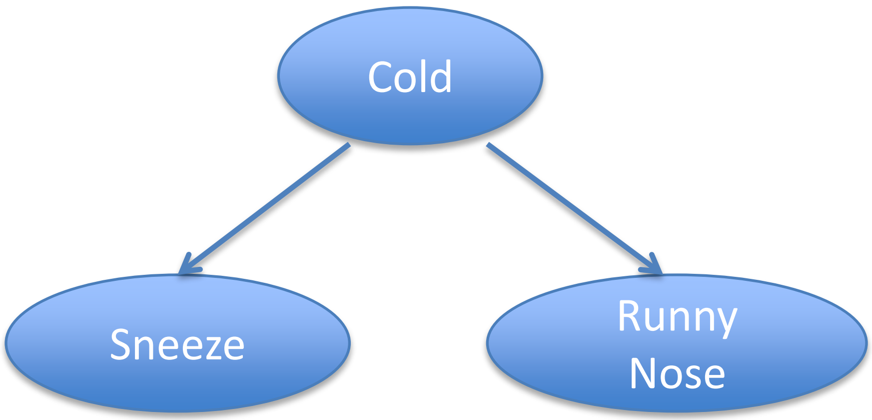
Bayesian Networks for Data Mining

Intro

- What is a Bayesian Network
  - Graphical method for representing conditional probabilities and causality
  - BayesTheorm
    - $P(H|E) = (P(E|H) P(H))/P(E)$
- Strengths of approach
  - Representing causality
  - Provide a method for dealing with missing data
  - Combine domain knowledge and data
  - Efficient approach to overfitting
- Bayesian Networks aren't specific to Bayesian techniques
  - DAGS
  - Causal Markov : Each node is independent of its non-descendants conditional on its parents

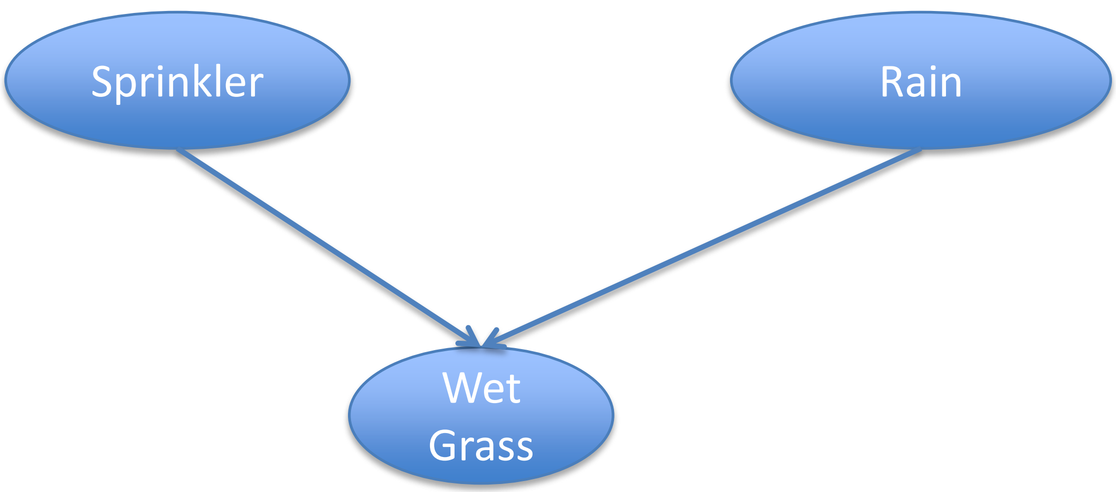
Examples

- $P(S|R)>P(S)$ $!I(S,R)$
- $P(S|R,C) = P(S|C)$ $I(S,R|C)$



Examples: “explaining away”

