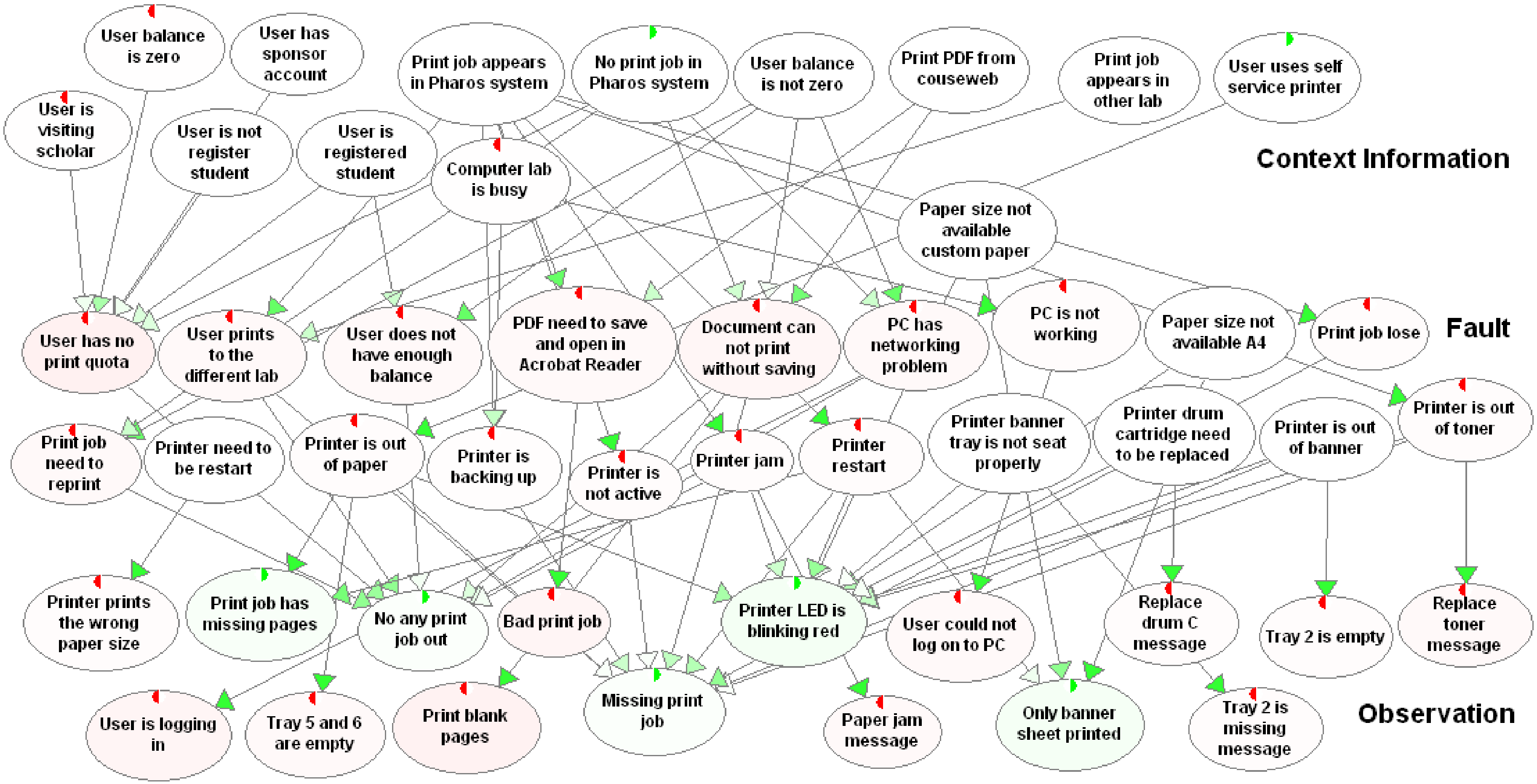
TAN (Tree Augmented Network) models



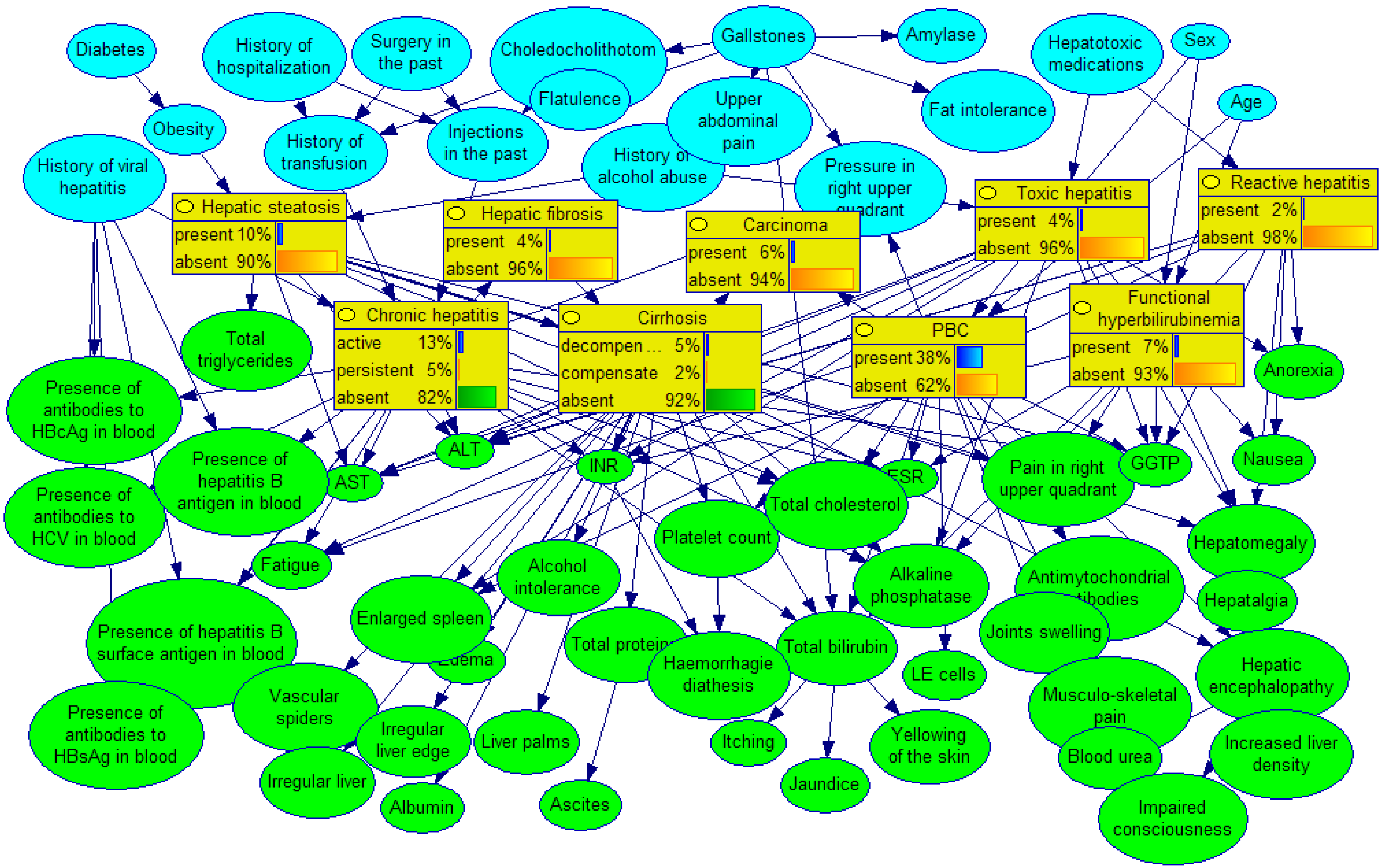
Bipartite models



BN3O Models


User balance is zero
User has sponsor account
Print job appears in Pharos system
No print job in Pharos system
User balance is not zero
Print PDF from couseweb
Print job appears in other lab
User uses self service printer
User is visiting scholar
User is not register student
User is registered student
Computer lab is busy
Context Information
Paper size not available custom paper
User has no print quota
User prints to the different lab
User does not have enough balance
PDF need to save and open in Acrobat Reader
Document can not print without saving
PC has networking problem
PC is not working
Paper size not available A4
Print job lose
Fault
Printer is out of toner
Print job need to reprint
Printer need to be restart
Printer is out of paper
Printer is backing up
Printer is not active
Printer jam
Printer restart
Printer banner tray is not seat properly
Printer drum cartridge need to be replaced
Printer is out of banner
Printer prints the wrong paper size
Print job has missing pages
No any print job out
Bad print job
Printer LED is blinking red
User could not log on to PC
Replace drum C message
Tray 2 is empty
Replace toner message
User is logging in
Tray 5 and 6 are empty
Print blank pages
Missing print job
Paper jam message
Only banner sheet printed
Tray 2 is missing message
Observation

