BAYESIAN NETWORKS

FUNDAMENTALS

AARHUS BSS
DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

MARCH 2024

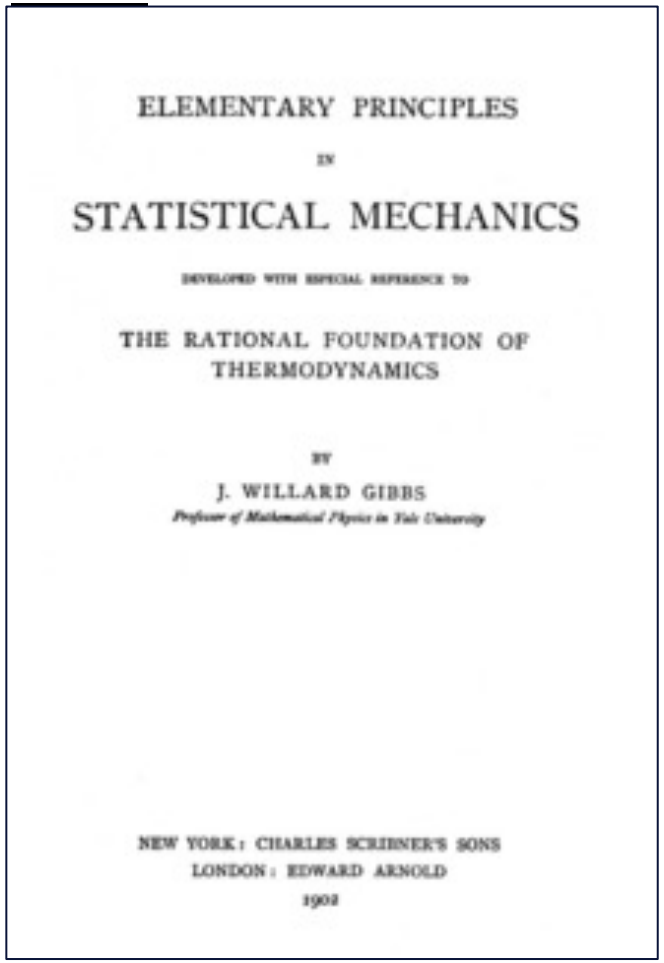
ANA ALINA TUDORAN, PHD
ASSOCIATE PROFESSOR

AGENDA

- Discuss BN within the map of analytical models and definitions
- Brush up basic concepts, including simple BN examples
- Discuss BN algorithms
- Discuss technical issues behind BN

ORIGINS

Statistical mechanics, physics (Gibbs, 1902)



ELEMENTARY PRINCIPLES

IN

STATISTICAL MECHANICS

DEVELOPED WITH SPECIAL REFERENCE TO

THE RATIONAL FOUNDATION OF THERMODYNAMICS

BY

J. WILLARD GIBBS

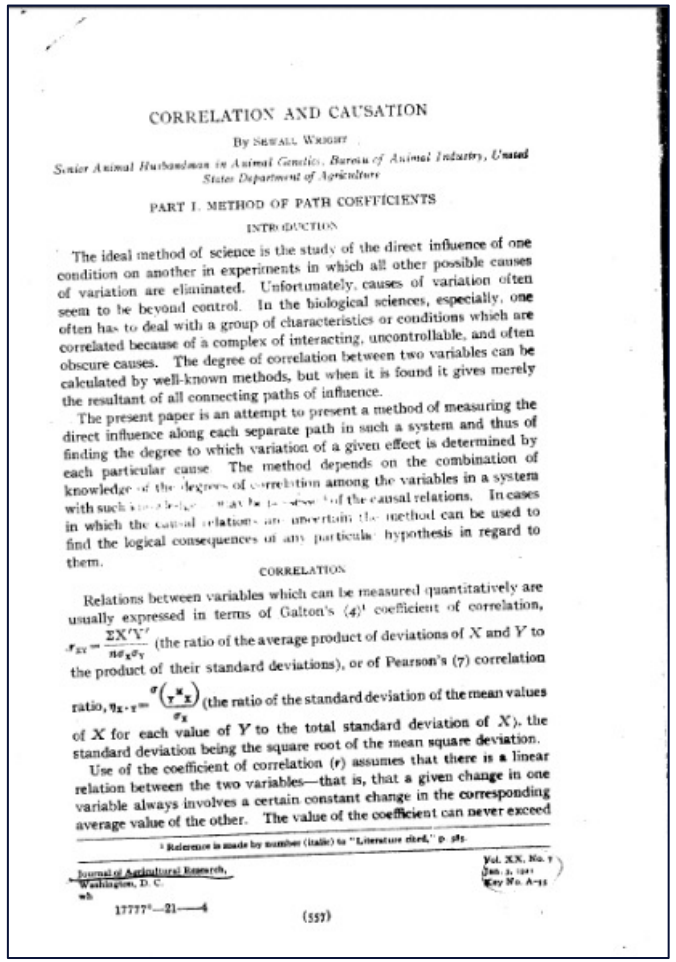
Professor of Mathematical Physics in Yale University

NEW YORK: CHARLES SCRIBNER'S SONS

LONDON: EDWARD ARNOLD

1902

Path analysis (Wright, 1921)



CORRELATION AND CAUSATION

By SEWALL WRIGHT

Senior Animal Husbandman in Animal Genetics, Bureau of Animal Industry, United States Department of Agriculture

PART I. METHOD OF PATH COEFFICIENTS

INTRODUCTION

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such [illegible] of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.

CORRELATION

Relations between variables which can be measured quantitatively are usually expressed in terms of Galton's (4)[1] coefficient of correlation, $r_{xy} = \frac{\Sigma x'y'}{n\sigma_x\sigma_y}$ (the ratio of the average product of deviations of X and Y to the product of their standard deviations), or of Pearson's (7) correlation ratio, $\eta_{x \cdot y} = \frac{\sigma\left(\frac{M_x}{y}\right)}{\sigma_x}$ (the ratio of the standard deviation of the mean values of X for each value of Y to the total standard deviation of X), the standard deviation being the square root of the mean square deviation.

Use of the coefficient of correlation (r) assumes that there is a linear relation between the two variables—that is, that a given change in one variable always involves a certain constant change in the corresponding average value of the other. The value of the coefficient can never exceed

[1] Reference is made by number (italic) to "Literature cited," p. 585.

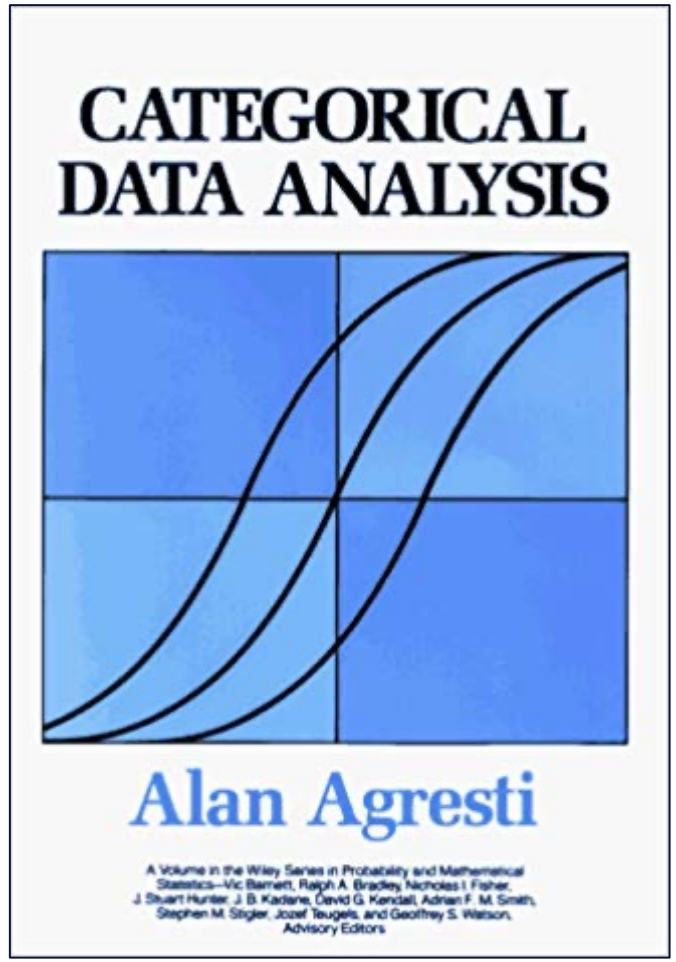
Journal of Agricultural Research,
Washington, D. C.
wb

Vol. XX, No. 7
Jan. 3, 1921
Key No. A-11

17777°—21——4

557

Loglinear models (Agresti 1984; 1990)



CATEGORICAL DATA ANALYSIS

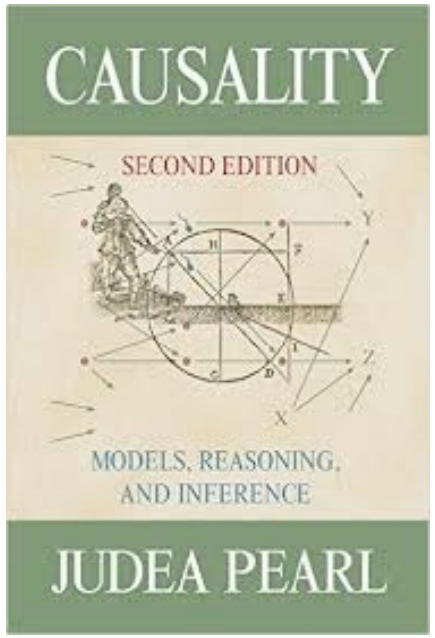
Alan Agresti

A Volume in the Wiley Series in Probability and Mathematical Statistics—Vic Barnett, Ralph A. Bradley, Nicholas I. Fisher, J. Stuart Hunter, J. B. Kadane, David G. Kendall, Adrian F. M. Smith, Stephen M. Stigler, Jozef Teugels, and Geoffrey S. Watson, Advisory Editors

AARHUS BSS

FOUNDATION

- Professor **Judea Pearl**
- 2011 **A.M. Turing Award**
- revolutionized the field of artificial intelligence
- important tool for engineering and the natural and social sciences



CAUSALITY IN OBSERVATIONAL DATA

- Search for causal models in observational data
- In practice most of the **search is informal** ☹
- For example:
  - Some background assumptions about the causal links
  - Set one single model and apply a statistical test
  - If the model is rejected based on the test, the model is modified until it will pass the test.
  - The **reliability** of the model depends on the correctness of the background assumptions

- But, for reliable causal inference, it is not sufficient to find one model that passes a statistical test.
- We need to find and evaluate all such models (Spirtes, 2010).

