Bayesian Networks Learning From Data

Marco F. Ramoni
Children's Hospital Informatics Program
Harvard Medical School

HST951 (2003)

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

Introduction

- Bayesian networks were originally developed as a knowledge representation formalism, with human experts their only source.
- Their two main features are:
  - ✓ The ability to represent deep knowledge (knowledge as it is available in textbooks), improving portability, reusability, and modularity.
  - ✓ They are grounded in statistics and graph theory.
- Late ’80s, people realize that the statistical foundations of Bayesian networks makes it possible to learn them from data rather than from experts.

Outline

- Learning from data.
- Learning Bayesian networks.
- Learning probability distributions.
- Learning network structures.
  - ✓ The classical way.
  - ✓ The Bayesian way.
- Searching the space of possible models.
- A couple of examples.
- Lurking variables, hidden variables, and causality.

Components

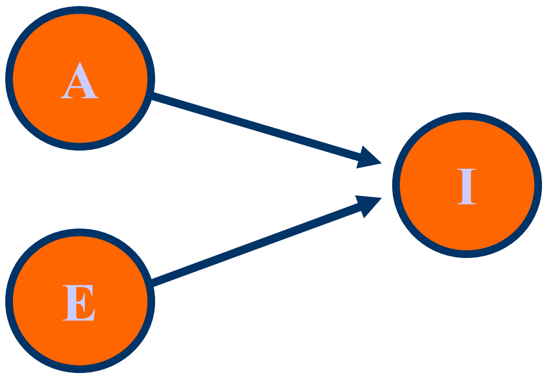
Qualitative: A dependency graph made by:

- Node: a variable X, with a set of states $\{x_1, \ldots, x_n\}$.
- Arc: a dependency of a variable X on its parents Π.

Quantitative: The distributions of a variable X given each combination of states π_i of its parents Π.

A	p(A)
Y	0.3
O	0.7

E	p(E)
L	0.8
H	0.2



A	E	I	p(I|A,E)
Y	L	L	0.9
Y	L	H	0.1
Y	H	L	0.5
Y	H	H	0.5
O	L	L	0.7
O	L	H	0.3
O	H	L	0.2
O	H	H	0.8

A=Age; E=Education; I=Income

The Age of the Experts

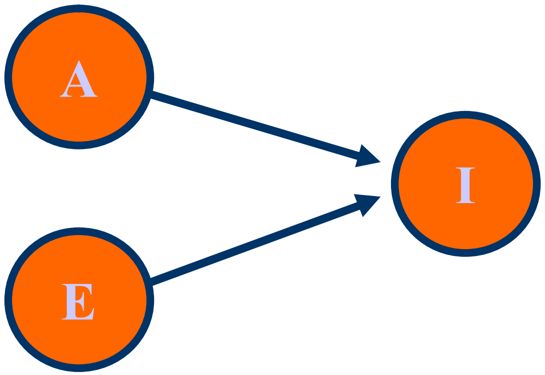
- The traditional source of knowledge is a human expert.
- The traditional trick is to ask for a “causal graph” and then squeeze the numbers out of her/him.
- The acquisition is easier than the traditional one but still… it can be painful.

Learning Bayesian Networks

- Learning a Bayesian network means to learn.
  - ✓ The conditional probability distributions,
  - ✓ The graphical model of dependencies.

A	p(A)
Y	0.3
O	0.7

E	p(E)
L	0.8
H	0.2



A	E	I	p(I|A,E)
Y	L	L	0.9
Y	L	H	0.1
Y	H	L	0.5
Y	H	H	0.5
O	L	L	0.7
O	L	H	0.3
O	H	L	0.2
O	H	H	0.8

Learning Probabilities

- Learning of probability distributions means to update a prior belief on the basis of the evidence.
- Probabilities can be seen as relative frequencies:

$$p(x_j \mid p_i) = \frac{n(x_j \mid p_i)}{\sum_j n(x_j \mid p_i)}$$

- Bayesian estimate includes prior probability:

$$p(x_j \mid p_i) = \frac{a_{ij} + n(x_j \mid p_i)}{\sum_j a_{ij} + n(x_j \mid p_i)}$$

α_{ij} / α_i represents our prior as relative frequencies.

Learn the Structure

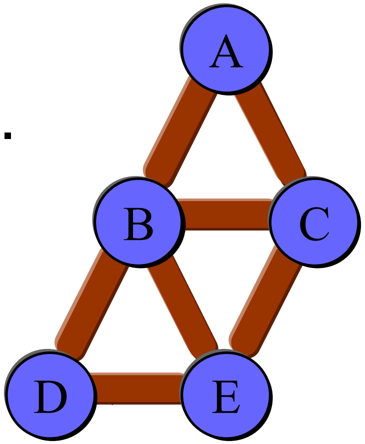
- In principle, the process of learning a Bayesian network structure involves:
  - Search strategy to explore the possible structures;
  - Scoring metric to select a structure.
- In practice, it also requires some smart heuristic to avoid the combinatorial explosion of all models:
  - ✓ Decomposability of the graph;
  - ✓ Finite horizon heuristic search strategies;
  - ✓ Methods to limit the risk of ending in local maxima.

Model Selection

- There are two main approaches to select a model:
  - Constraint-based: use conditional independence test to check assumptions of independence and then encode the assumptions in a Bayesian network.
  - Bayesian: all models are a stochastic variable, the network with maximum posterior probability.
- Bayesian approach is more popular:
  - Probability: it provides the probability of a model.
  - Model averaging: predictions can use all models and weight them with their probabilities.

Constraint Based

- A network encodes conditional independence (CI).
- A DAG has an associated undirected graph which explicitly encodes these CI assumptions.
- Associated undirected graph: the undirected graph obtained by dropping the direction of links.
- Moral graph: the undirected graph obtained by.
  - ✓ Marring parents of a child.
  - ✓ Dropping the directions of the links.
- How to read this graph?



