

Analysis of Survival Times Using Bayesian Networks

Helge Langseth

Presented at ESREL '98

Trondheim, Norway, 16-19 June 1998



Two types of statistical models

Neyman categorises statistical models into two groups:

- **Interpolating models**

Used merely to capture rough effects in the data

- **Explorative models**

Used to explore the underlying process which generates the data we have observed



Scope

With a database as a starting point, we want to build an explorative model to pinpoint how to *reduce the rate of critical failures* in a system components.

Our main goal is to build a model to gain *understanding* about how the covariates contribute to the system's survival times.

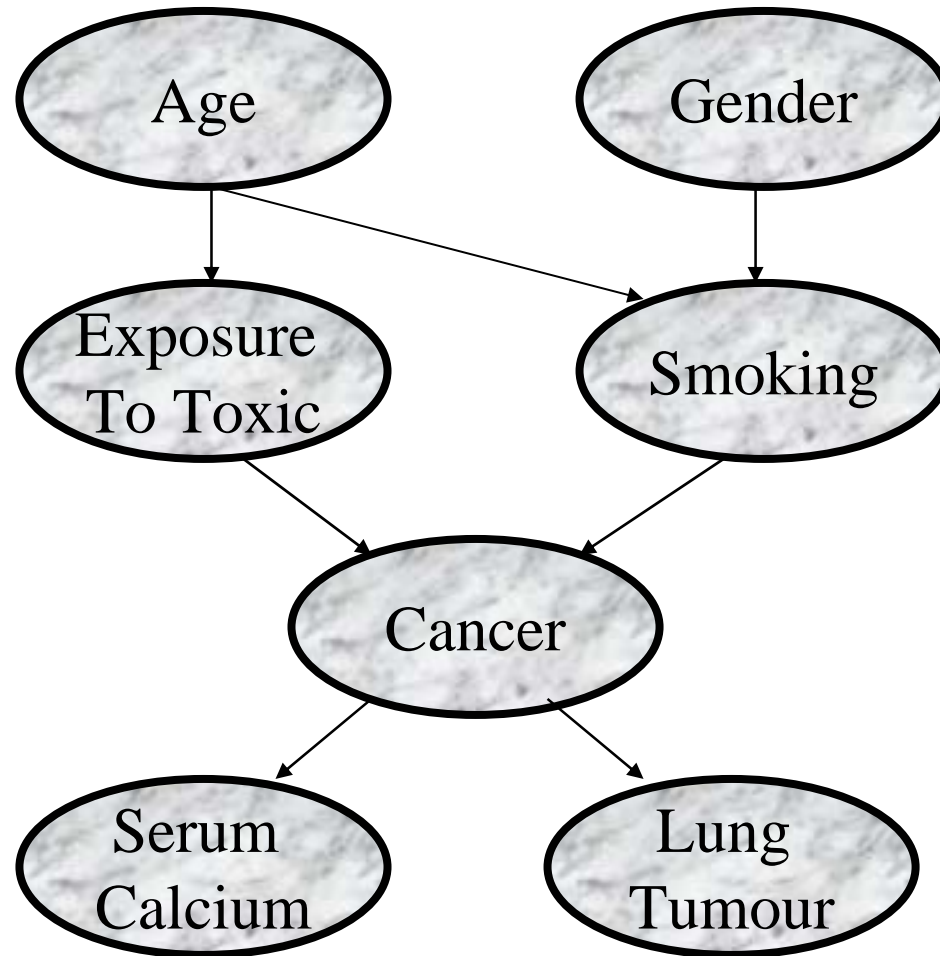


The History of Graphical Models

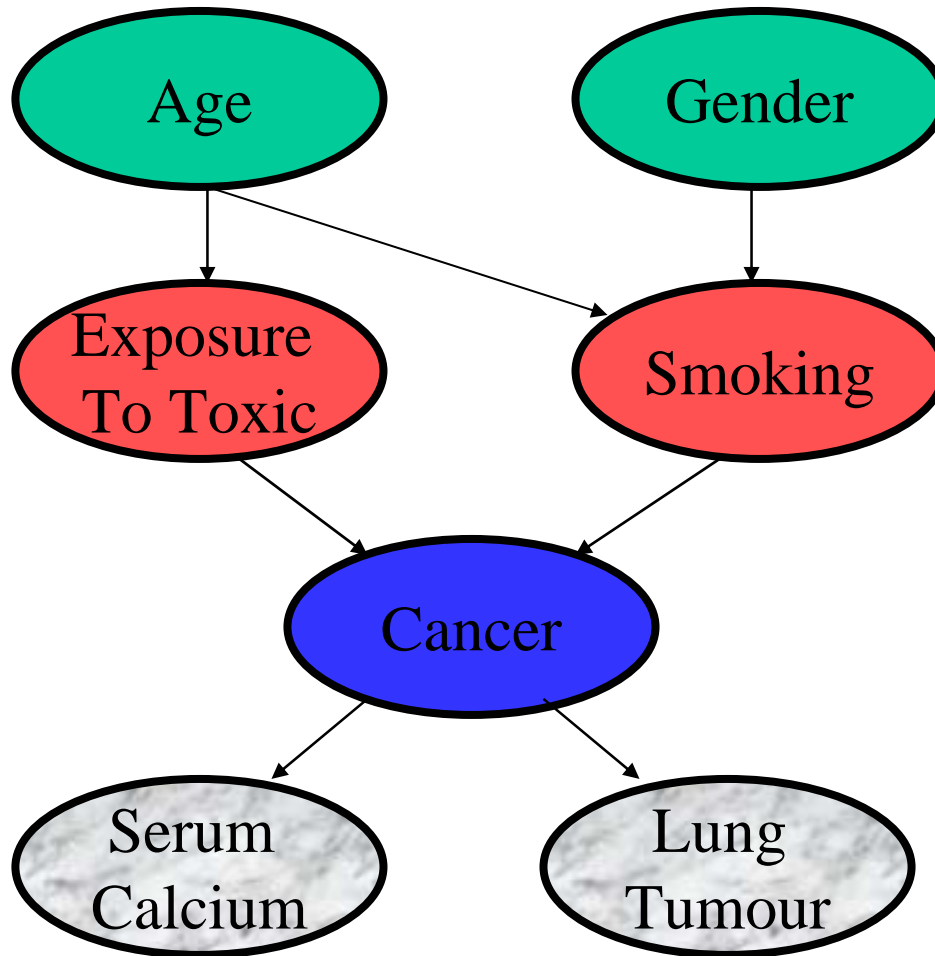
- Graphical models in statistics can be dated back to Wright's notation in 1921.
- The calculation complexity did however, render the Bayesian Networks neglected for 60 years
- In the 1980'ties, effective algorithms for exact calculations on graphs, and later on computer intensive methods like Markov-Chain Monte-Carlo brought the Bayesian Networks back into the light, and up on the *Top 5 Statistical Buzz-Word of the Week*.



Bayesian Networks



Conditional Independence



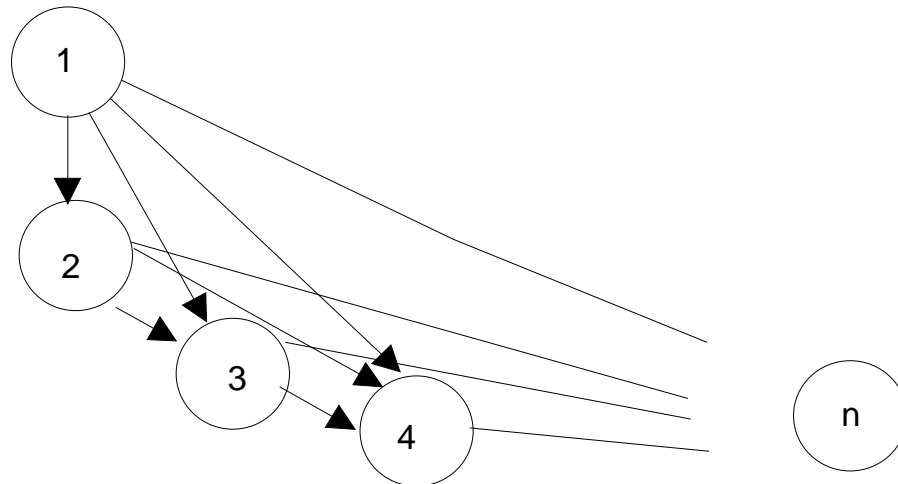
Cancer is independent of **Age** and **Gender** given **Exposure To Toxic** and **Smoking**