BN AS PARTIAL SOLUTION

- BN were designed as partial solutions to causal inference from observational data
- BN are automated search algorithms over large search spaces
- Some challenges remain when working with observational data if:
    - small sample sizes
    - unmeasured causes in the dataset

DEFINITIONS

1. A class of graphical models that allow a concise representation of the probabilistic dependencies between a given set of random variables $X = \{X_1, X_2, ..., X_p\}$ as a directed acyclic graph (DAG) (Nagarajan et al 2013).

2. A form of multivariate analysis that uses graphs to represent a model as a network of interactions between variables.

3. BN is a stochastic data-mining technique that applies knowledge from computer science, probability theory, information theory, logic, machine learning, and statistics, in order to obtain information for the construction of decision support systems.

WHY “BAYESIAN”?

Based on Rev. Thomas Bayes´s Theorem who proposed a rule to update probabilities in the light of new evidence

$$P(A|B) = P(A) \bullet \frac{P(B|A)}{P(B)}$$

PROOF OF BAYES RULE

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \qquad p(B|A) = \frac{p(A \cap B)}{p(A)}$$

$$p(A|B) \times p(B) = p(B|A) \times p(A)$$

$$p(A|B) = p(A) \bullet \frac{p(B|A)}{p(B)}$$

TERMINOLOGY

$$p(A|B) = p(A) \bullet \frac{p(B|A)}{p(B)}$$

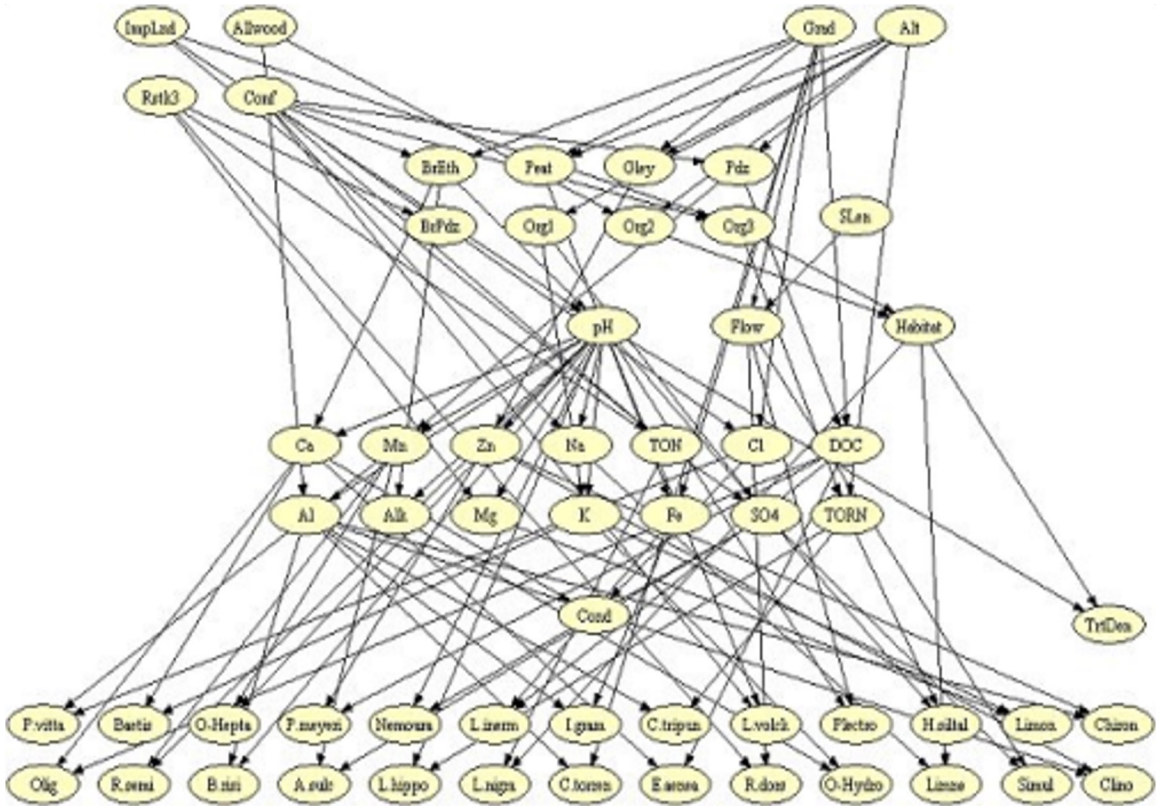
A – effect

- P (A) is the prior (unconditional probability), without taking into account any information about B.
- P(A|B) is the posterior (conditional probability of A, given B). It takes into account information about B.

B – plausible cause or predictor

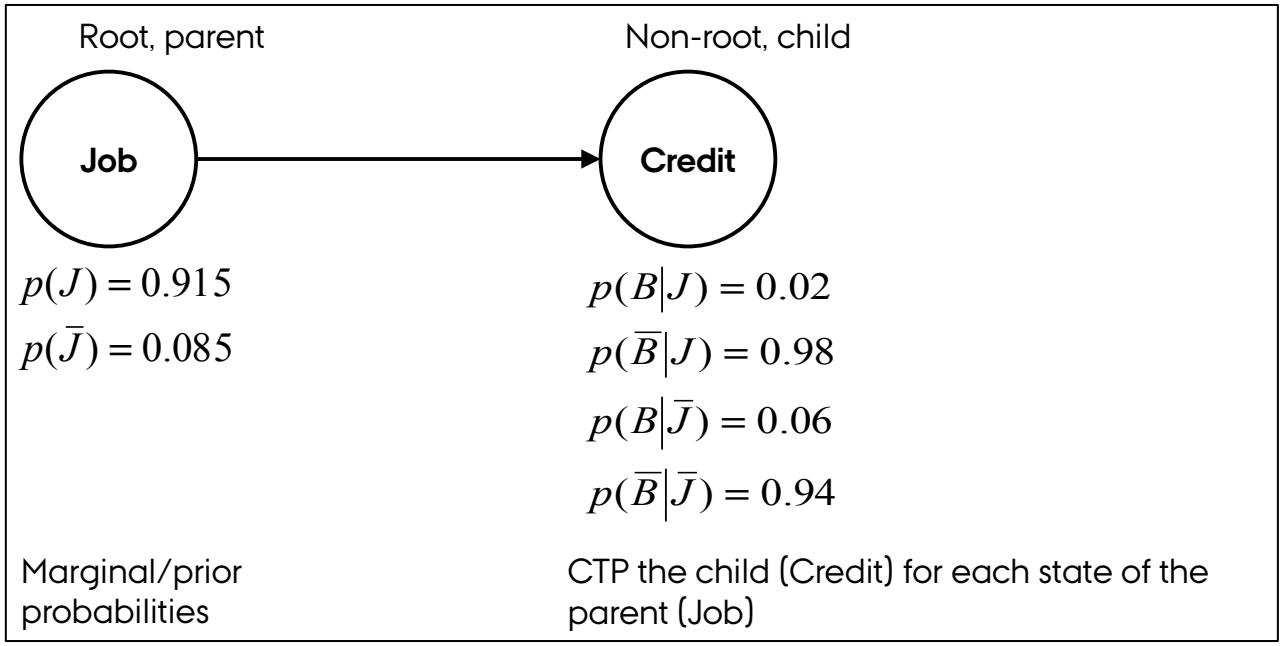
- P(B) is the marginal probability of the B.
- P(B|A) is the likelihood (conditional probability of B, given A).

BN - OTHER NAMES



- Causal probabilistic networks (Jensen et al., 1990b)
- Probabilistic graphical models (Lauritzen,1995)
- Bayesian belief networks (Cheng et al., 1997)
- Belief networks (Darwiche, 2002)
- Causal networks (Heckerman, 2007)
- Directed graphs (Wasserman, 2010)
- Probabilistic expert systems (Glenn Shafer (1996)
- Influence diagrams (Shachter, 1986a)

BN CHARACTERISTICS



① The structure/DAG where

nodes = variables

arcs = links = relationships

② The strength of these relationships is defined by the Conditional Probability Tables attached to each node.