Learning CI Constraints

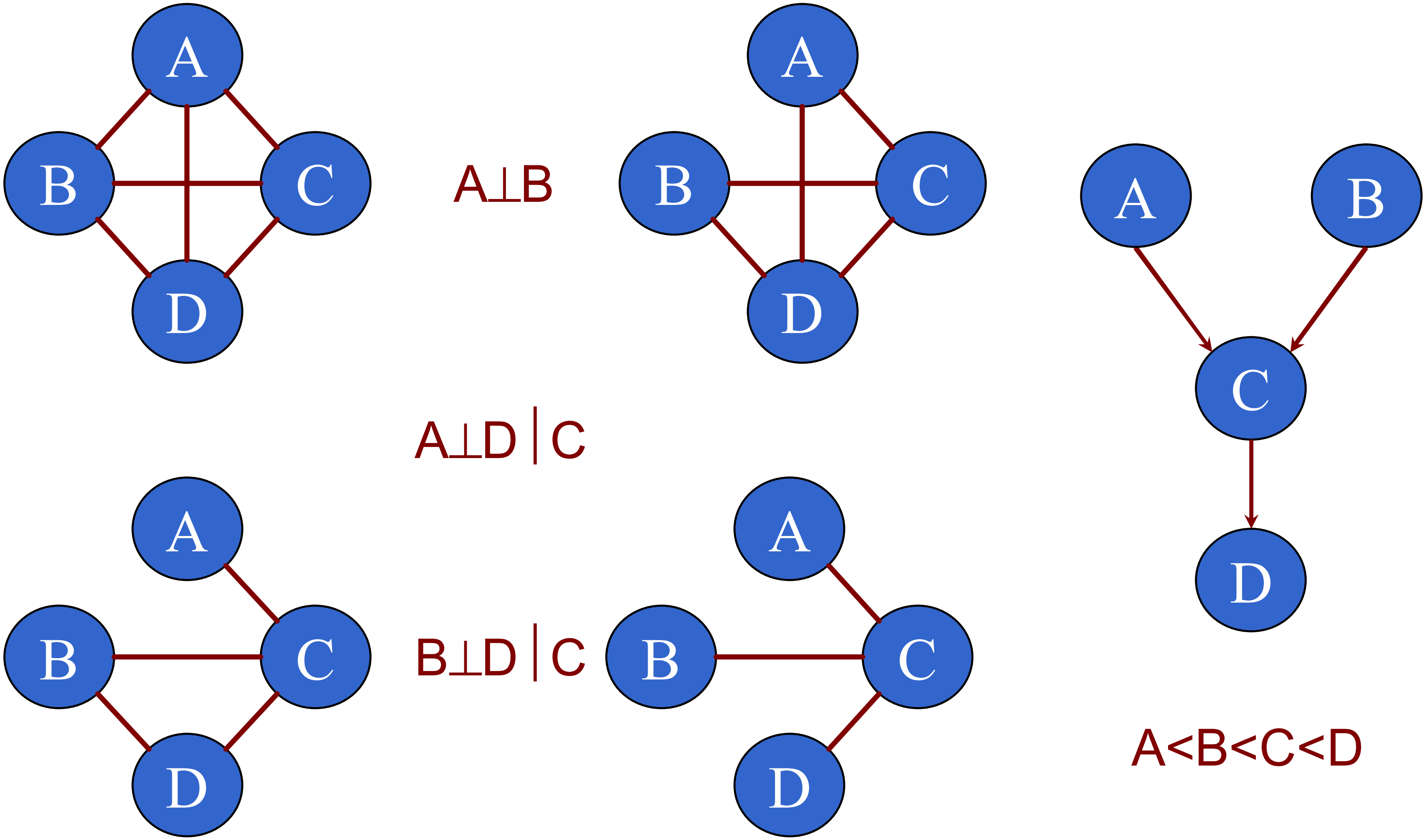
Search strategy: top-down.

1. Start with the saturated (undirected) graph.
2. Go link by link and test the independence.
3. If independence holds, remove the arc.
4. Swing the variables to assess the link direction.

Scoring metric: independence tests.

- ✓ Compute the expected frequencies under the assumption that the variables are independent.
- ✓ Test the hypothesis with some statistics (G^2).
- ✓ Assume no structural 0.

Example



A<B<C<D

Bayesian Model Selection

- The set of possible models *M* is a stochastic variable with a probability distribution *p(m)*.
- We want to select the model m_i with the highest posterior probability given the data *D*.
- We must search all models and find the one with highest posterior probability.
- We can use Bayes' theorem:

$$p(M \mid \mathrm{D}) = \frac{p(\mathrm{D}, M)}{p(\mathrm{D})} = \frac{p(\mathrm{D} \mid M)\, p(M)}{p(\mathrm{D})}$$

Scoring Metric

Result: we just need the posterior probability.

First note: all model use the same data:

$$p(m_i | D) \propto p(D | m_i) p(m_i).$$

Second note: models have equal prior probability:

$$p(m_i | D) \propto p(D | m_i).$$

Conclusion: as we need only a comparative measure, we need just the marginal likelihood.

Assumptions: this scoring metric works under certain assumptions (complete data, symmetric Dirichlet distributions as priors).

