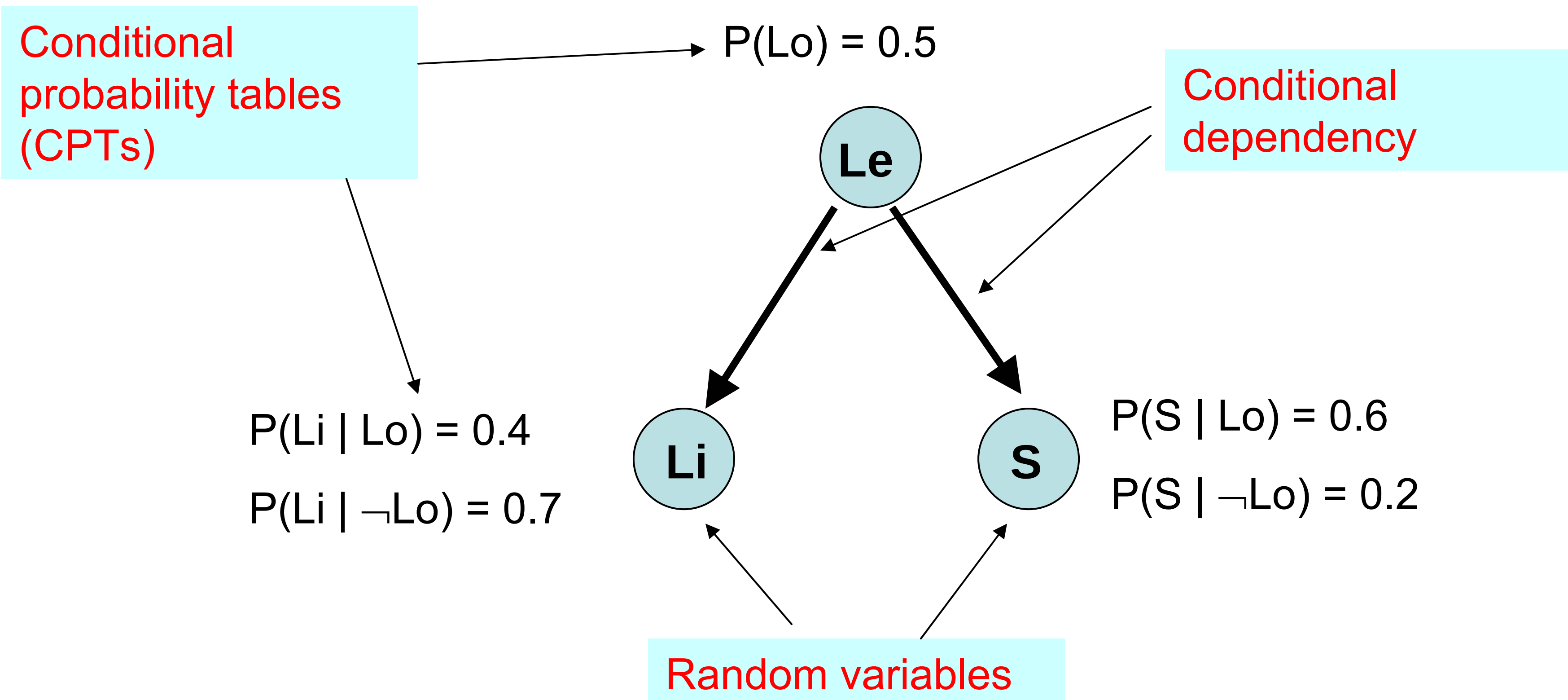
15-780: Graduate Artificial Intelligence

Inference in Bayesian networks

Bayesian networks: Notations



A example problem

- An alarm system

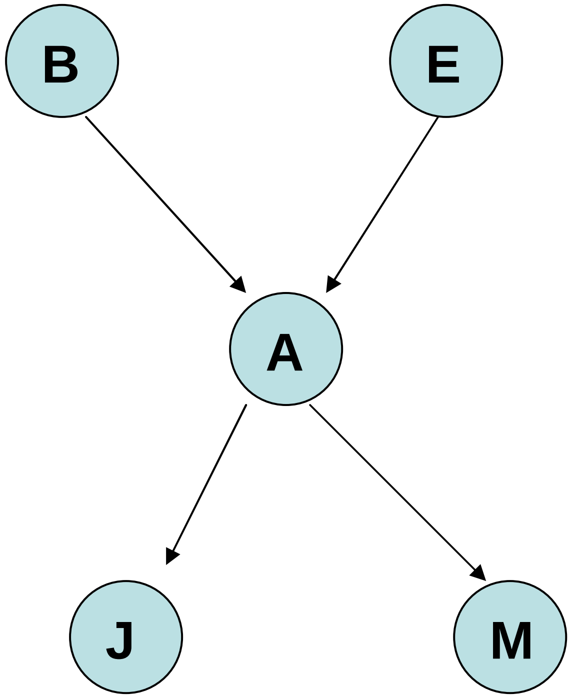
  B – Did a burglary occur?