AARHUS BSS

EXAMPLE

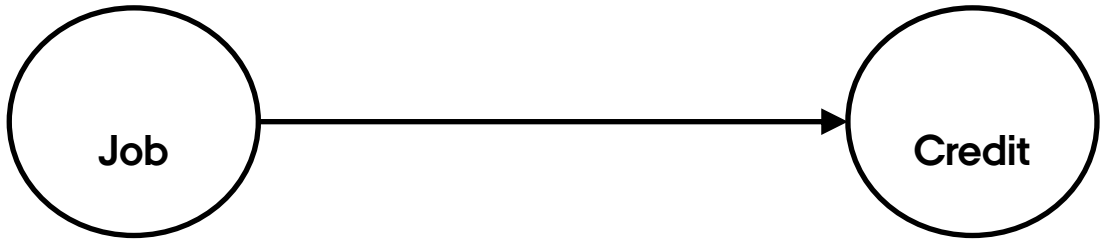
Data: n = 1000 customers from a financial institution

ID	Name	**Job (J)**	Age	Other variables	**Credit (B)**
1	J.B.B.	Yes	22	…	Bad
2	R.E.	Yes	54	…	Good
3	T.U.	No	35	…	Bad
…	…	…	…	…	…
i	…	…	…	…	…
…	…	…	…	…	…
1000	J.Ø.	Yes	54	…	Good

BUILDING (CONT.)

Marginal probabilities for the root node

Conditional probabilities for the non-root nodes



BUILDING (CONT.)

Marginal probabilities

J	$\bar{J}$
91.5%	8.5%

$$p(J) = 0.915$$
$$p(\bar{J}) = 0.085$$

$\bar{B}$	B
98%	2%

$$p(\bar{B}) = 0.98$$
$$p(B) = 0.02$$

Joint probabilities

	$\bar{B}$	B	
J	90%	1.5%	91.5%
$\bar{J}$	8%	0.5%	8.5%
	98%	2%	100%

$$p(\bar{B} \cap J) = 0.90$$
$$p(\bar{B} \cap \bar{J}) = 0.08$$
$$p(B \cap J) = 0.015$$
$$p(B \cap \bar{J}) = 0.005$$

BUILDING (CONT.)

- Conditional probabilities obtained using conditional probability formula:

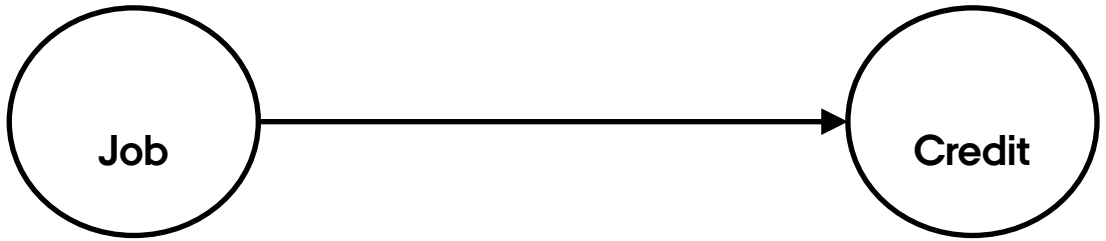
$$p(B|\bar{J}) = \frac{p(B \cap \bar{J})}{p(\bar{J})} = \frac{0.005}{0.085} = 0.06$$

"the conditional probability that the applicant will be a good credit, given he has no permanent job."

BUILDING (CONT.)

Marginal probabilities for the root node

Conditional probabilities for the non-root nodes



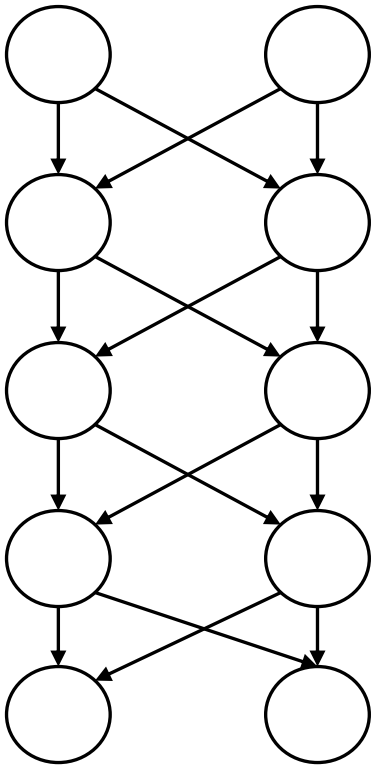
$$p(J) = 0.915$$
$$p(\bar{J}) = 0.085$$

$$p(B|J) = 0.02$$
$$p(\bar{B}|J) = 0.98$$
$$p(B|\bar{J}) = 0.06$$
$$p(\bar{B}|\bar{J}) = 0.94$$

CONDITIONAL VS. JOINT PROBABILITIES

- The number of joint probabilities required to build the net can be huge in big networks

Assume the figure attached is a simple BN and that all variables are binary.



# of Joint probabilities:

$$2^{10} = 1024 \quad (actually\ 2^{10} - 1)$$

# of Conditional probabilities:

$$2 + 2 + 8 * 8 = 68$$

Note: if we eliminate the redundant parameters (p(A)=1-p(nonA)), there are only 34 values required.

Q&A

AARHUS
BSS

INFERENCE

- When making inference, the analyst uses the BN to **update the posterior probability** in the light of new evidence.

- Although not so evident at this point, this is one of **the essence of BN:** In the light of new evidence and conditioning on multiple variables, the use of BN facilitates the estimation of conditional probabilities in big networks

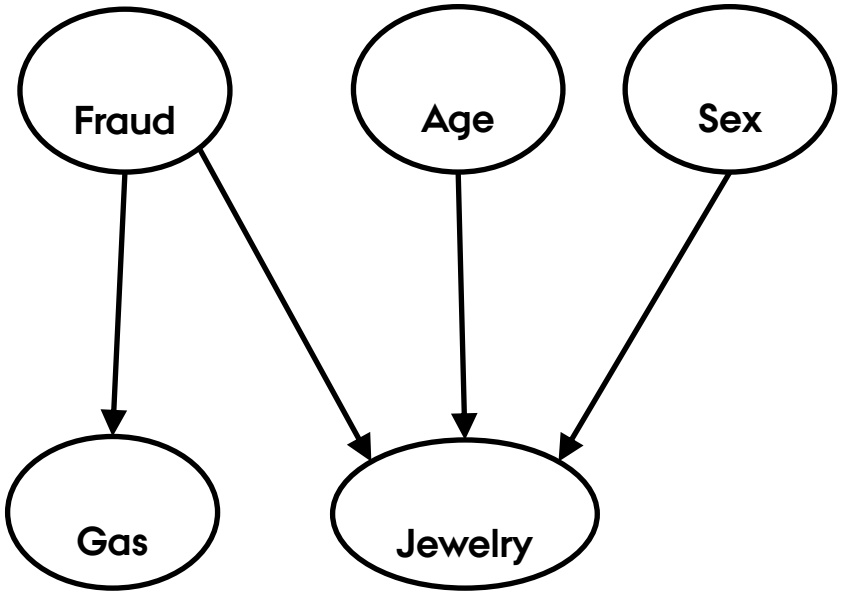
EXAMPLE

Objective:

- A financial institution wants to build a BN to predict fraudulent use of a credit card by its customers

Previous knowledge:

- A credit card thief is more likely to buy gas and jewelry
- A middle age women is more likely to purchase jewelry



- Build the structure
- Define prior probabilities of the root nodes (Fraud, Age, Sex) and conditional probabilities of the children nodes (Gas, Jewlery) for every combination of the states of its parents

BUILDING

$p(A =< 30) = 0.25$

$p(A = 30\,to\,50) = 0.40$

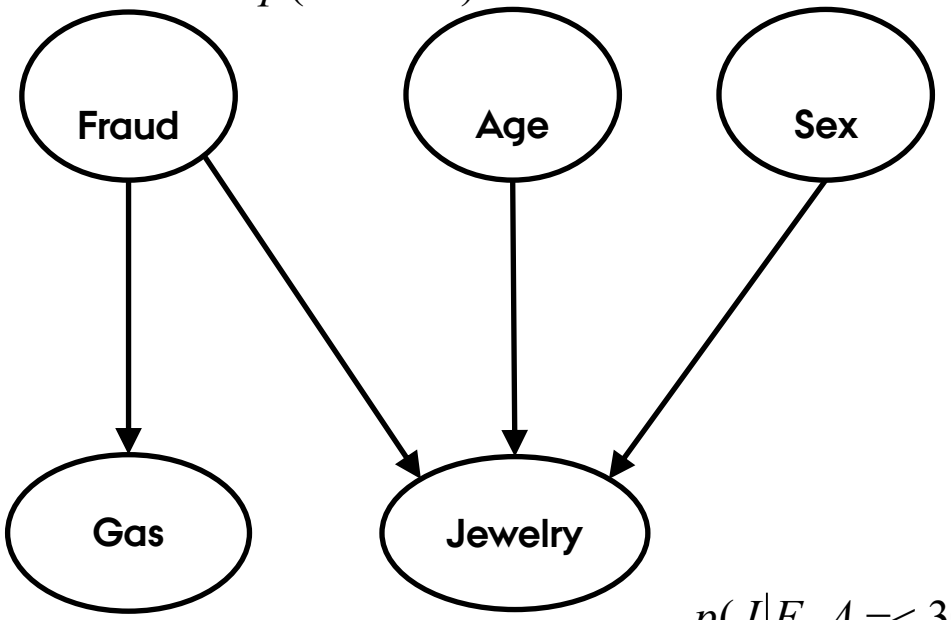
$p(A => 50) = 0.35$

$p(S = male) = 0.5$

$p(S = female) = 0.5$

$p(F) = 0.00001$

$p(\overline{F}) = 0.99999$



$p(G|F) = 0.2$

$p(\overline{G}|F) = 0.8$

$p(G|\overline{F}) = 0.01$

$p(\overline{G}|\overline{F}) = 0.99$

$p(J|F, A =< 30, S = male) = 0.05$

$p(\overline{J}|F, A =< 30, S = male) = 0.95$

.....