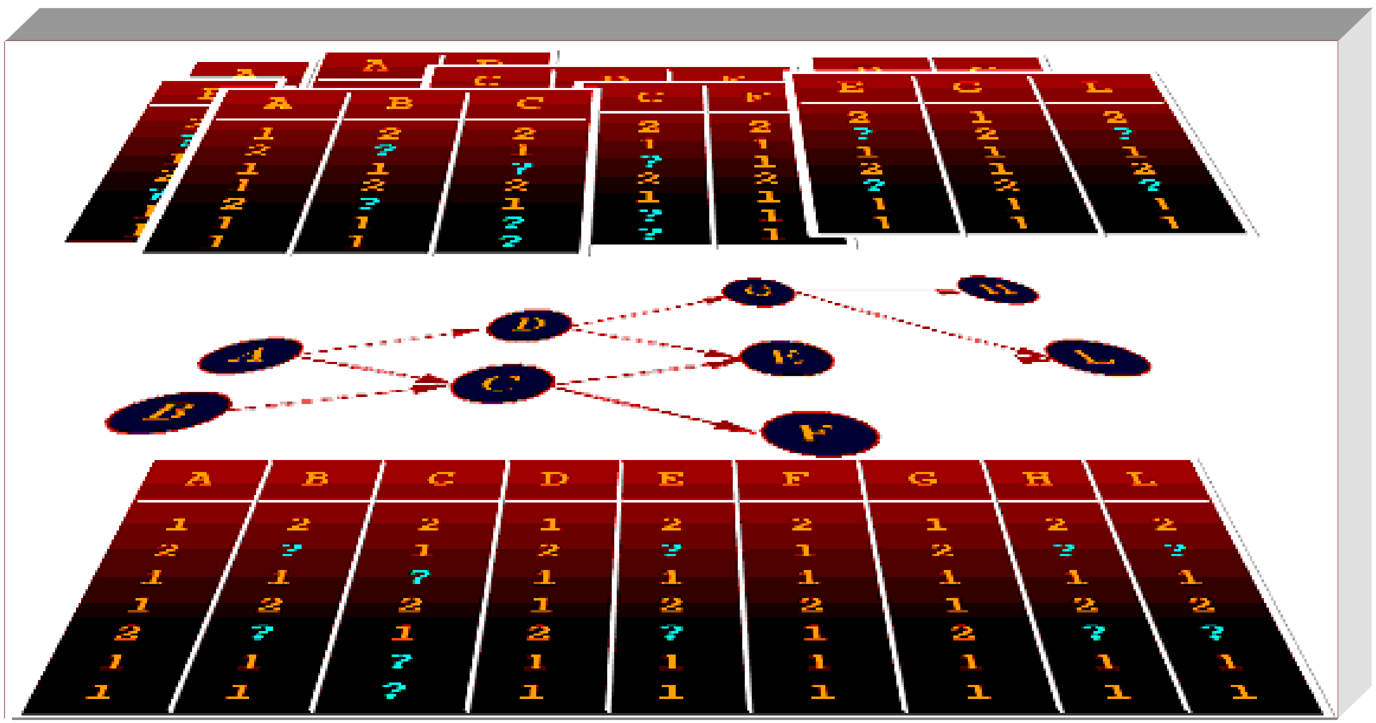
Bayes Factor

- The marginal likelihood (linear or log) is a measure proportional to the posterior probability.
- This is good enough to identify the best model but not to say how better is a model compared to another.
- This may be important to take into account criteria of parsimony or to assess confidence.
- Bayes factor computes how many times a model is more likely than another as the ratio of their marginal likelihood (or marginal log likelihood):

$$BF(m_i, m_j) = p(D \mid m_i) / p(D \mid m_j) \propto p(m_i \mid D) / p(m_j \mid D).$$

Factorization

- The graph factorize the likelihood: the “global” likelihood is the product of all local likelihood.

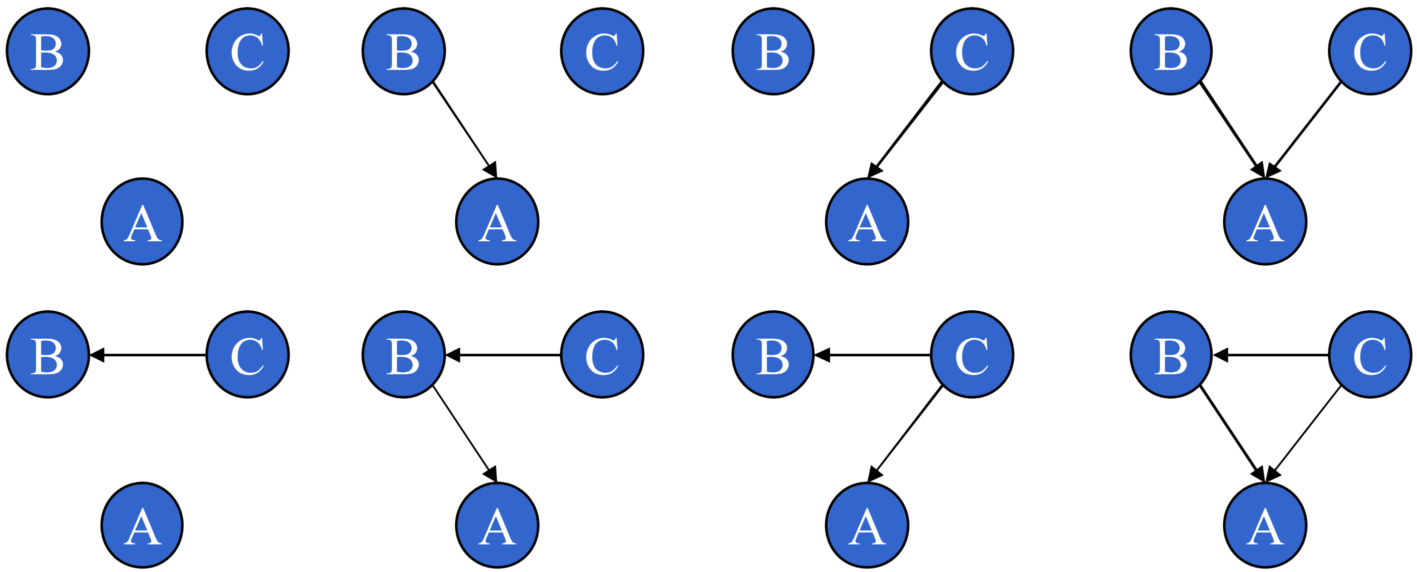


Search

Strategy: Bottom up.