  E – Did an earthquake occur?

  A – Did the alarm sound off?

  M – Mary calls

  J – John calls



Constructing a Bayesian network: Revisited

- Step 1: Identify the random variables
- Step 2: Determine the conditional dependencies

  - Select on ordering of the variables

  - Add them one at a time

  - For each new variable X added select the minimal subset of nodes as parents such that X is independent from all other nodes in the current network given its parents.

- Step 3: Populate the CPTs

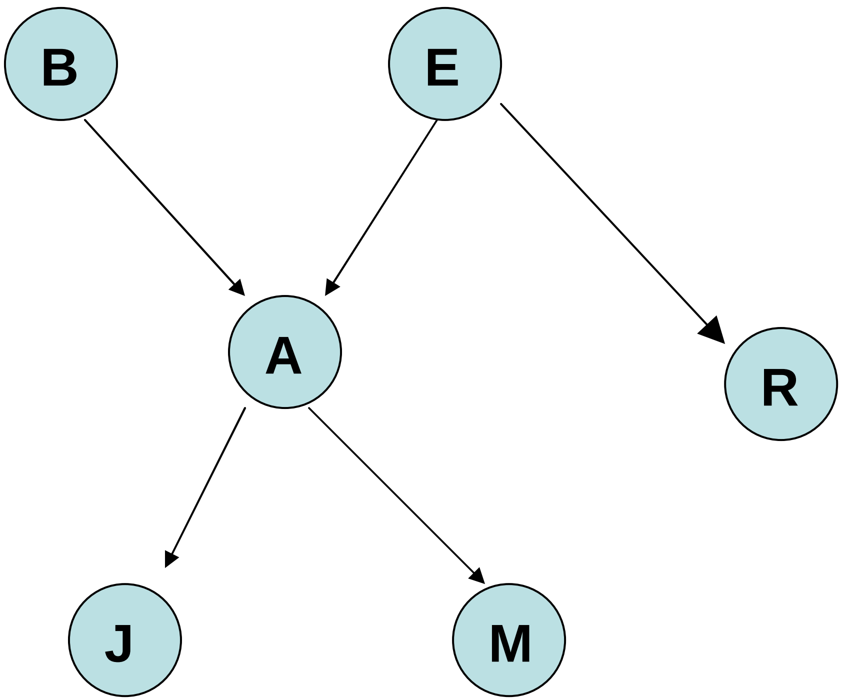
  - We will discuss this when we talk about density estimations

Reconstructing a network

Suppose we wanted to add a new variable to the network:

R – Did the radio announce that there was an earthquake?

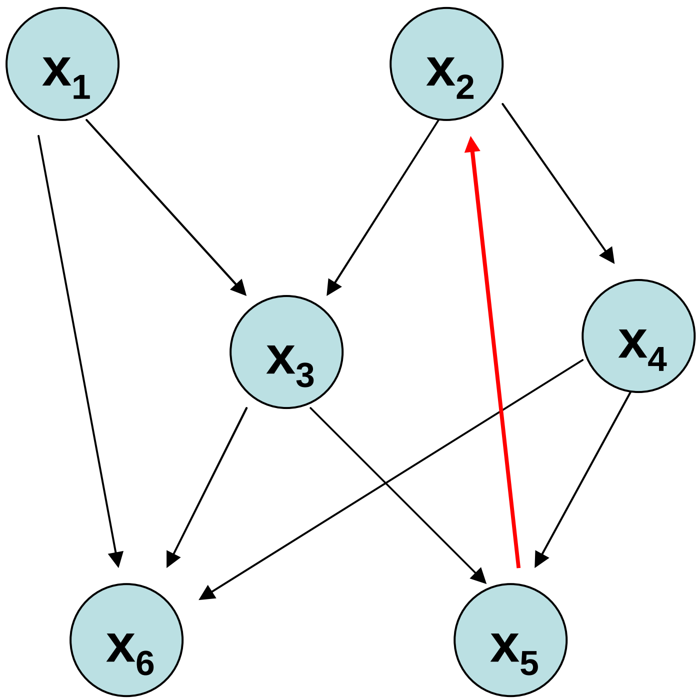
How should we insert it?



Bayesian networks: Restrictions and joint distributions

- Bayesian networks are directed acyclic graphs (DAGs)

  - Otherwise a node will impact (indirectly) its own probability making inference hard



This is NOT a valid Bayesian network!