$p(J|\overline{F}, A => 50, S = female) = 0.001$

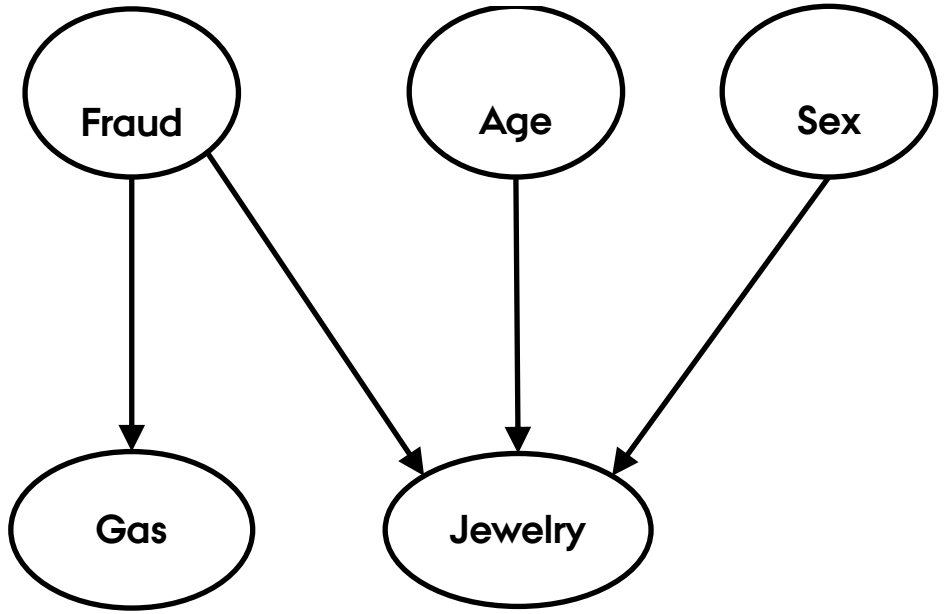
$p(\overline{J}|\overline{F}, A => 50, S = female) = 0.999$

INFERENCE

P(Fraud = yes| G= yes, A<=30, s= female) = ?

If this probability is sufficiently high, the analyst can take measures such as deny the current purchase or require additional identification.

$p\ (A =< 30) = 100\%$ $p\ (S = female) = 100\%$



$p\ (G = Yes) = 100\%$

Q&A

AARHUS
BSS

BN STRUCTURE

Algorithms for classification

1. Naïve Bayes
2. Tree Augmented Naïve Bayes (TAN)

Algorithms focused on causality discovery

3. Constrained-based algorithms

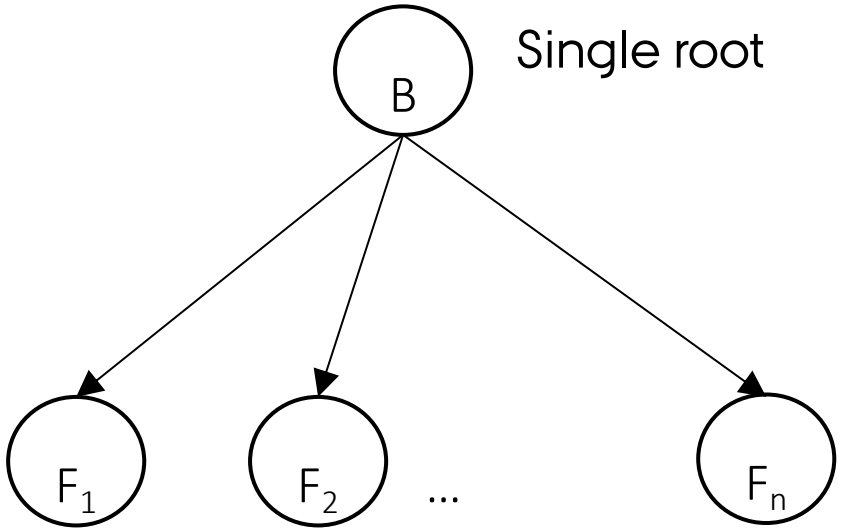
Algorithms focused on prediction

4. Score-based algorithms

AARHUS BSS

1. NAÏVE BAYES

Used to predict class membership probabilities based on the common aspects of each subject data (similar to logistic regression and discriminant analysis)

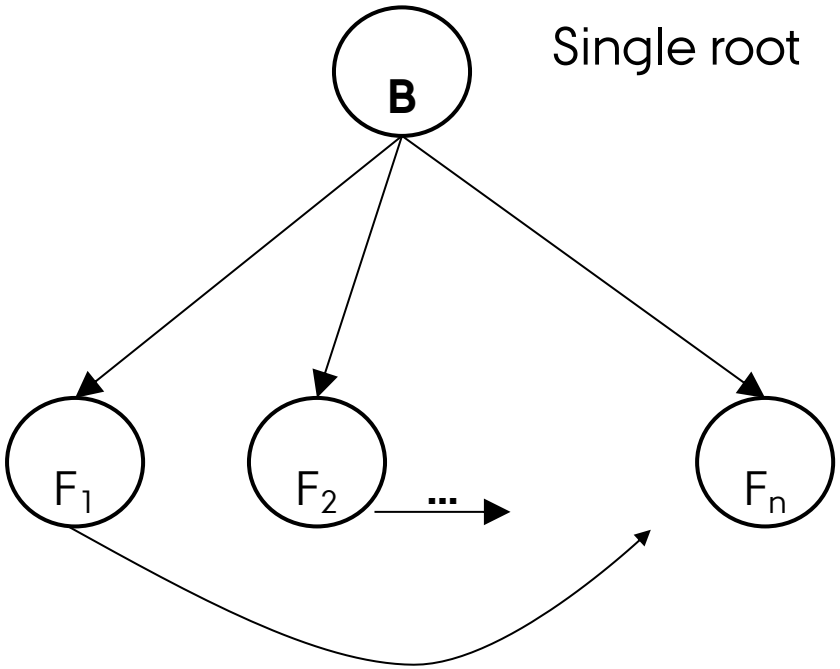


Cf. Markov property:

Any F_i *and* F_j are conditionally independent given B, i ≠ j

This implies: $P(F_1 \ldots F_n | B) = \prod_{i=1..n} P(F_i | B)$

2. TREE AUGMENTED NAÏVE BAYES (TAN)



Sames as Naive Bayes but allows relationships between features

3. CONSTRAINED-BASED ALGORITHMS

To identify the causal structures in the observational data, when causes are present in the data

- Data => Conditional Independence tests => DAG faithful to CI tests
- Faithfulness = all and only the conditional independencies found in P are entailed by DAG
- Several algorithms:
    - PC (Spirtes et al., 2001): the first practical application of the inductive causation (IC) algorithm by Verma and Pearl (1991)
    - Grow-Shrink (gs) (Margaritis, 2003)
    - Incremental Association (iamb) (Tsamardinos et al., 2003)
    - Fast Incremental Association (fast.iamb) (Yaramakala and Margaritis, 2005)
    - Interleaved Incremental Association (inter.iamb) (Tsamardinos et al., 2003)

CAUSALITY INFERENCE

- Reliance on bivariate associations may be very misleading

- It is necessary to take a multivariate approach by including all relevant variables in the analysis so as to study both marginal and conditional associations.

- This is the basis of graphical modelling and BN.

INDEPENDENCE

$X, Y - independent\ events\ if$

$$\begin{cases} P(X \cap Y) = P(X) \cdot P(Y) \ \ or \\ P(Y|X) = P(Y) \end{cases}$$

$X \in \{x_1, x_2\}\ and\ Y \in \{y_1, y_2\}$

$independent\ r.v.\ if$

$$\begin{cases} P(X \cap Y) = P(X) \cdot P(Y) \ \ or \\ P(Y|X) = P(Y) \end{cases}$$

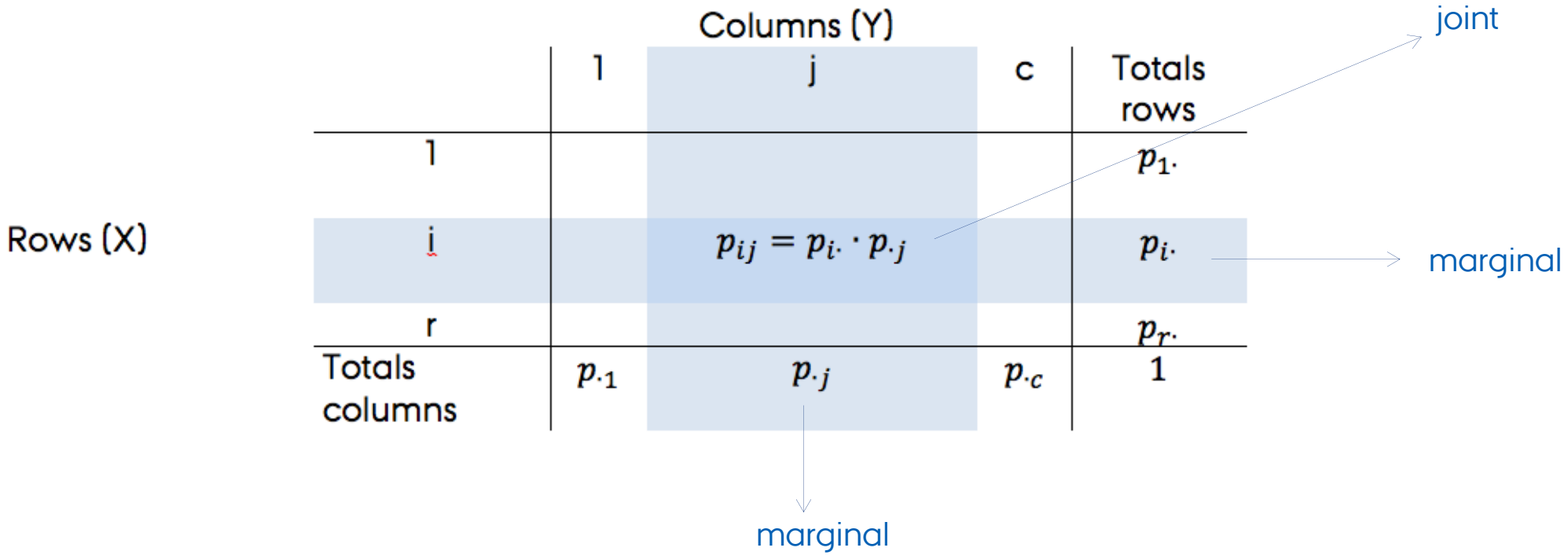
$for\ all\ values\ of\ X\ and\ Y$

$If\ \ X, Y\ \ are\ indep.\ continuous\ r.v.$

$$\begin{cases} f(x,y) = f(x) \cdot f(y) \ \ or \\ f(y \mid x) = f(y) \end{cases}$$

INDEPENDENCE

If the two categorical variables are **independent**, in each cell of the contingency table, the joint probability is equal to the product of the corresponding marginal probabilities



INDEPENDENCE

Often, it is preferred the characterization of independent variables that does not involve the density of X. That is:

$$P(Y|X) = P(Y)$$

$$f(y \mid x) = f(y)$$

Literally meaning that the marginal probability of Y does not change as a function of X.