“Fundamental Theorem”

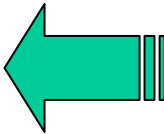
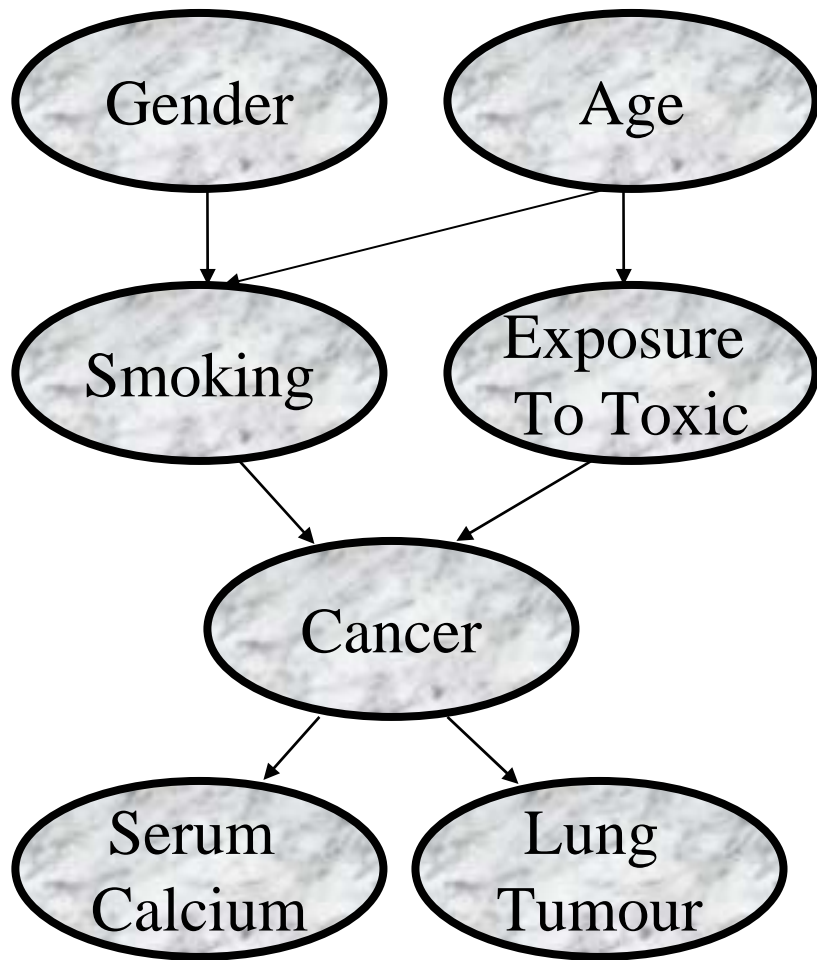
Every multidimensional statistical distribution function can be represented by a Bayesian Network.

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i | x_1, x_2, \dots, x_{i-1})$$

$$= \prod_{i=1}^n f(x_i | \text{"All predecessors"})$$



Nodes are Probability Tables



Exposed To Toxic Material	Age	
	<i>In</i> (25,65)	<i>Not In</i> (25, 65)
<i>True</i>	5 %	1%
<i>False</i>	95 %	99%

Where do the Networks come from?