Bayesian networks: Restrictions and joint distributions

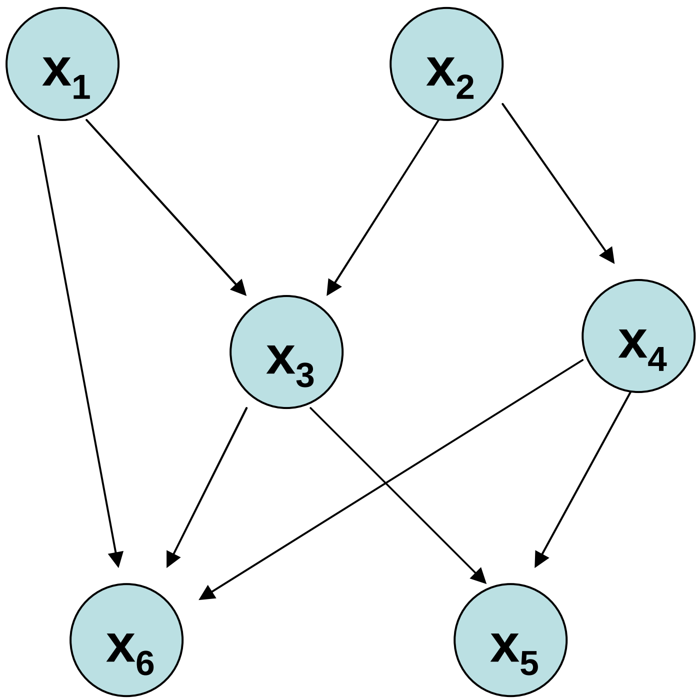
- Bayesian networks are directed acyclic graphs (DAGs)

  - Otherwise a node will impact (indirectly) its own probability making inference hard

- Given a Bayesian network the joint probability distribution can be factored as:

$$P(X) = \prod_i p(x_i \mid Pa(x_i))$$

where X is a vector of observations and $Pa(x_i)$ is the set of parent nodes of x_i

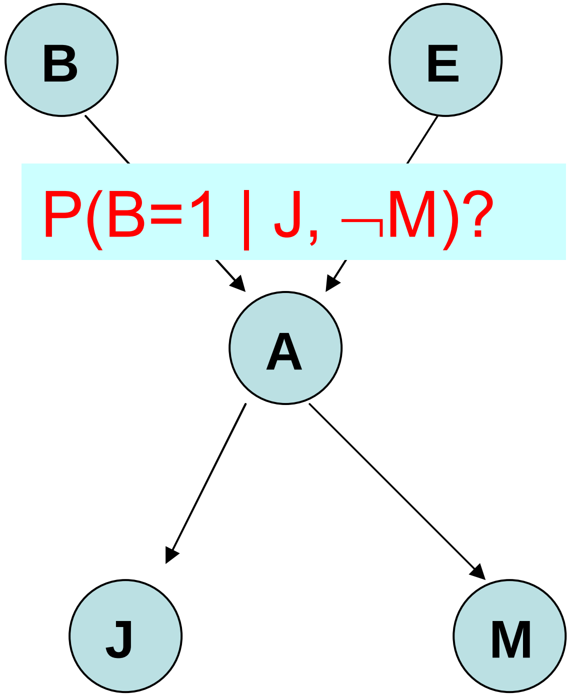


Using Bayesian networks

- Inference
  - Computing joint distributions
  - Inferring values of unobserved variables
- Structure learning

Bayesian network: Inference

- Once the network is constructed, we can use algorithms for inferring the values of unobserved variables.
- For example, in our previous network the only observed variables are the phone calls. However, what we are really interested in is whether there was a burglary or not.
- How can we determine that?



Inference

- Lets start with a simpler question
  - How can we compute a joint distribution from the network?
  - For example, $P(B,\neg E,A,J,\neg M)$?
- Answer:
  - That's easy, lets use the network

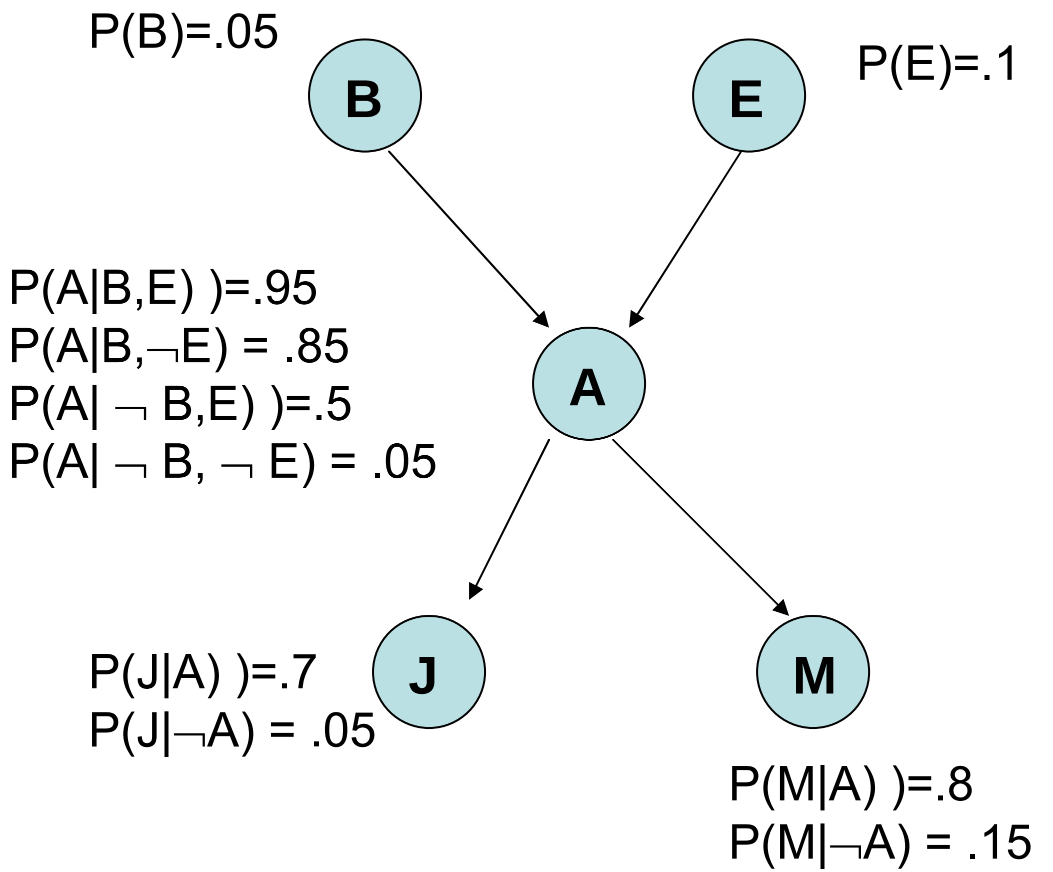
Computing: P(B,¬E,A,J, ¬M)