CONDITIONAL INDEPENDENCE

X and Y are **conditionally independent** given Z, if for each value of Z, X and Y are independent.

$$X \perp Y | Z$$

EXAMPLE

As illustration, consider some data from a study of health and social characteristics of Danish 70-year-olds. Representative samples were taken in 1967 and again— on a new cohort of 70-year-olds— in 1984 (Schultz-Larsen et al., 1992). Body mass index (BMI) is a simple measure of obesity, defined as weight/height^2. It is of interest to compare the distribution between males and females, and between the two years of sampling.

Three variables:

- BMI - continuous
- Gender - discrete (Males, Females)
- Year – discrete (1964, 1967)

BMI DISTRIBUTION

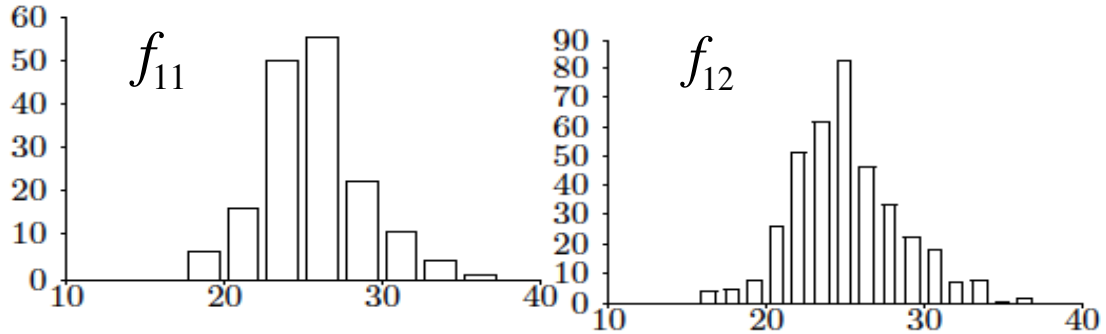


Figure 2.1: Males, 1967 sample.

Figure 2.2: Males, 1984 sample.

$$f_{B|G,Y}(b|G=i,Y=j)=f_{ij}$$

$i: 1 = male,\ 2 = female$

$j: 1 = 1967,\ 2 = 1984$

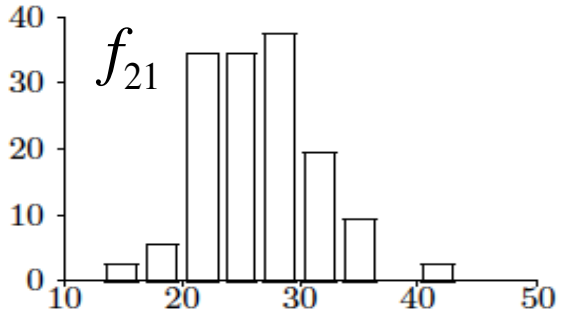


Figure 2.3: Females, 1967 sample.

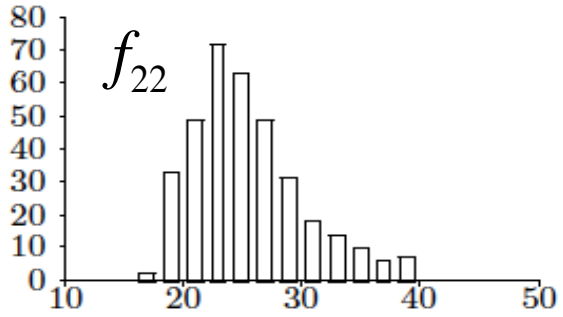


Figure 2.4: Females, 1984 sample.

1. BMI distribution does not change **neither by gender nor by year** => var are independent

$$f_{11} = f_{21} = f_{12} = f_{22}$$

$$BMI \perp (Gender, Year)$$

2. If BMI distribution changes by gender but **does not change by year** given we control for gender

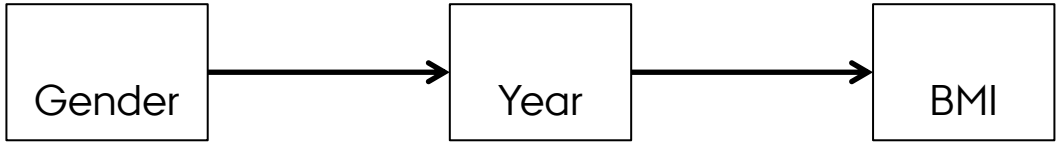
$$f_{11} = f_{12} \quad and \quad f_{21} = f_{22}$$

$$BMI \perp Year | Gender$$

3. If BMI distribution changes by year and **does not change by gender** given we control for year:

$$f_{11} = f_{21} \quad and \quad f_{12} = f_{22}$$

$$BMI \perp Gender \mid Year$$



CHALLENGES OF CONSTRAINED-BASED

- These algorithms require sufficient data to learn conditional independencies with certainty
- The errors in conditional independence test can introduce biases (e.g. which is the best level of significance 1%, 5%, 10% ?)

AARHUS
BSS

4. SCORE-BASED ALGORITHMS

General heuristic optimization techniques useful for prediction

- All possible DAGs are given the same probability of occurrence
- Given the data, the DAG with the highest network score is chosen
- **Several scores:**
  - Bayesian Information Criterion (BIC)
  - Akaike Information Criterion (AIC)
  - Bayesian Dirichlet equivalent (BDe)
- **Several algorithms:**
  - Greedy search algorithms – Hill Climbing (hc) (Bouckaert, 1995)
  - Genetic algorithms (Larranaga et al., 1997)
  - Simulated annealing (Bouckaert, 1995)
  - A review of score algorithms can be found in Russel & Norvig (2009)

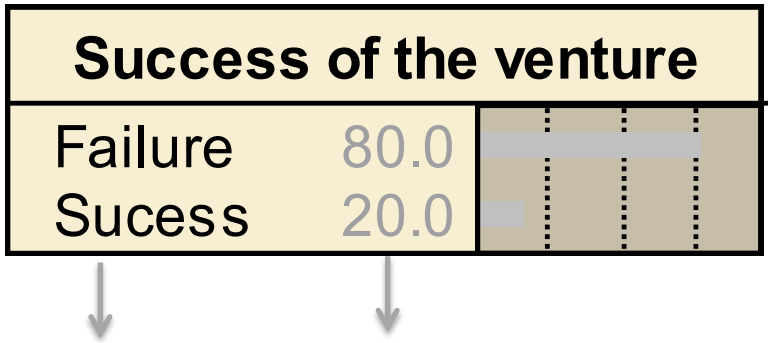
AARHUS BSS

Q&A

AARHUS
BSS

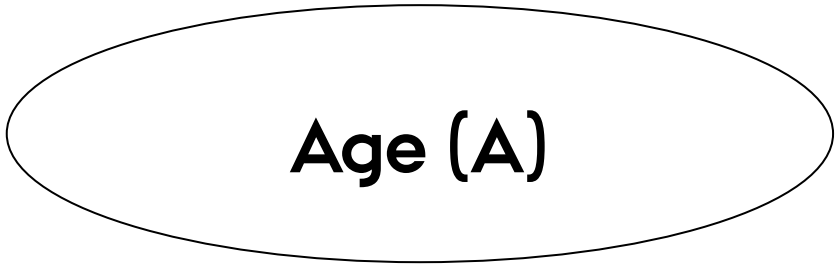
DAG ELEMENTS

Discrete variable (in NETICA)



States *Marginal probabilities*

Continuous variable

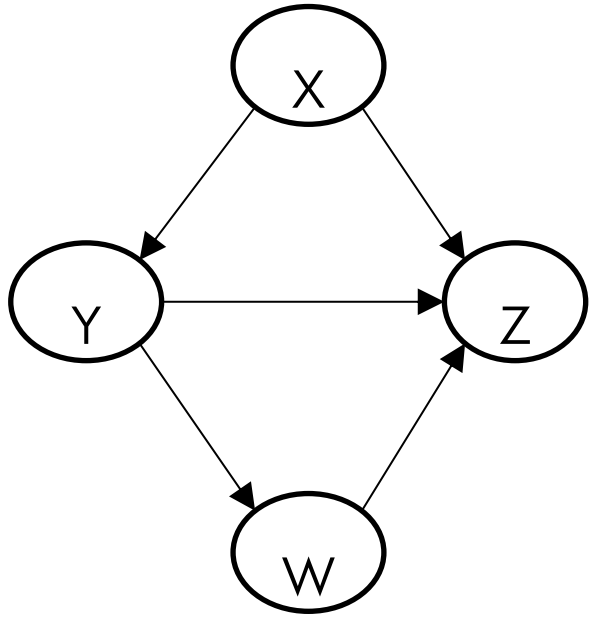


$$A \approx N(50,10^2)$$

DAG ELEMENTS

Set of nodes : $V = \{X, Y, Z, W\}$

Directed edges or arcs: $E = \begin{Bmatrix} (X,Y); (X,Z); (Y,Z); \\ (Y,W); (W,Z) \end{Bmatrix}$



A path: $[X, Y, W, Z]$

A chain: $[Y, W, Z, X]$

A cycle: a path from a node to itself. BNs are acyclic.