- $P(S|R)=P(S)$ $I(S,R)$
- $P(S|R,W) < P(S|W)$ $!I(S,R|W)$



Examples

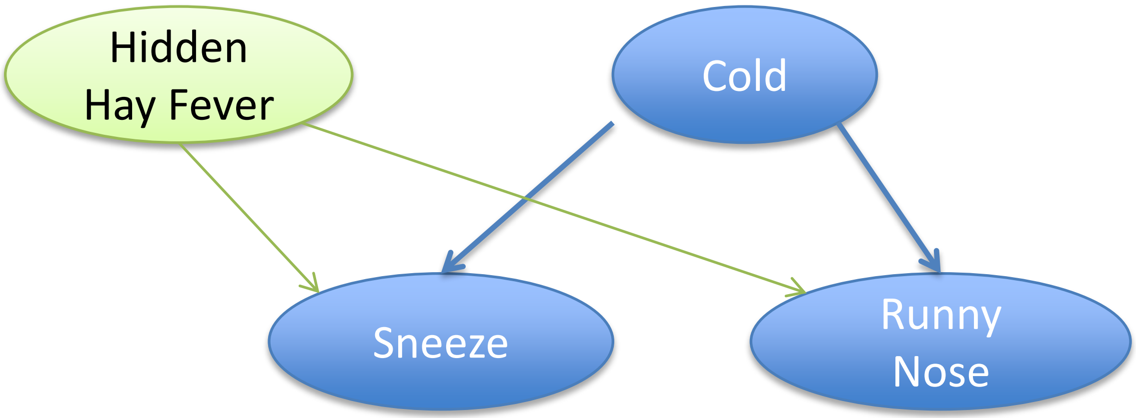
- $P(D|T)>P(D)$ $!I(D,T)$
- $P(D|T,S) = P(D|S)$ $I(D,T|S)$

Exceptions to Causal Markov

- Hidden common causes
- Causal Feedback
- Selection bias

Hidden Variables

- Hidden variables break conditional independence
  - Its no longer true that I(S,R|C)



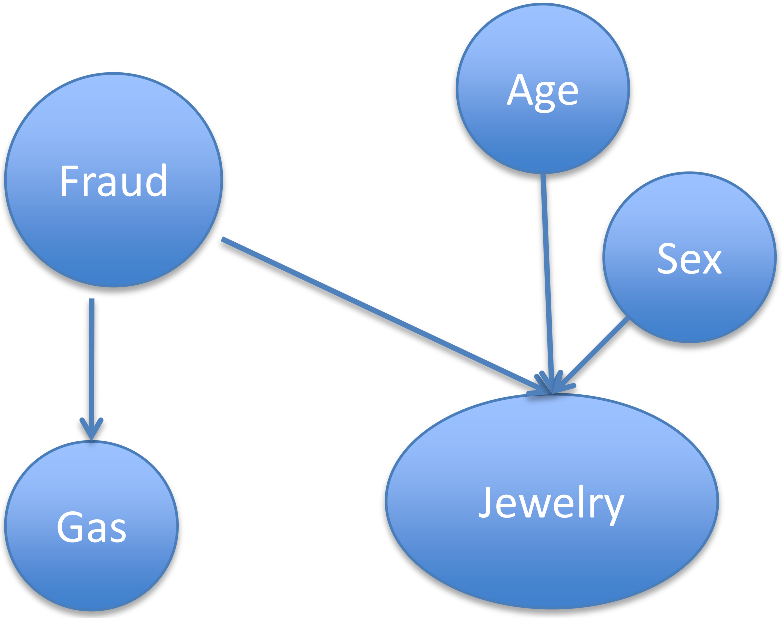
Bayesian vs. Frequentist

- Probability as "a degree of belief" (vs. a physical property, i.e. physical probability)
- Frequentist: given model parameters, W, (and est., W^*) imagine data sets D of size N that may be generated:
  - $E_{p(D|W)}(W) = \Sigma_D p(D_i|W)W^*(D_i)$
- Bayesian: given all the data D, imagine parameters, W, which could have generated D
  - $E_{p(W|D)}(W) = \Sigma_W p(W_i|D)W_i$