P(B,¬E,A,J, ¬M) =

P(B)P(¬E)P(A | B, ¬E)
P(J | A)P(¬M | A)

= 0.05*0.9*.85*.7*.2

= 0.005355



Computing: $P(B,\neg E,A,J,\neg M)$

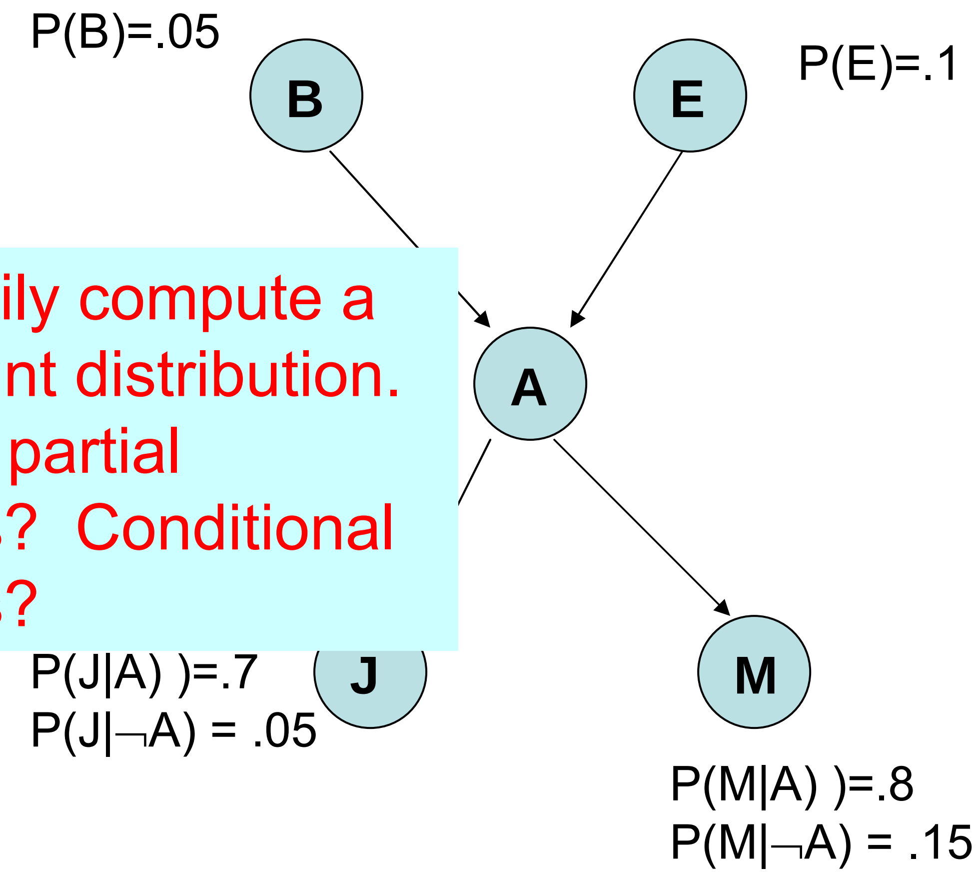
$P(B,\neg E,A,J,\neg M) =$

$P(B)P(\neg E)P(A \mid B,\neg E)$
$P(J \mid A)P(\neg M \mid A)$

$= 0.05*0.9*.85*.7*.2$

$= 0.005355$

We can easily compute a complete joint distribution. What about partial distributions? Conditional distributions?