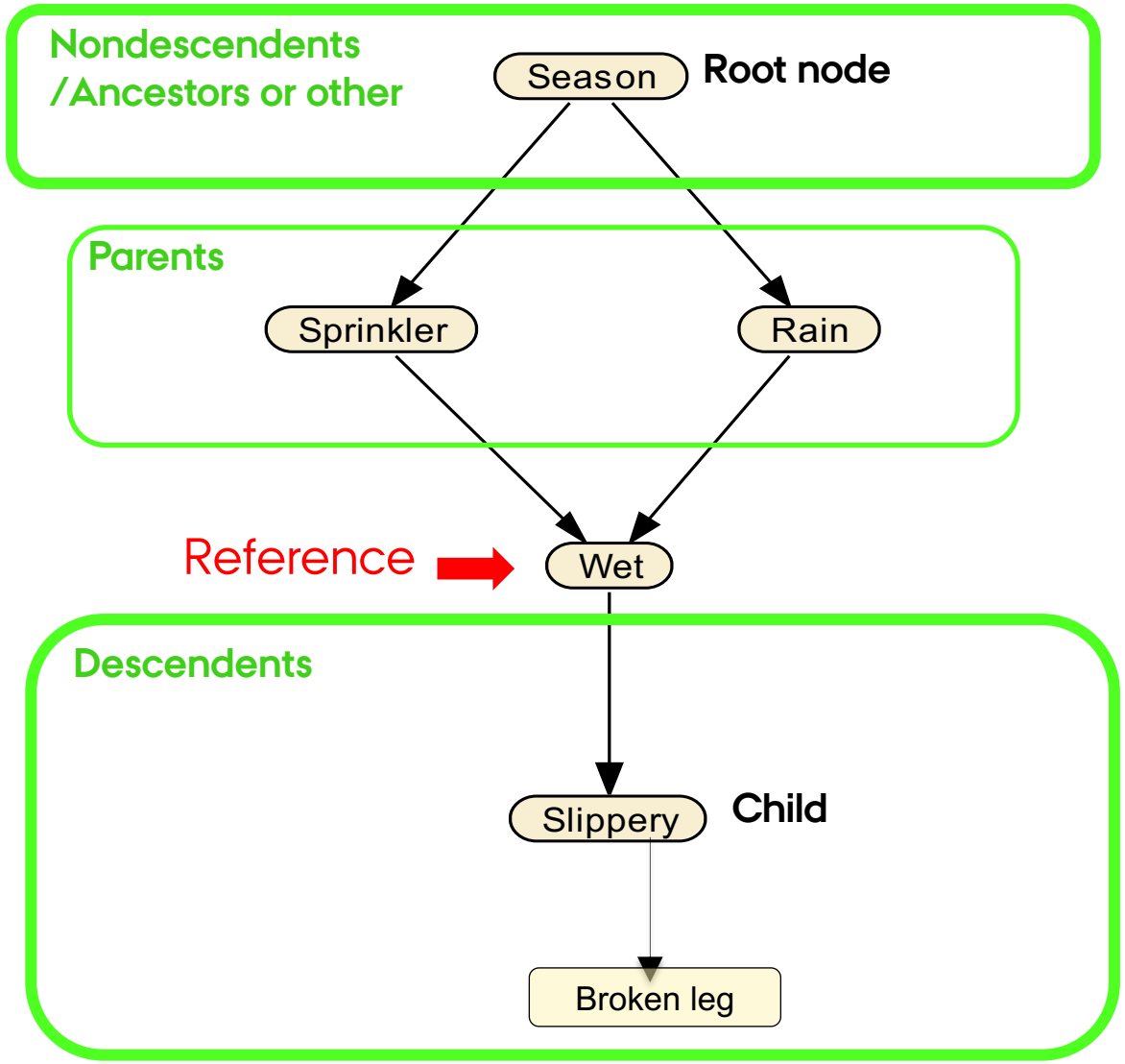
Complete (saturated) graph : if there is an arc between every pair of nodes

Connected graph: if there is a path between every pair of nodes

Empty graph: a graph with no arcs

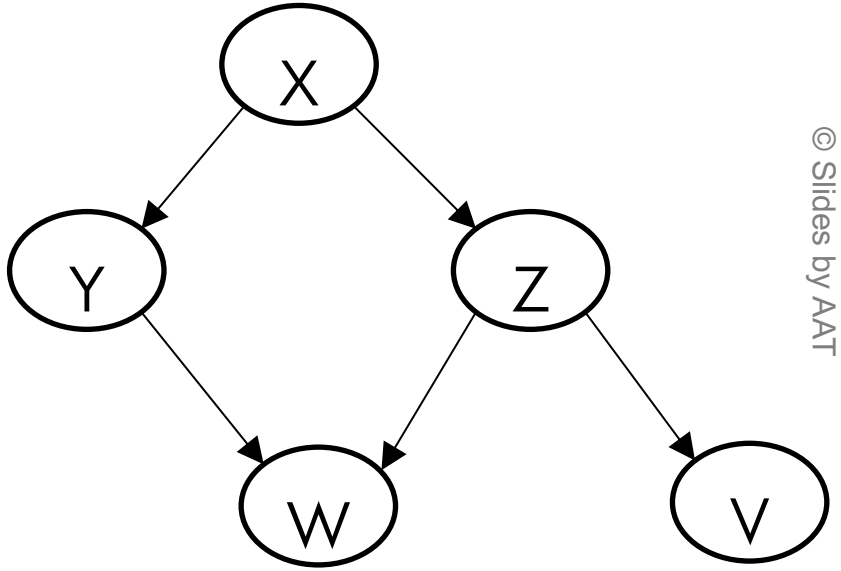
DAG ELEMENTS

Neighbourhood:
(Parents & Children)



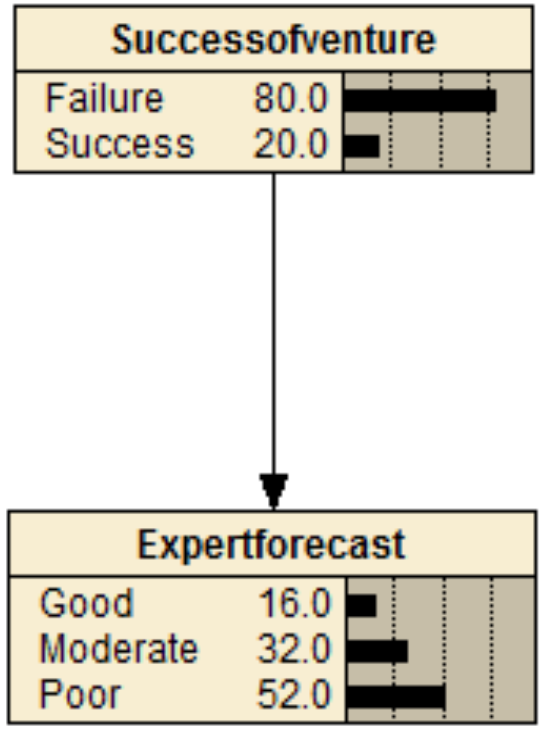
EXAMPLE

Node	Parents	Nondescendents
X	∅	∅
Y	X	Z , V
Z	X	Y
W	Y , Z	X , V
V	Z	X, Y, W

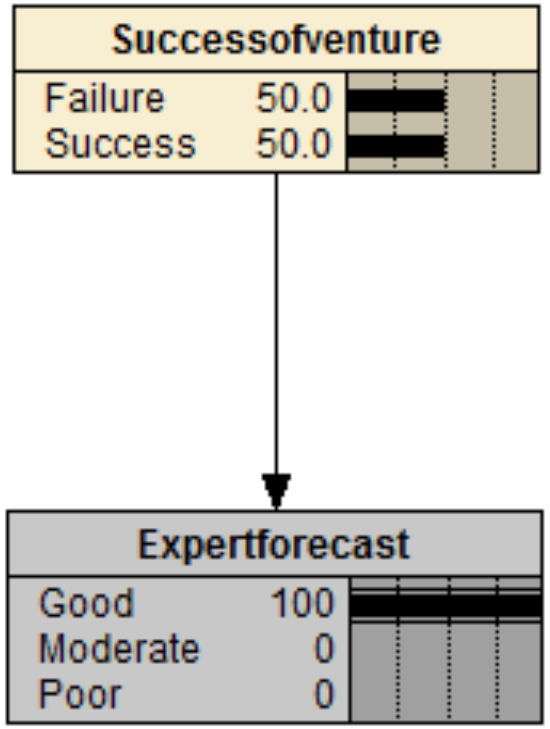


SETTING EVIDENCE (INFERENCE)

No evidence in the net



Expert forecast node has hard evidence set



Q&A

AARHUS
BSS

HOW THE INFORMATION FLOWS?