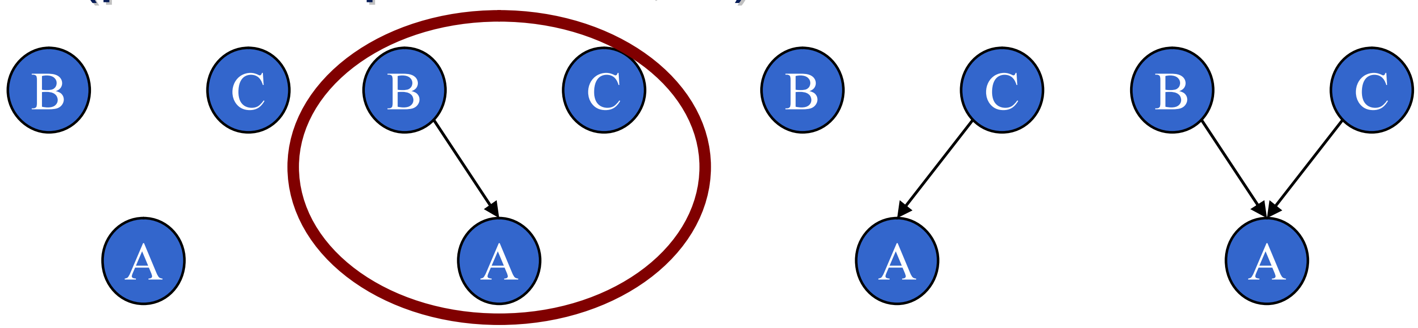
Variables: $X_i < X_j$, if X_i cannot be parent of X_i.

Example: A<B<C.

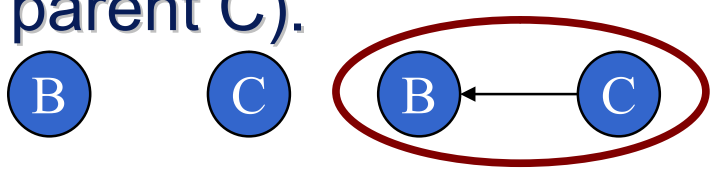


Local Model Selection

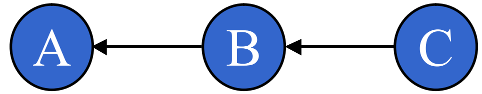
A (possible parents B; C):



B (possible parent C).



The model:



Survival Analysis

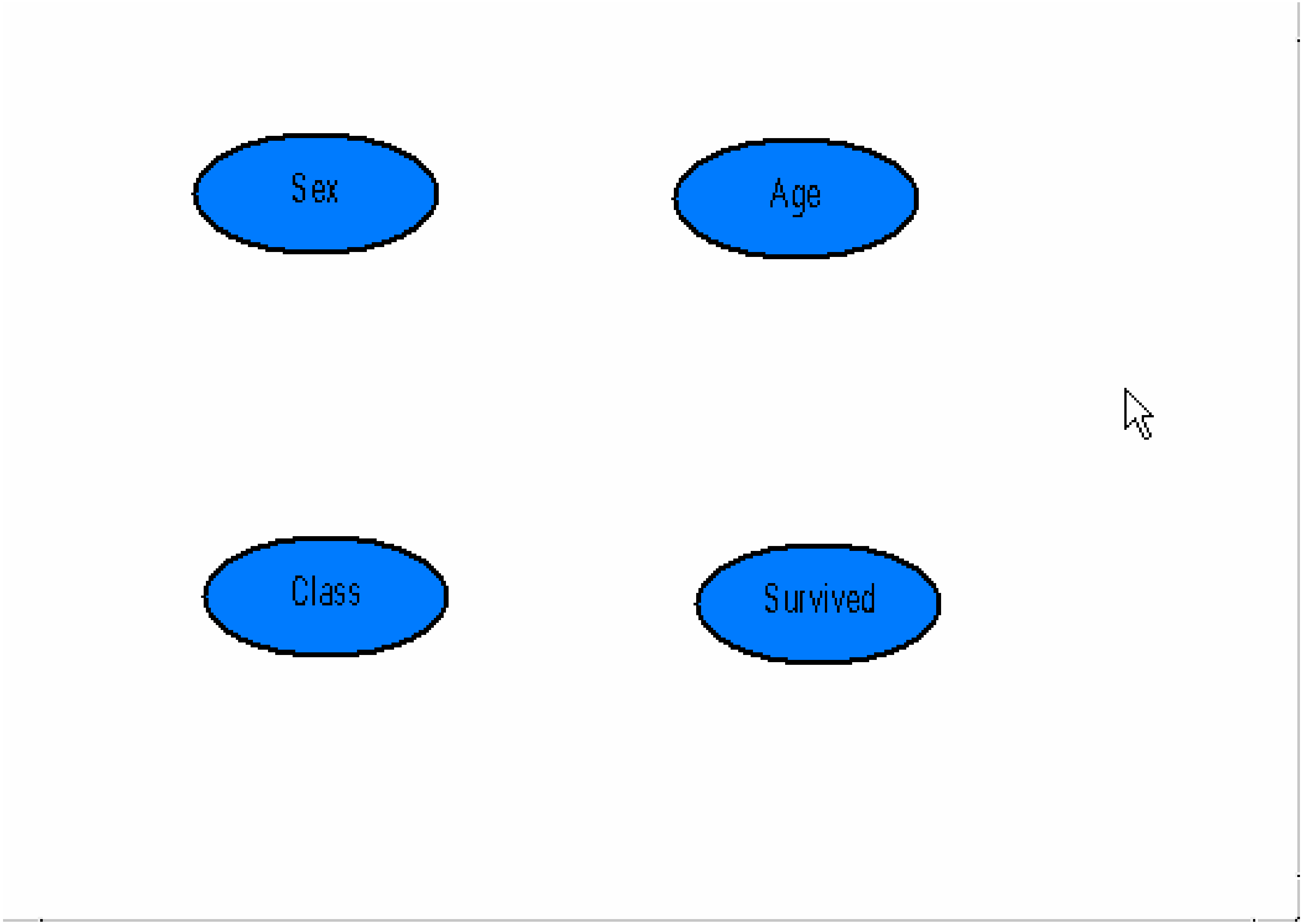
Topic: Survival analysis of the Titanic disaster.