Complete models



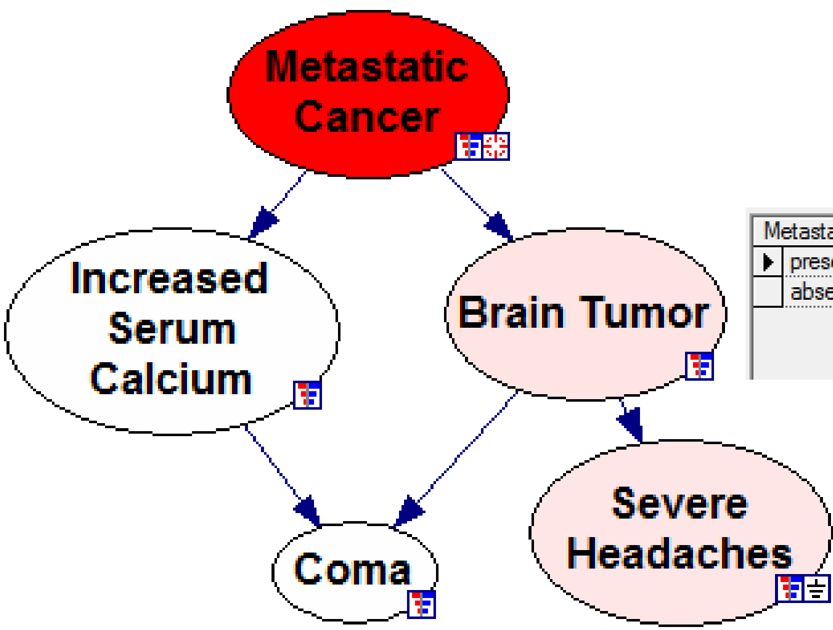
Is precision real or illusory?

- When getting the parameters from experts, we may well get better models when eliciting fewer parameters.
- When learning, the same may happen!

Other Relevant Issues

Sensitivity analysis in Bayesian networks

- Given a target node (or a set of target nodes) and a possible set of evidence nodes, we can identify the parameters that matter most for those target(s)'s posteriors.
- We compute essentially the derivative of the posterior over each of the parameters.