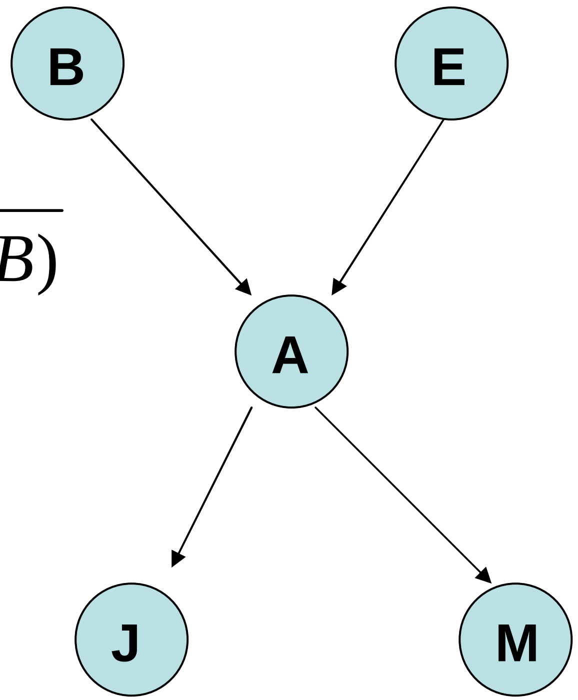
Inference

- We are interested in queries of the form:
  P(B | J,¬M)
- This can also be written as a joint:

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M, B) + P(J, \neg M, \neg B)}$$

chain rule

- How do we compute the new joint?



Computing partial joints

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(B, J, \neg M) + P(\neg B, J, \neg M)}$$

Sum all instances with these settings (the sum is over the possible assignments to the other two variables, E and A)

Computing: P(B,J, ¬M)

P(B,J, ¬M) =

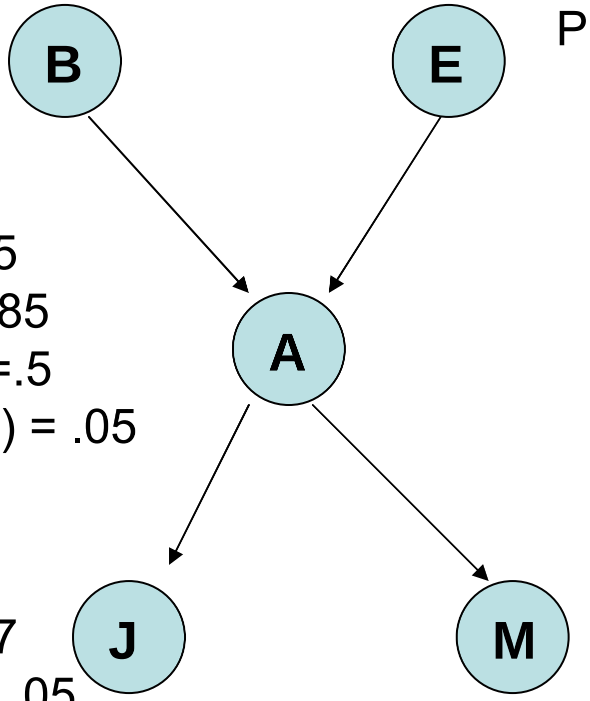
P(B,J, ¬M,A,E)+

P(B,J, ¬M, ¬ A,E) +
P(B,J, ¬M,A, ¬ E) +
P(B,J, ¬M, ¬ A, ¬ E) =

0.0007+0.00001+0.005+0.
0003 = 0.00601

P(B)=.05

P(E)=.1



P(A|B,E))=.95
P(A|B,¬E) = .85
P(A| ¬ B,E))=.5
P(A| ¬ B, ¬ E) = .05

P(J|A))=.7
P(J|¬A) = .05

P(M|A))=.8
P(M|¬A) = .15

Computing partial joints

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(B, J, \neg M) + P(\neg B, J, \neg M)}$$

Done!