Input: 2022 cases on four variables.

- ✓ Class: first, second, third, crew;
- ✓ Gender: male, female;
- ✓ Age: adult, child;
- ✓ Survived: yes, no.

Output: the model of interactions and its likelihood.

The Titanic



Example

Database: Breast Cancer Database (UCI Archive).

Source: University of Wisconsin, W. H. Wolberg.

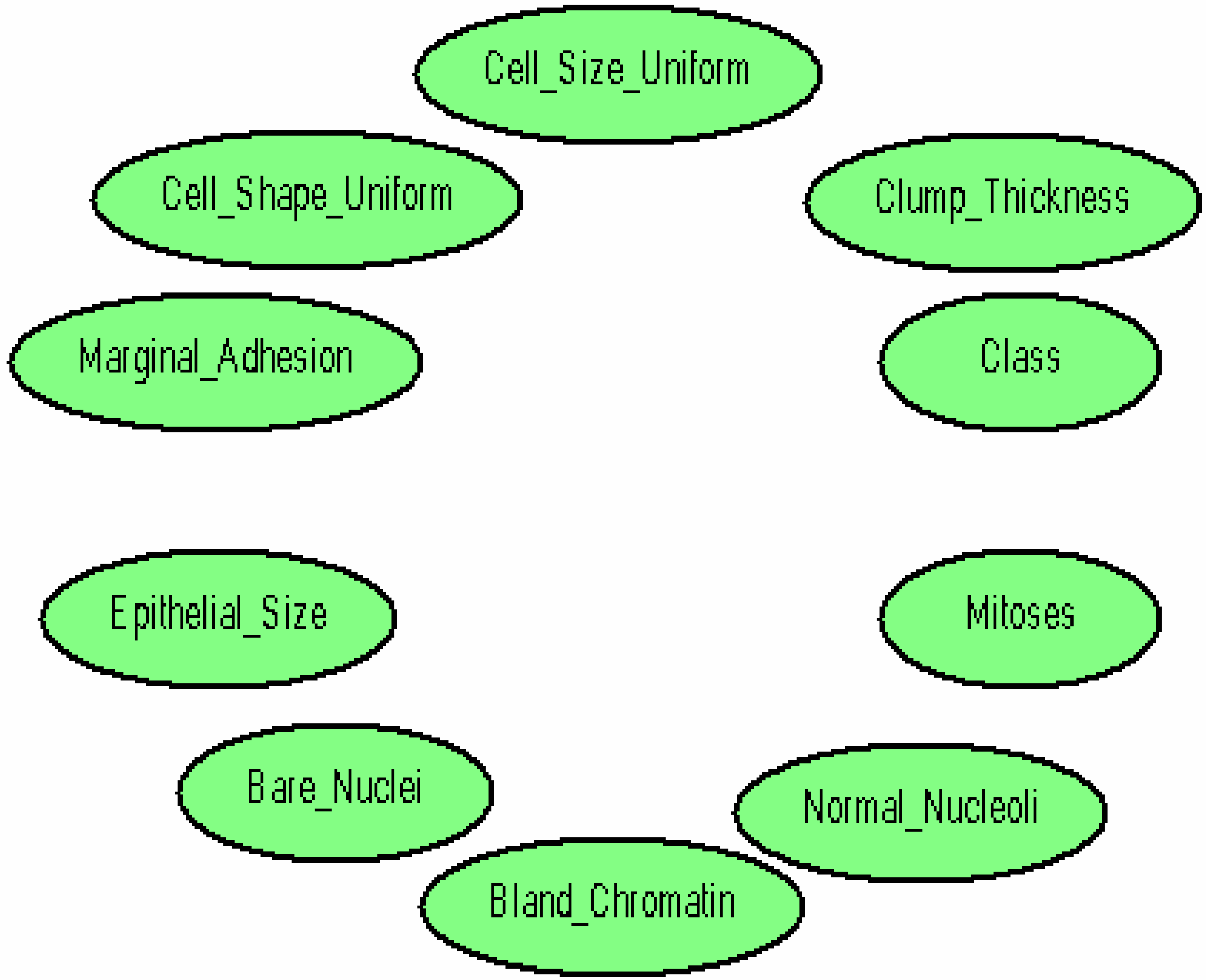
Topic: Breast cancer malignancy classification.

Cases: 699 cases.

Variables: 10 with 10 states + malignancy class:

1	Clump Thickness	6	Bare Nuclei
2	Uniformity of Cell Size	7	Bland Chromatin
3	Uniformity of Cell Shape	8	Bland Chromatin
4	Marginal Adhesion	9	Normal Nucleoli
5	Single Epithelial Cell Size	10	Mitoses

Breast Cancer



Causality

- What the arrows in a Bayesian network mean?
- The received definition of causal sufficiency (Suppes, 1970) states that a relation is causal if:
  - ✓ There is correlation between the variables;
  - ✓ There is temporal asymmetry (precedence);
  - ✓ There is no hidden variable explaining correlation.
- Hidden variables explain statisticians' reluctance to use the word causal.
- Yule (1899) on the poverty causes in England. Note: "Strictly speaking, for 'due to' read 'associated with'."

Richard III

- Naïve (Aristotelian) causality:

  For want of a nail the shoe was lost,
  For want of a shoe the horse was lost,
  For want of a horse the rider was lost,
  For want of a rider the battle was lost,
  For want of a battle the kingdom was lost,
  And all for want of a horseshoe nail.

- Modern causality among variables not events:

  Galilean equation: $d=t^2$.

- When we talk causality, we talk Causal Laws.