- 3 principles or type of connections

- Following these principles, it is possible to decide for any pair of variables whether they are independent given the evidence entered in the network

SERIAL CONNECTION

In Figure 1a, if we get evidence for C, that changes the probability of B, which in turn changes the probability of A

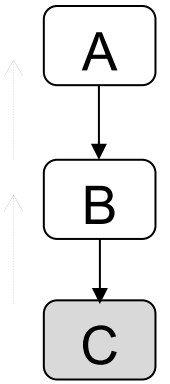


Figure 1a

Yet, if we set evidence for B (Figure 1b), then any change to C have no effect on A. Nor can any changes to A affect C. We say that the certainty of B blocks any dependence formerly shared between A and C. To be precise, there may be dependencies introduced by other relationships in the net, but not via B.

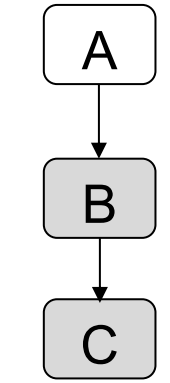


Figure 1b

DISCOUNTING (EXPLAINING AWAY)

Figure 1a shows A and B are independent, but they are competing explanations of C.

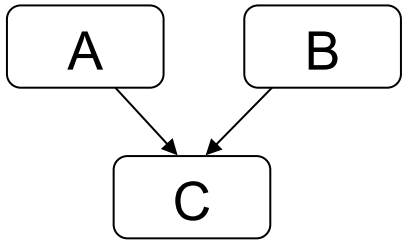


Figure 1a

In Figure 1b, if we set evidence for C, any later change in the probability of A, has an opposite change in B (and vice versa).

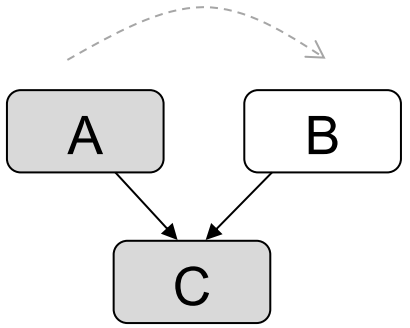


Figure 1b

DIVERGENT (COMMON CAUSE)

In Figure 1a, if we get evidence for A, this increases the chances of B which in turn increases the chances of C

Yet, in Figure 1b, if we set evidence for B, then any changes to A have no effect on C (and by symmetry, no changes in C can affect A). Again, to be precise, there may be dependencies introduced by other relationships in the net, but not via B.

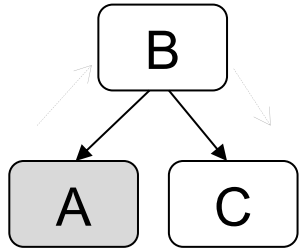


Figure 1a

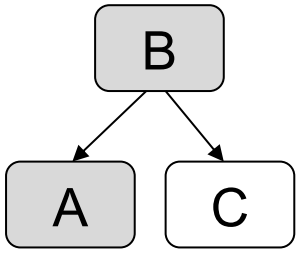


Figure 1b

D – SEPARATION CONCEPT

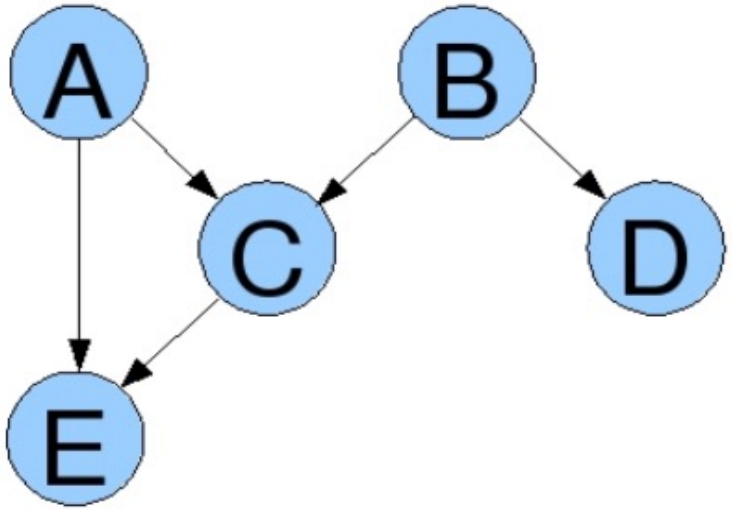
d-separation is a criterion for deciding whether a set X of variables is independent of another set Y, given one controls for a third set Z, using the three principles.

EXAMPLE 1

Which pairs of variables are independent in the graphical model below, given that none of them have been observed (no evidence)?

Answer:

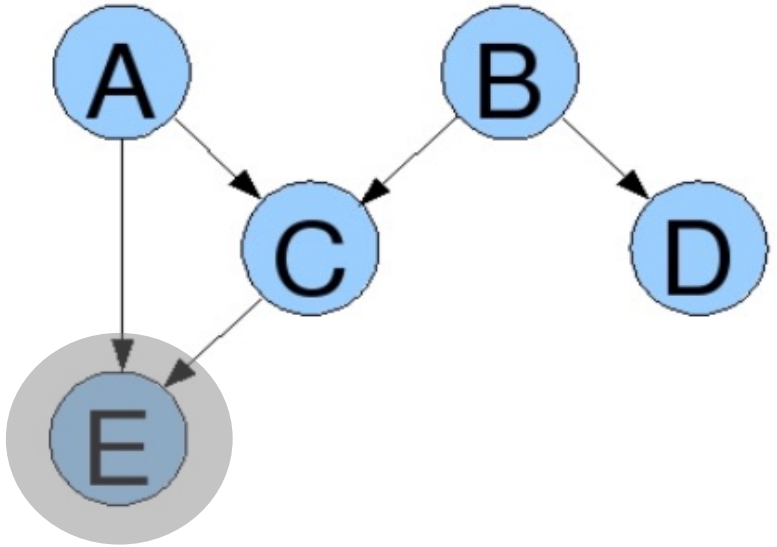
If no evidence set, cf. principle 1,
A, B are independent as there are no active trails between them.



EXAMPLE 2

Now assume that the value of **E is known** (we set evidence for E). Which pairs of variables (not including E) are independent in the same graphical model, given E?

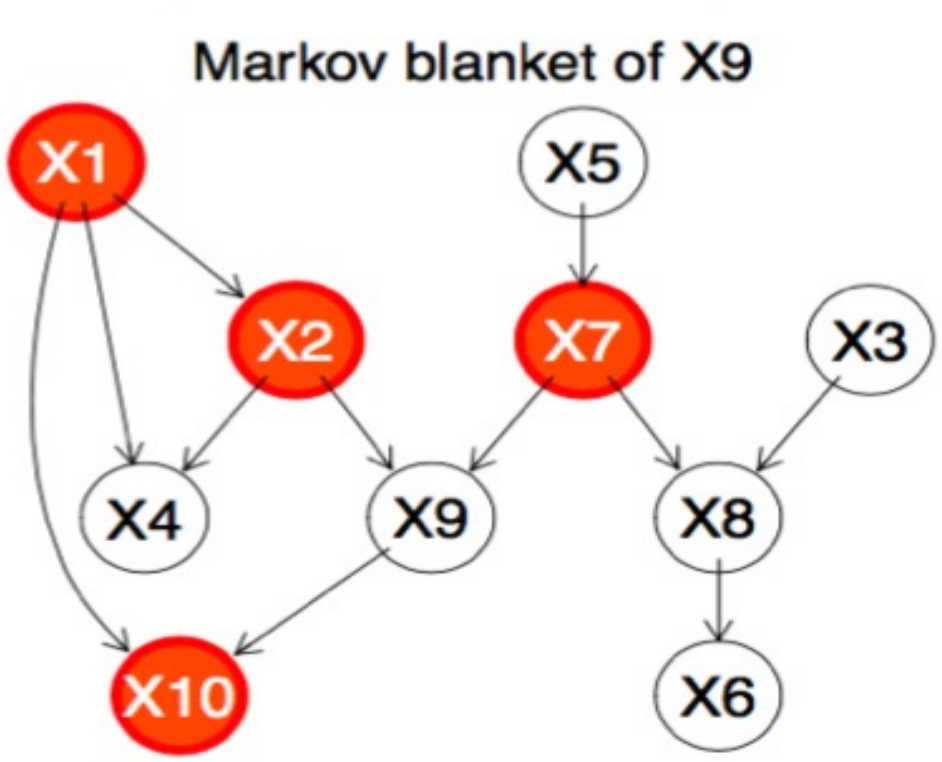
Answer: There are no pairs of variables that are independent. Observing E, activates the V-structures around C and E, giving rise to active trails between every pair of variables in the network.



MARKOV BLANKET

The Markov Blanket of a node includes:

- the parents of the node,
- the children
- other parents of those children



AARHUS BSS

MARKOV BLANKET

- The Markov Blanket of the node X_i contains all the nodes that, if we know their states (i.e. we have evidence for these nodes), it will isolate the node X_i from the rest of the network (i.e. it will make X_i independent of all the other nodes given its Markov Blanket).

- It is the set of nodes that includes all the knowledge needed to do inference on the current node.

- Learning a Markov Blanket is particularly helpful when there is a large number of variables to select from in a dataset. It can also serve as a highly-efficient variable selection method in preparation for other types of modeling e.g. Regression, Neural Nets, etc.

MARKOV PROPERTY

For each variable X in the graph, X is conditionally independent of the set of all its nondescendents given the set of all its parents.