Sum all instances with these settings

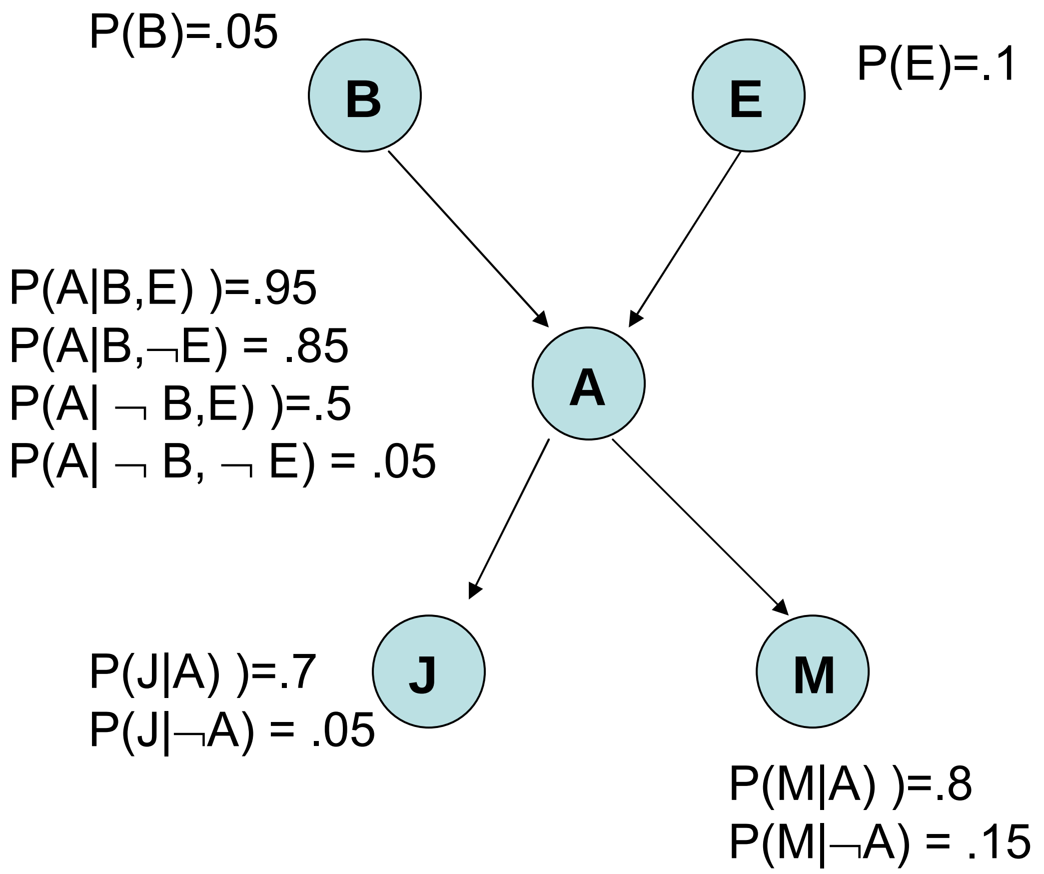
Computing: $P(\neg B, J, \neg M)$

$P(\neg B, J, \neg M) =$

$P(\neg B, J, \neg M, A, E) +$

$P(\neg B, J, \neg M, \neg A, E) +$
$P(\neg B, J, \neg M, A, \neg E) +$
$P(\neg B, J, \neg M, \neg A, \neg E) =$

0.00665+0.002+0.006+0.0
345 = 0.049



Computing partial joints

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(B, J, \neg M) + P(\neg B, J, \neg M)}$$

$$= \frac{0.006}{0.006 + 0.049} = 0.11$$

Computing partial joints

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(B, J, \neg M) + P(\neg B, J, \neg M)}$$

Sum all instances with these settings (the sum is over the possible assignments to the other two variables, E and A)

But the number of possible assignments is exponential in the unobserved variables?

That is, unfortunately, the best we can do. General querying of Bayesian networks is NP-complete

Inference in Bayesian networks if NP complete (sketch)

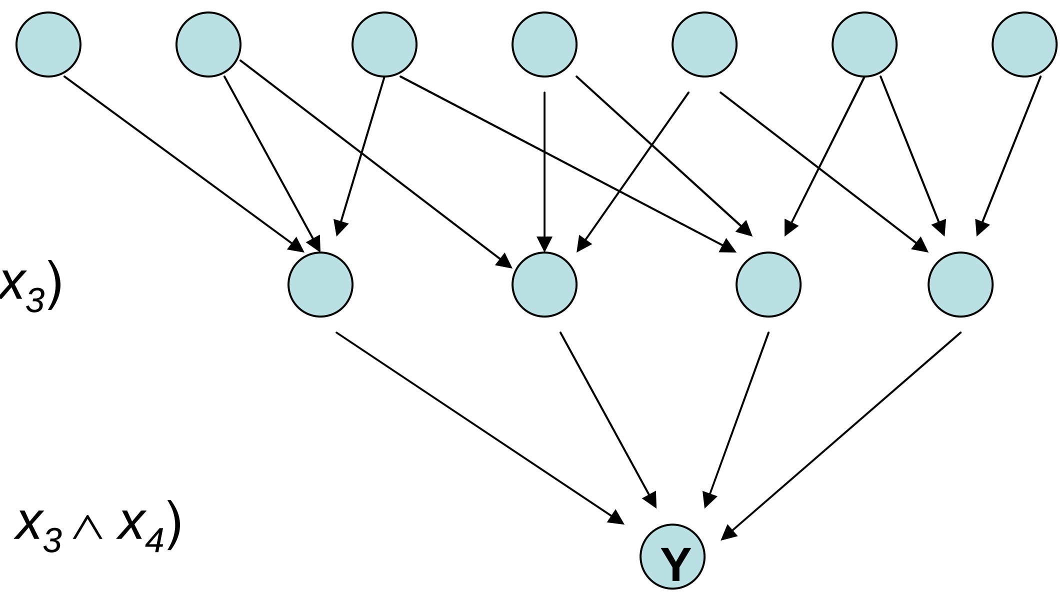
- Reduction from 3SAT
- Recall: 3SAT, find satisfying assignments to the following problem: $(a \vee b \vee c) \wedge (d \vee \neg b \vee \neg c) \ldots$

What is P(Y)?