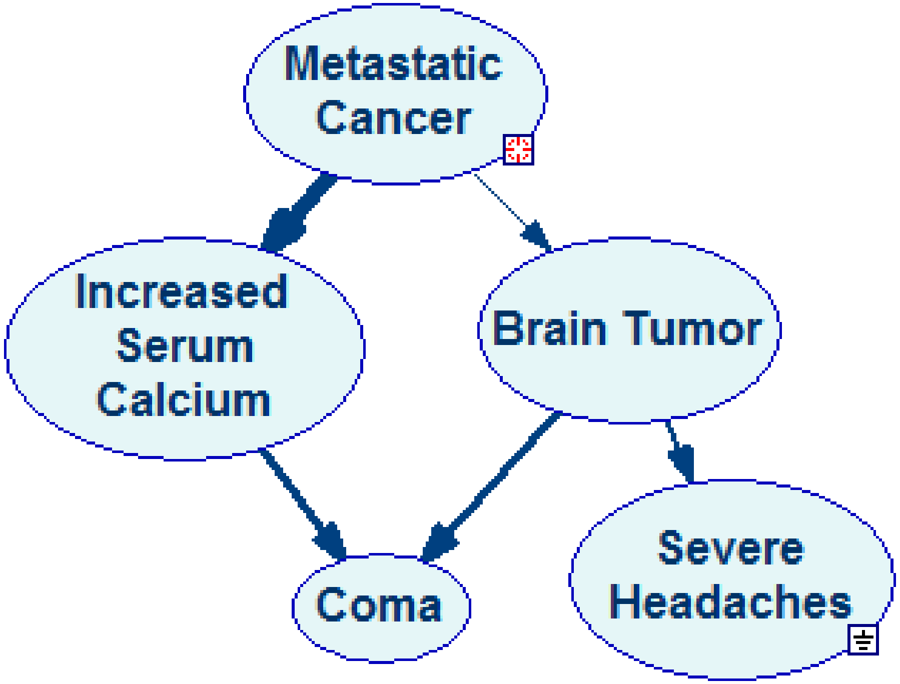


Metastatic Ca...	present	absent
present	0.2	0.05
absent	0.8	0.95

S = 0.053972
max(abs(S)) = 0.053972

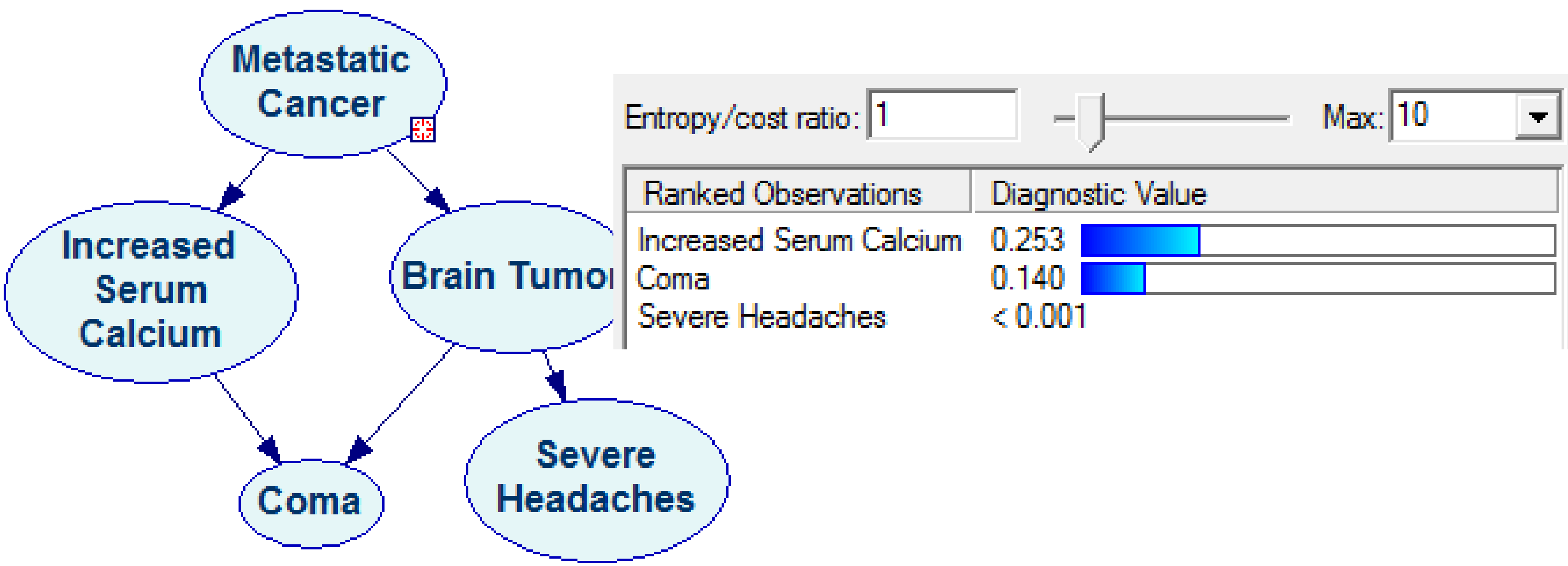
Strength of influence in Bayesian networks

- Because we have the entire joint probability distribution, we can compute individual strengths of influences between nodes
- It is essentially a measure of difference between conditional probability distributions in a CPT.



Value of information in Bayesian networks

- Because we have the entire joint probability distribution, we can compute the expected value of observations and rank-order them from the most to least informative.
- It is essentially expected cross-entropy between the targets and the individual observations.



Clarity test

- "Gas price in 1999" vs. "average regular unleaded gas price taken over all gas stations within the city of Pittsburgh on January 1 1999".
- "Market up or down" vs. "the market goes up means that the Standard & Poor's 500 Index rises".
- The matter of clarifying definitions of alternatives, outcomes, and consequences is absolutely crucial in real-world decision problems. The clarity test forces us to define all aspects of a problem with great care.