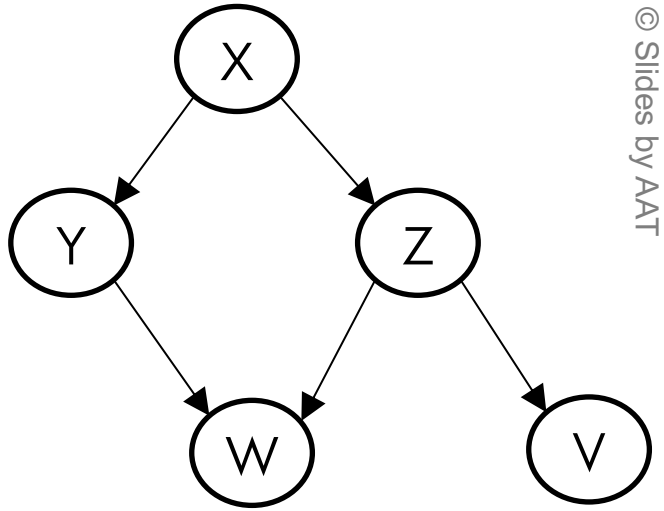
$$I_P(X, ND \mid PA)$$

$ND = \text{set of all nondescendents}$

$PA = \text{set of all parents}$

EXAMPLE

Node	Parents	Nondescendents	Conditional Independency
X	$\varnothing$	$\varnothing$	None
Y	X	Z , V	$I_P(Y,\{Z,V\}\|X)$
Z	X	Y	$I_P(Z,Y\|X)$
W	Y , Z	X , V	$I_P(W,\{X,V\}\|\{Y,Z\})$
V	Z	X, Y, W	$I_P(V,\{X,Y,W\}\|Z)$



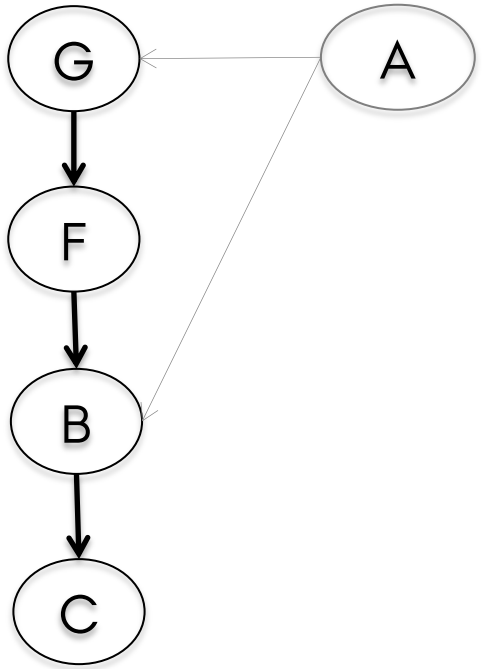
NOTE

- Technically, the Markov condition is guaranteed by learning the DAG from the data. The algorithms will construct a directed graph by catching the conditional independencies and drawing edges between variables pairs that are not conditionally independent.

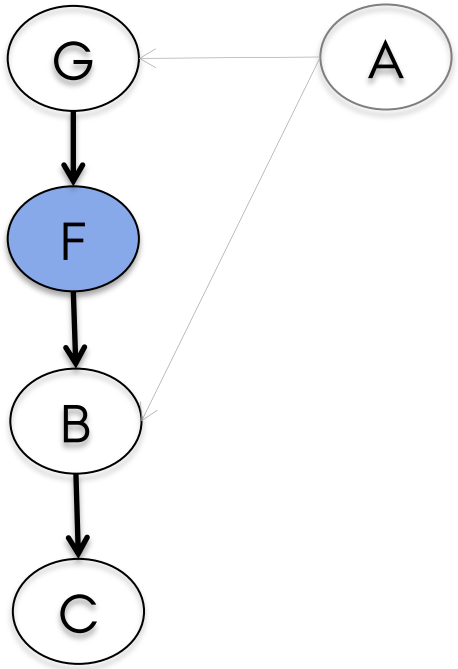
- However, if we had reason to believe that there is a hidden cause and this was not recorded in the dataset, and consequently not included in the graph, the Markov property will not hold.

- In that case, we cannot claim it is a causal BN network, although we can still use it as a predictor tool. See examples next.

EXAMPLES



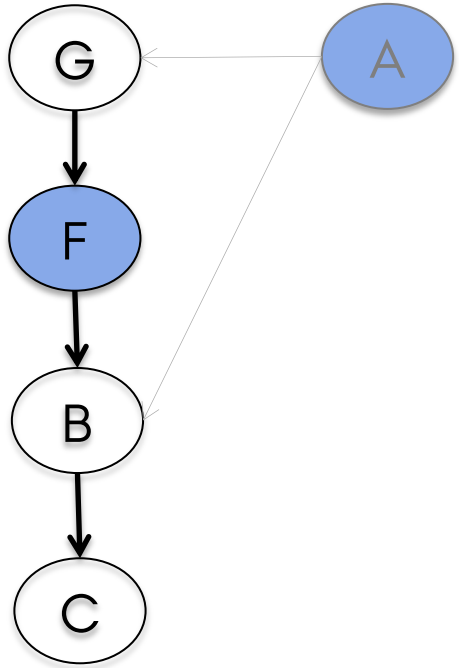
A is hidden/
unknown



A is hidden/
unknown

$I_P(C, G \mid F)$?

Re : *no*



A is hidden but controlled
in some way e.g. by randomization

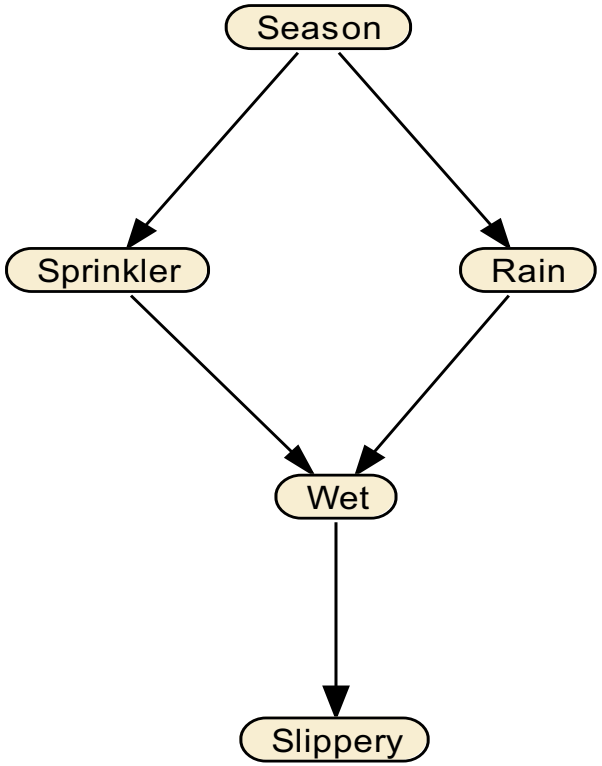
$I_P(C, G \mid \{F, A\})$?

Re : *yes*

CONCLUSION

In a causal BN, the Markov property is satisfied given that **all common causes are represented** in the graph (i.e. there are no hidden common causes acting as confounding factors).

In our Season example, we can say there is no way for Rain to influence Slippery except by way of causing Wet or not. Thus we can assume there is no hidden variable connecting Rain and Slippery. Every independency suggested by the lack of an arrow between Rain and Slippery is actually real in the system.



THEOREM 3.1

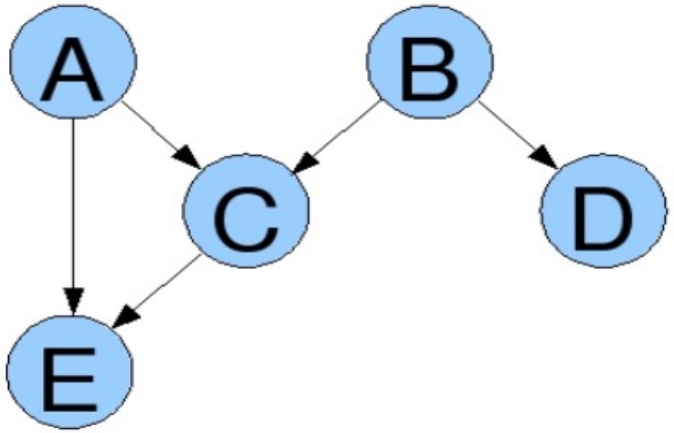
Owing to Markov property (and Bayes theorem), the representation of the **joint probability** of a set of random variables (global distribution) can be expressed as a **product of the conditional distributions** of each variable given its parents in G (graph), whenever these conditional distributions exist.

$$P(\mathrm{X}) = \prod_{i=1}^{p} P\left(X_i \middle| \Pi_{X_i}\right) \quad , \quad \text{where } X = \{X_i\},\ i = 1:n$$

$$\Pi_{X_i} = \{\text{parents of } X_i\}$$

EXAMPLE

Given model attached, which of the alternative (a), b), or c) is an appropriate decomposition of the joint probability distribution P (A, B, C, D, E)?



$a)$ $P(A,B,C,D,E) = P(A) \cdot P(B) \cdot P(C) \cdot P(D) \cdot P(E)$

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

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

SUMMARY

- Internally, BN are represented as a DAG and a set of marginal and conditional probabilities
- We prefer using the conditional probabilities to joint probabilities because they are significantly less numerous
- Joint probabilities values increase exponentially with the number of variables
- Conditional probabilities values increase linearly with the number of variables
- Theorem 3.1: the joint distribution values can be easily computed from the conditional probabilities.
- Representing the joint probability distribution of the variables in the net, as a product of local probability distributions (either marginal, for root nodes, or conditional for nodes with parents), according to the arcs present in the graph makes BN **a very efficient way** to define the joint probability of multiple variables.

MAIN REFERENCE

Neapolitan, R. E., & Jiang, X. (2007). Bayesian Networks. In *Probabilistic methods for financial and marketing informatics* (1st edition ed.). San Fransisco, CA: Morgan Kaufmann Publishers.

AARHUS
BSS