Situation:

We want to build a model to analyse a multidimensional vector X .

Aid:

To do so, we have N *i.i.d.* realisations of X , x_1, \dots, x_N

AND / OR

a domain expert.

Unknowns:

- The network structure
- The parameters in the local node tables



Generating Networks:

- Initialize Network

repeat

- Propose some Change to the structure
- Fit Parameters to the new structure
- Evaluate the new network according to some measure (like BIC, AIC, MDL)
- If the New network is Better than the previous, then Keep the Change

until Finished



Bayesian Networks are used in:

- In “expert systems”, mostly in medical domains (e.g. the *MUNIN* system)
- In decision support systems (e.g. for *NASA*)
- In analysis of dynamic systems (e.g. speech recognition, the *BAT-Mobile*)
- ...



Bayesian Networks, Summary:

- An estimate of the multidimensional density
- Easy to understand for non-statisticians (e.g. a domain expert)
- The representation is optimized for tasks like
 - Prediction
 - Classification
 - Decision support
- Can incorporate prior domain knowledge:
 - “Top down analysis”: Expert knowledge
 - “Bottom up” analysis: Data driven system verification

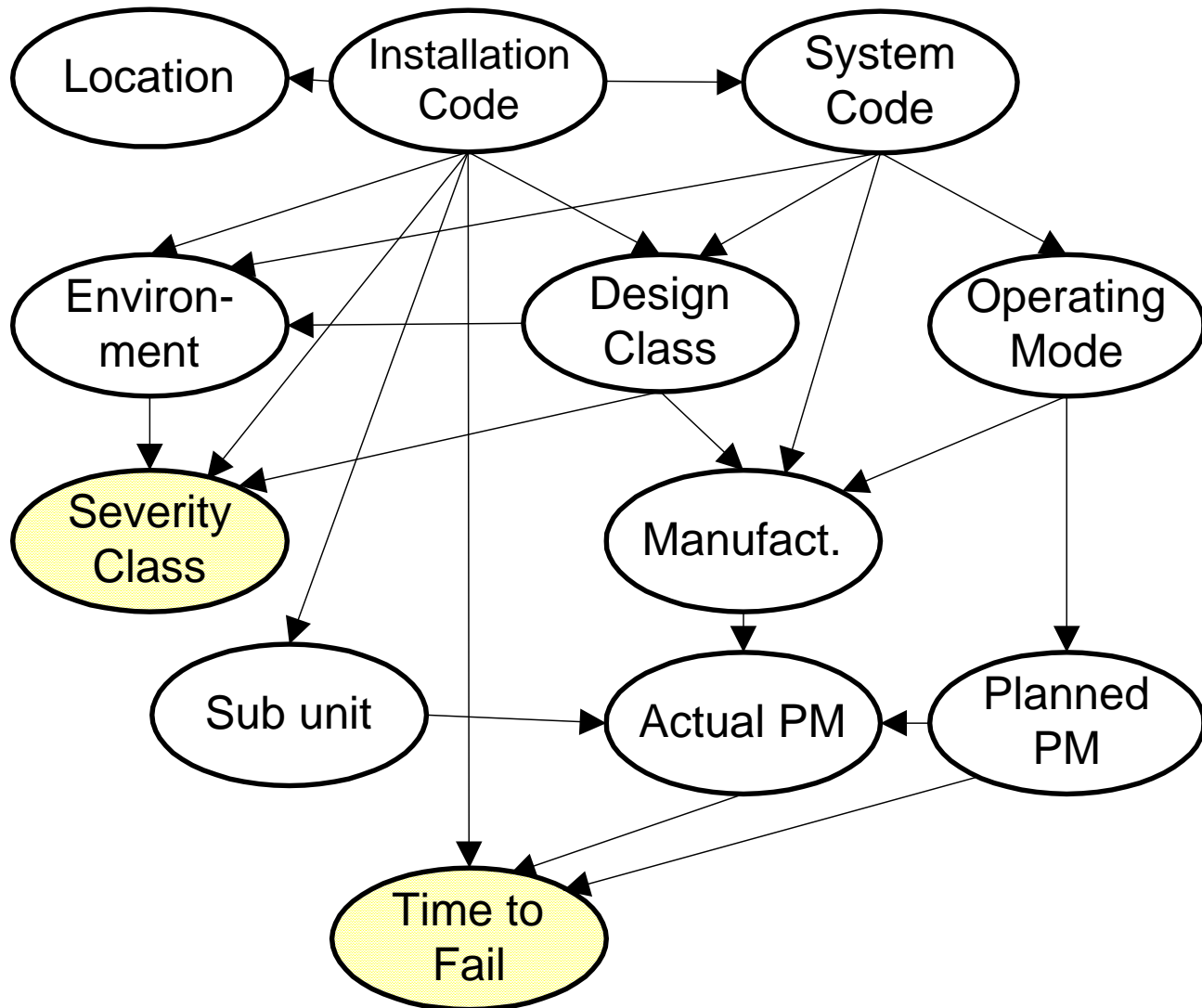


Reliability Analysis

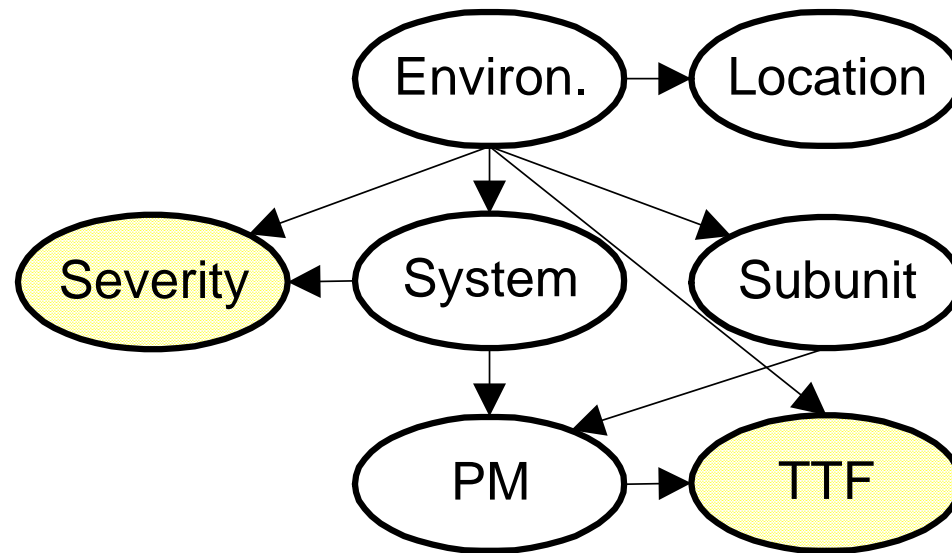
- Data-set: 219 Gas-Turbines with 2921 failures and 300 censored survival times from the OREDA-IV database
- Each failure is described by ten covariates, e.g., *System Type, Manufacturer, Actual/Planned PM,...*
- We have special interest in *Time To Fail* and *Failure Severity* (**Critical** or **Degraded**)
- Problem to solve: “*How can we reduce the frequency of critical failures?*”