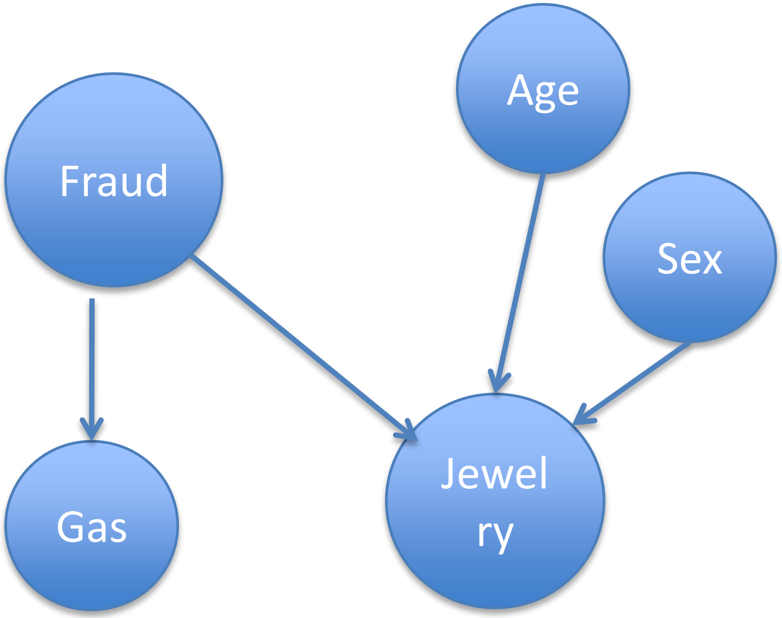
Probability Factorization:

- P(f)
- P(a|f) = P(a)
- P(s|a,f) = P(s)
- P(g|f, a, s) = P(g|f)
- P(j|f,a, s, g) = P(j|f, a, s)

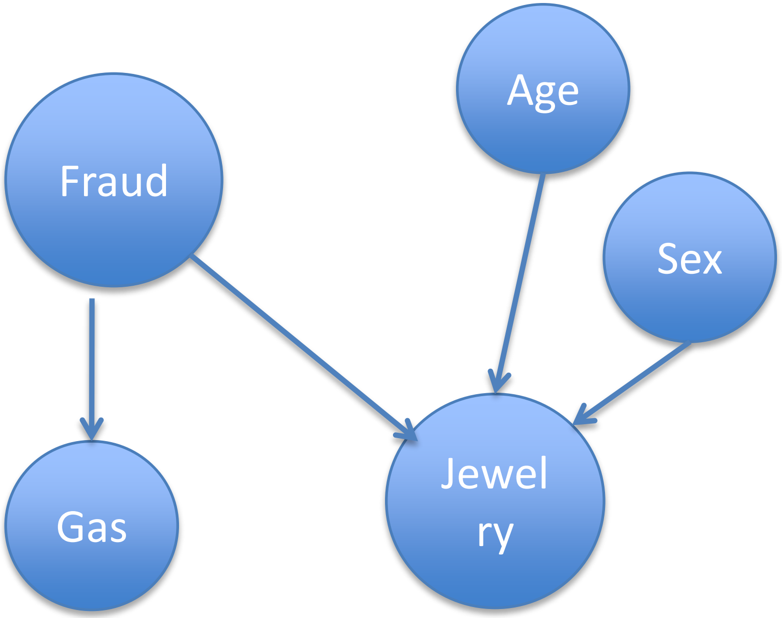
- n! orderings…



Inference



Inference



Learning Probabilities

- Local distribution function – $p(x_i|pa_i,w_i,S^h)$
  - :

- Compute $p(w_s|D,S^h)$
  - $w_s= (w_1,w_2,\ldots,w_n)$
  - $S^h$is the hypothesized network
  - w_i are the parameters on the ith node
- Assume no missing data, and the parameter vectors w_i are independant

- Incomplete Data: Absent data independent of state
  - Estimate missing x_i using an estimated p(x) based on x_{i-1} points (e.g. using Monte Carlo and Gibbs sampling, or Gaussian approximation, or MAP/ML