$P(x_i=1) = 0.5$

$P(x_i=1) = (x_1 \vee x_2 \vee x_3)$

$P(Y=1) = (x_1 \wedge x_2 \wedge x_3 \wedge x_4)$



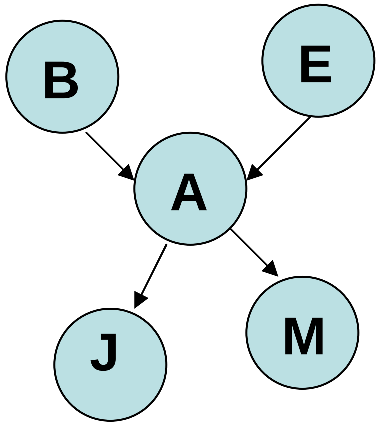
Other inference methods

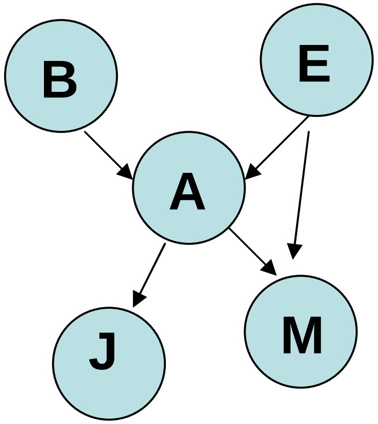
- Convert network to a polytree

  - In a polytree no two nodes have more than one path between them

  - For such a graph there is a linear time algorithm

  - However, converting into a polytree requires a large increase in the size of the graph (number of nodes)



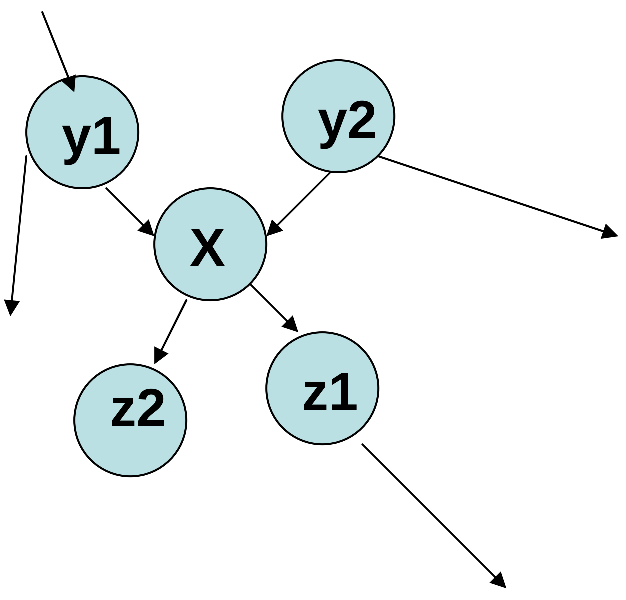


Why is inference in polytrees easy?

- In polytrees, given a variable X we can always divide the other variables into two sets:

  E+: Variables ‘above’ X

  E-: Variables ‘below’ X
- These sets are mutually exclusive (why?)
- Using these sets we can efficiently compute conditional and joint distributions