Generated Network



“Clique” Graph



Location: Installation
Code Location

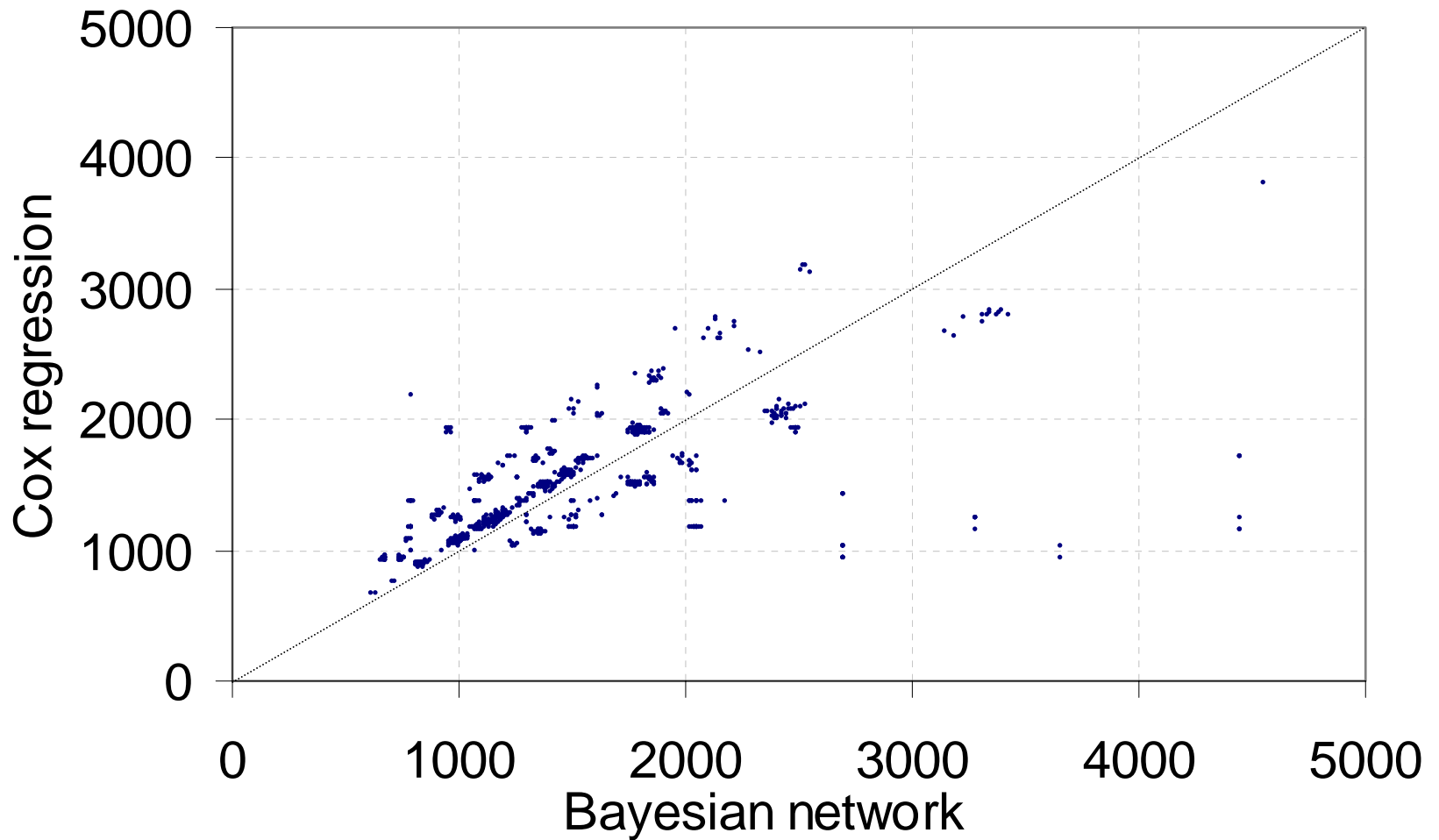
PM: Planned PM
Actual PM

System: System Code
Operating Mode
Manufacturer

Environment: Installation
Code
Environment
System Code
Design Class



Model Verification



Conclusions

- We have generated a Bayesian Network to analyse a data-set from the OREDA IV database.
- The Bayesian Network enabled both *Qualitative* and *Quantitative* analysis of the data-set.
- To verify the calculations, the numerical results were compared to those found by Cox regression. The results of the two methods were at the same level.