Discovering the network

See example in section 10.3

Model Selection

- Criterion defined measures degree to which a network fits the prior knowledge and data
  - Log of posterior probablility: log p(D,S) = log p(S) + log p(D|S)
  - Section 9 discusses methods for calculating log p(D|S)
  - Section 10 covers methods for calculating priors for S
- Local Criteria
  - Ignore relationships between the children
  - Predict for the lth child using each parent-child relationship for l-1 children

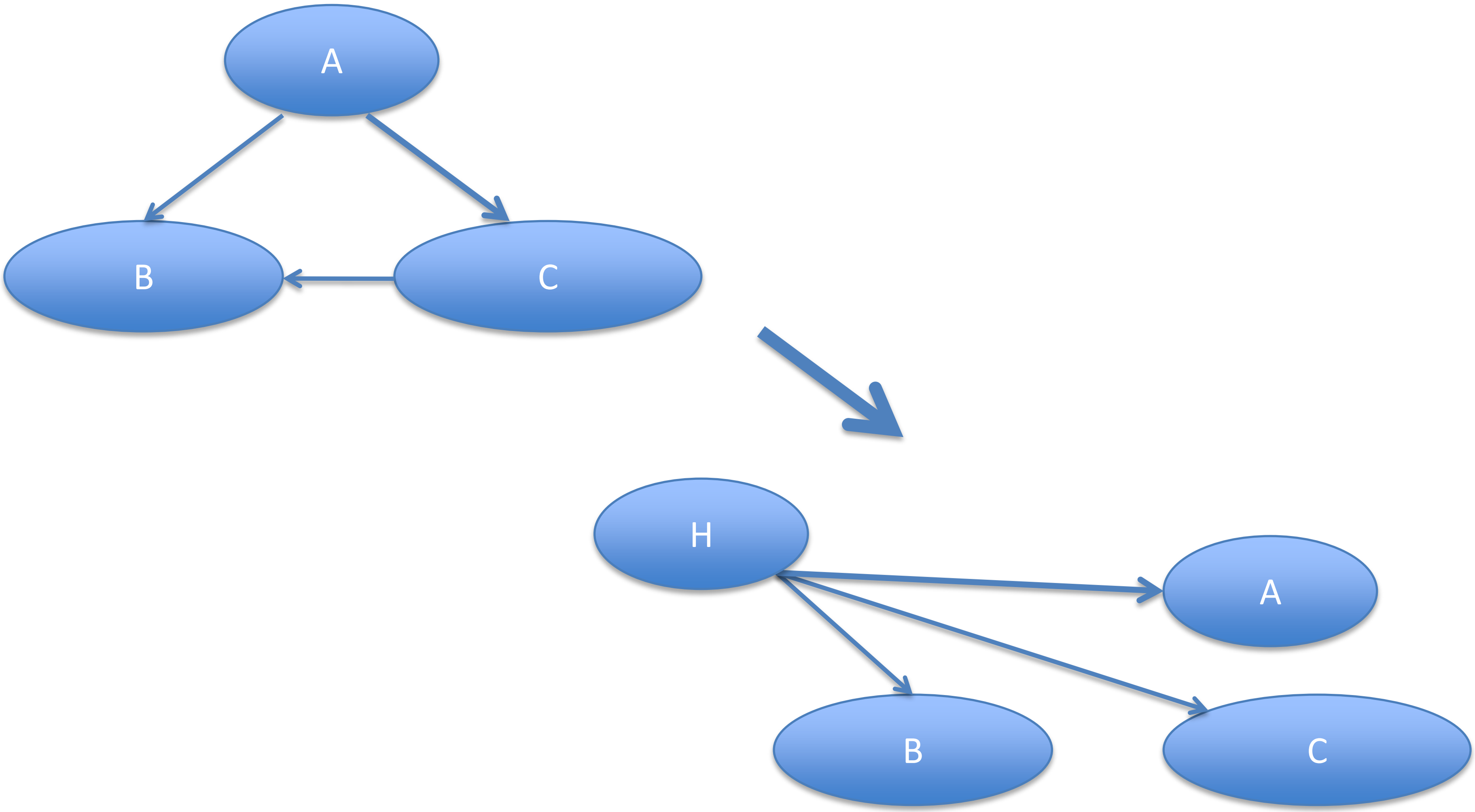
Supervised Learning

- Local distribution function – $p(x_i | pa_i, w_i, S^h)$
  - Each dist. Is like a classification problem (given the parents, predict the child
  - Train for each dist you need
  - Complete data means Bayesian and other classifiers are essentially equivalent