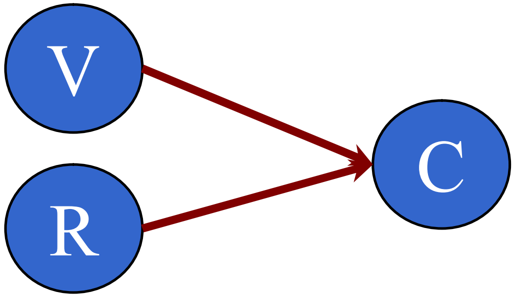
The Enemies

- The critical problem here is the Simpson's paradox: getting stuck in a local maximum.
- 674 murder defendants in Florida between 1976 and 1987. Are capital sentences racially fair?

	No Death	Death	Total
White	141 88.1%	19 11.9%	160 49.1%
Black	149 89.8%	17 10.2%	166 50.9%

Lurking Variable: The Victim

Victim	Defendant	Non Death	Death
White	White	132 87.4%	19 12.6%
White	Black	52 82.5%	11 17.5%
Black	White	9 100%	0 0%
Black	Black	97 94.2%	6 5.8%

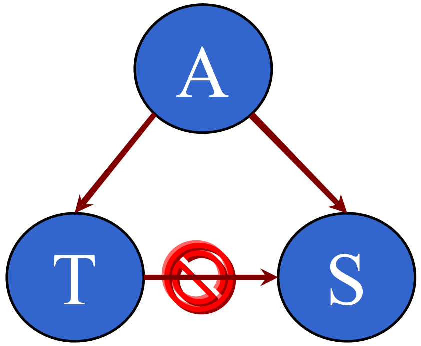


Hidden Variables

- Hidden variables can also prevent independence.
- Consider a database of children, reporting their T-shirt size and their running speed.

T-shirt	Fast	Slow
Small	0.32	0.68
Large	0.35	0.65

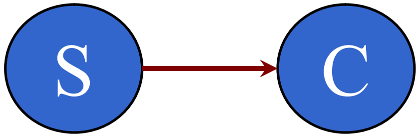
T-shirt	Age	Fast	Slow
Small	<5	0.3	0.7
Large	<5	0.3	0.7
Small	>5	0.4	0.6
Large	>5	0.4	0.6



Does Smoking Cause Cancer?

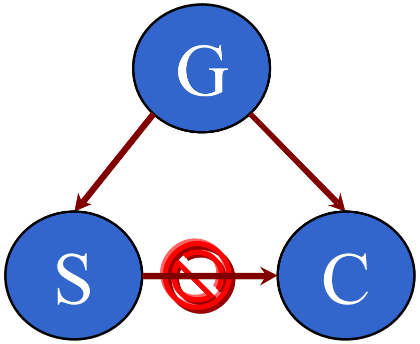
- In 1964, the Surgeon General issued a report linking cigarette smoking to lung cancer based on correlation between smoking and cancer in observational data.
- Based on these results, the report claimed causality: If we ban smoking, the rate of cancers will be the same as the one in the non-smoking population.

Note: Observational data are data collected without design (all St Valentine customers of Stephanie's).



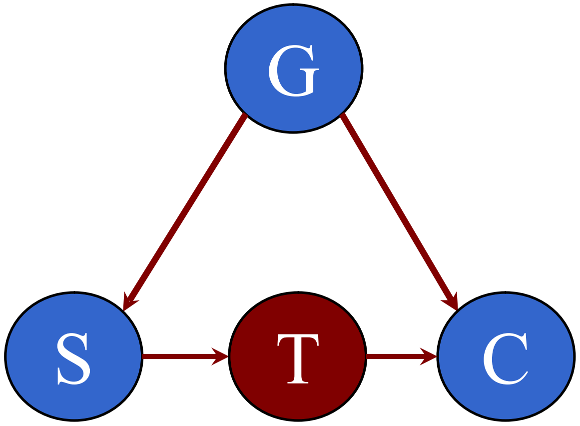
"Of Course Not!"

- Sir Ronald Fisher said.
- The correlation can be explained by a model in which there is no causal link between smoking and cancer but an unobserved genotype simultaneously causes cancer and craving for nicotine.
- Only a controlled experiment (once impossible now also illegal) could have the last word.



Auxiliary Variables

- The causal model rests on the assumption that smoking affects lungs through tar accumulation.
- This accumulation is a measurable quantity and can be used as a proxy of the causal dependency.



Measuring the Immeasurable

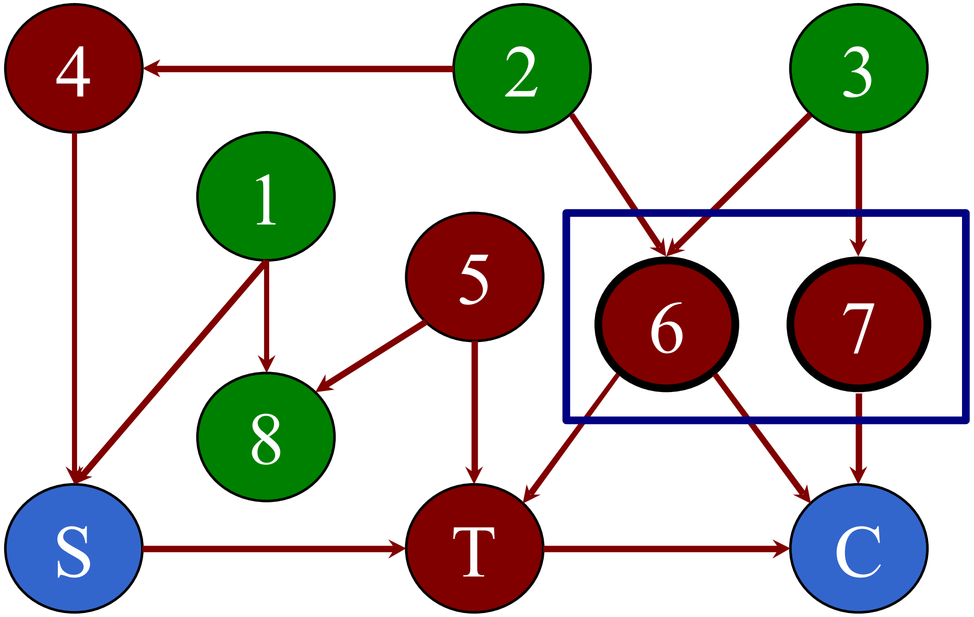
- Not all factors are measurable:
  - Measurable: tar concentration, age, income.
  - Non measurable: lifestyle, affective state, genotype.
- Can we use only measurable factors to rule out both measurable and non measurable factors and avoid the appearance of hidden variables and Simpson's paradox with them?
- This seems to be an experimental design problem but it can be used in observational studies as well.
- In statistics it is called the Adjustment Problem.