Unsupervised Learning

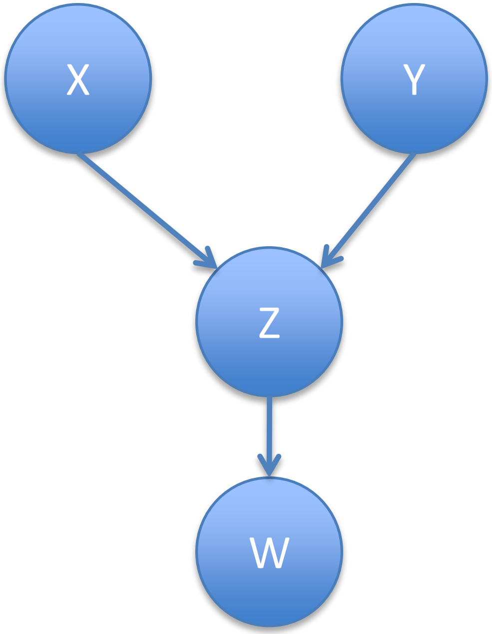
- Identifying hidden variables
  - Model selection assuming no hidden variables
  - Given model, look for sets of mutually dependant variables
  - For each dependant set, add a hidden variable
  - Rescore model to see if we're better off

Adding a Hidden Variable



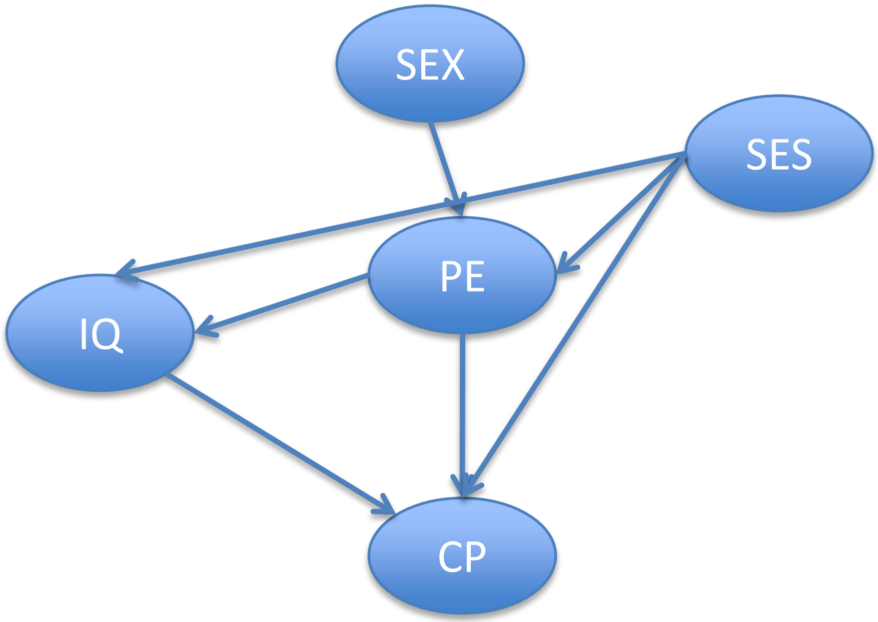
Causality

- I(x,y) and I(w, {x,y} | z)
- Does W cause Z or does Z cause W?
  - Note that we don't see I(X,W)



Case Study

- Discover a network



Case Study

- Hypothesize a hidden variable: “parental quality”