The Adjustment Problem

Adjustment: Which factors should be measured (or which experimental conditions should be kept still?).

Problem: Are factor 6 and 7 enough to avoid paradox?

Solution: Model the interaction of factors with a BBN.



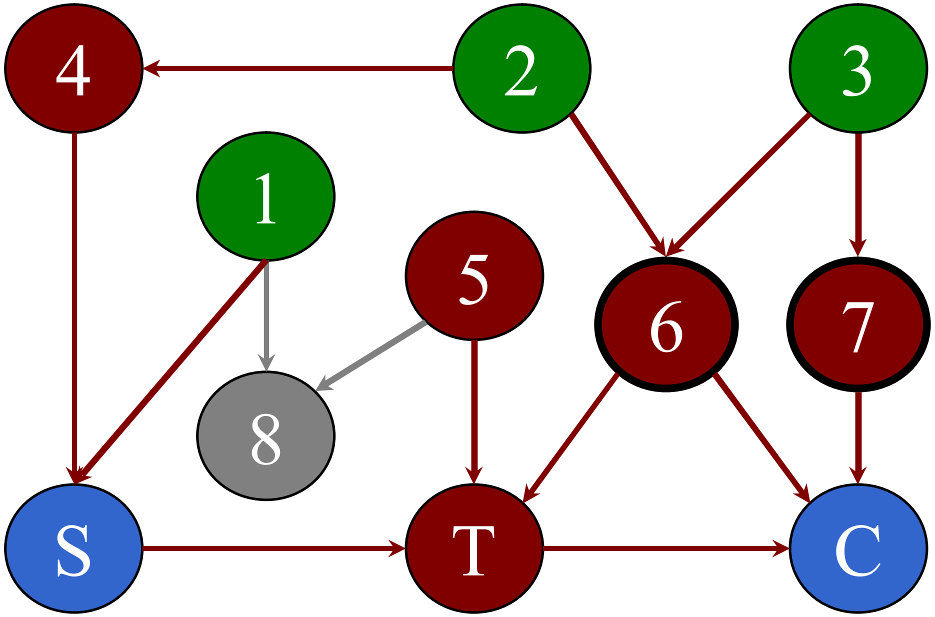
(blue)	Variables of Interest
(dark red)	Observable
(green)	Not Observable

The Adjustment Problem

Step 1: Build the model.

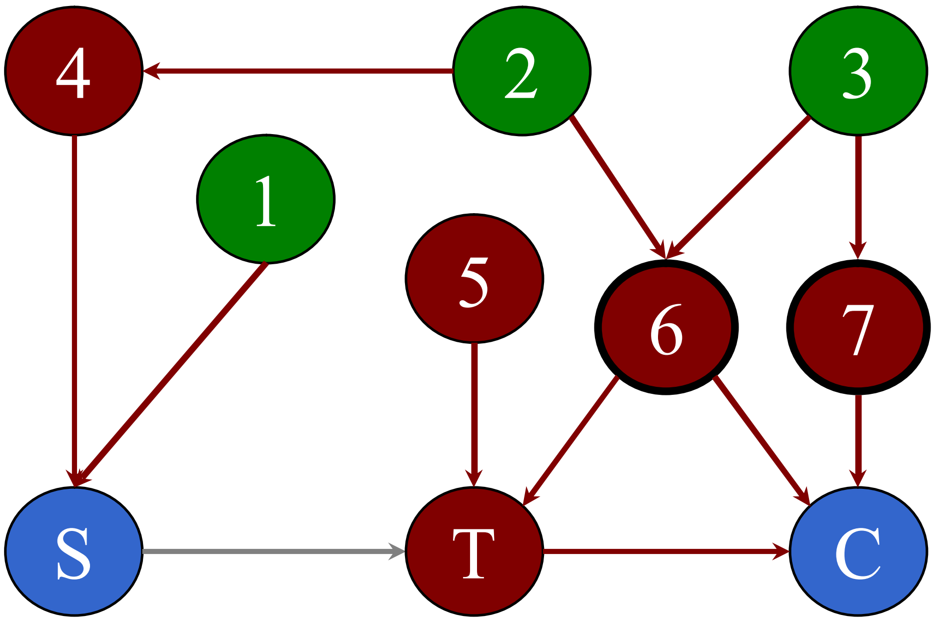
Note: Measurements should not be children of S and C.

Step 2: Remove all non ancestors of S, C, 6 and 7.



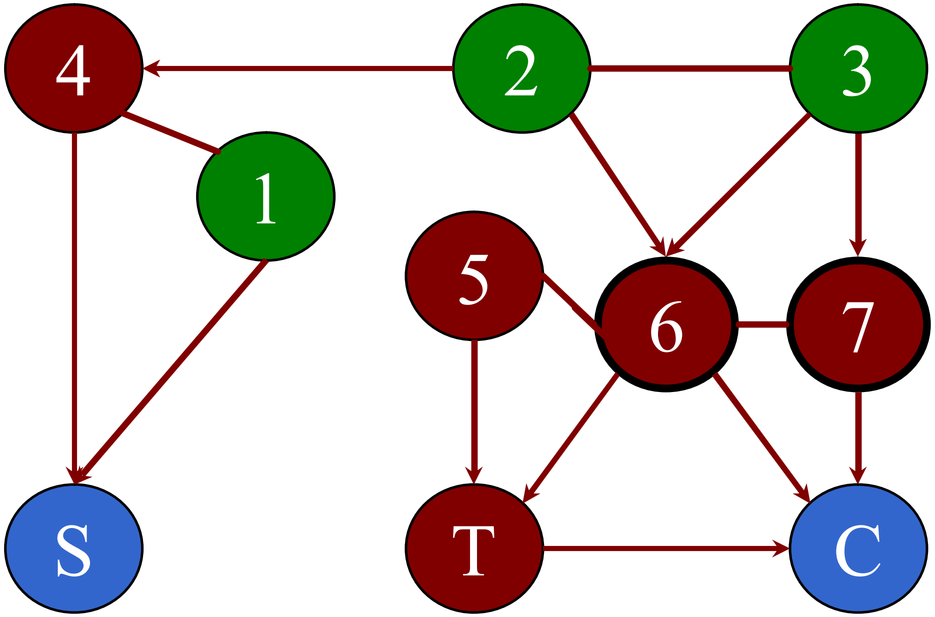
The Adjustment Problem

Step 3: Delete all arcs starting from S.



The Adjustment Problem

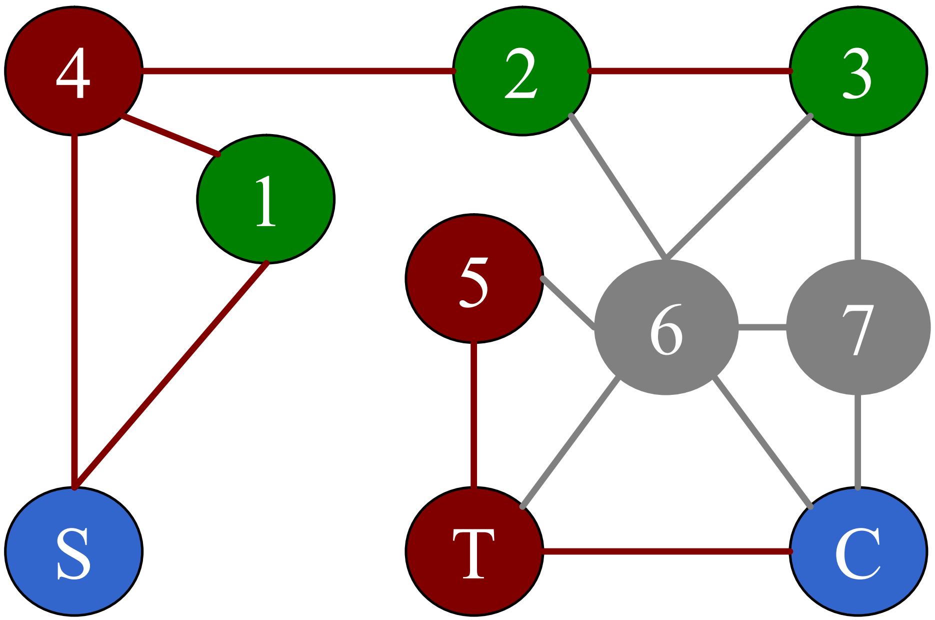
Step 4: Moralize (marry parents of a common child).



The Adjustment Problem

Step 5: Drop the directionality of the links (arrows).

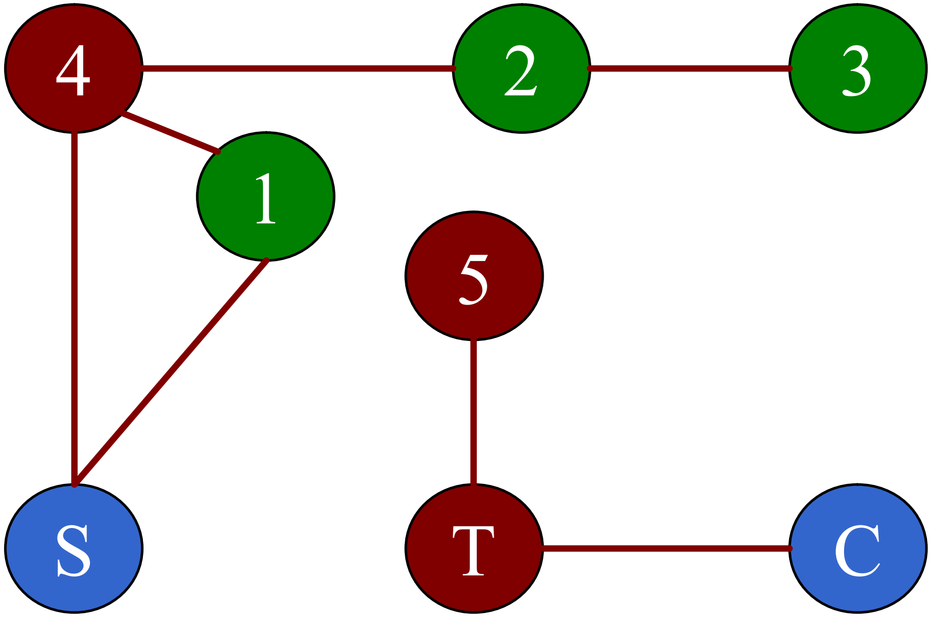
Step 6: Remove the factors to measure (6 and 7).



Solution

Test: If the variables of interest (S and C) are disconnected, then measurements are appropriate.

Answer: Yes.



Take Home

- Bayesian networks are a knowledge representation formalism to reason under uncertainty.
- They provide a semantics understandable to humans and facilitate the acquisition of modular knowledge.
- Bayesian networks can be learned from data.
- Hidden variables and not measurable quantities are major obstacles to causal discovery.
- Bayesian networks can be used to model causality, although their arcs aren’t necessarily causal.