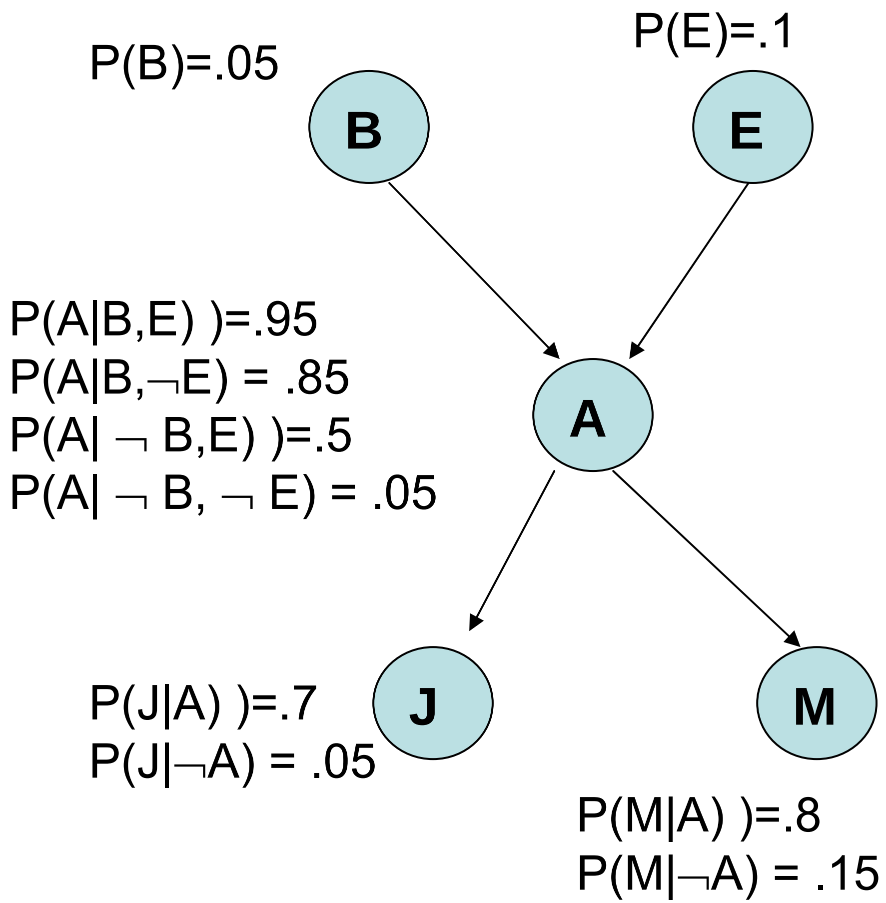
Stochastic inference

- We can easily sample the joint distribution to obtain possible instances

1. Sample the free variable
2. For every other variable:

   - If all parents have been sampled, sample based on conditional distribution

We end up with a new set of assignments for B,E,A,J and M which are a random sample from the joint



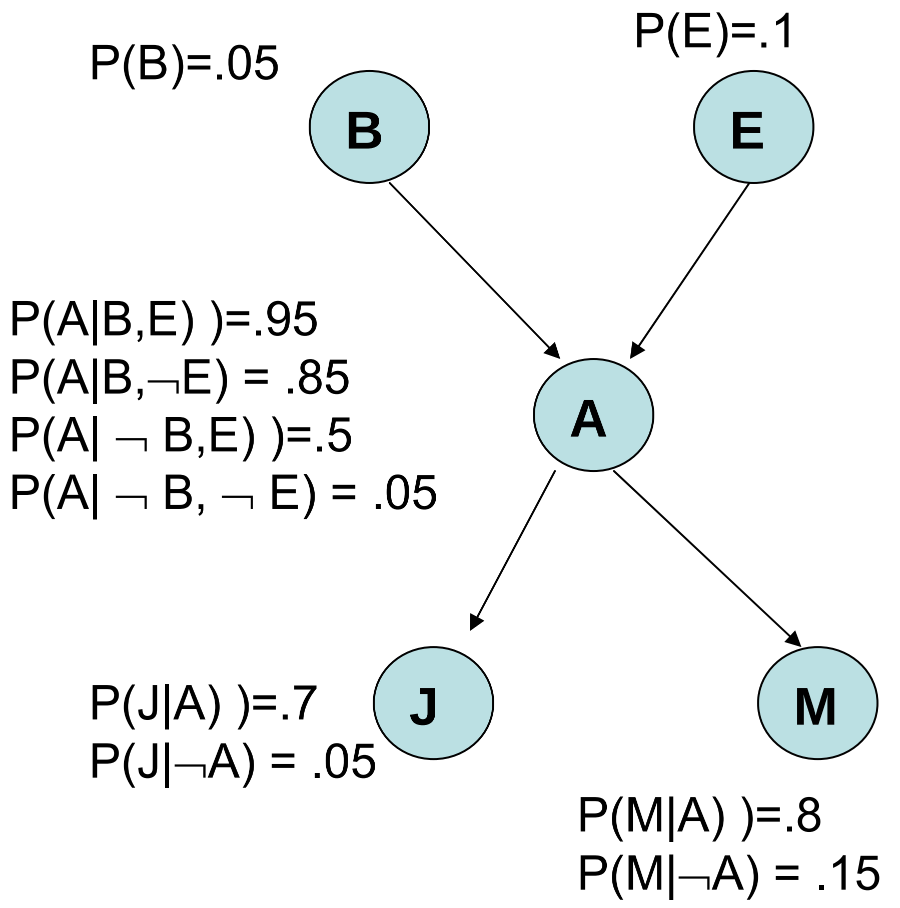
Stochastic inference

- We can easily sample the joint distribution to obtain possible instances

1. Sample the free variable
2. For every other variable:

   - If all parents have been sampled, sample based on conditional distribution

Is it always possible to carry out this sampling procedure? why?



Using sampling for inference

- Lets revisit our problem: Compute $P(B \mid J, \neg M)$
- Looking at the samples we can cound:

  - N: total number of samples

  - N_c : total number of samples in which the condition holds $(J, \neg M)$

  - N_B: total number of samples where the joint is true $(B, J, \neg M)$
- For a large enough N
  - N_c / N $\approx P(J, \neg M)$
  - N_B / N $\approx P(B, J, \neg M)$
- And so, we can set

$$P(B \mid J, \neg M) = P(B, J, \neg M) / P(J, \neg M) \approx N_B / N_c$$

Using sampling for inference

- Lets revisit our problem: Compute P(B | J,¬M)
- Looking at the samples we can cound:
  - N: total number o
  - N_c : total number
  - N_B: total number
- For a large enoug
  - N_c / N ≈ P(J,¬M)
  - N_B / N ≈ P(B,J,¬M)
- And so, we can set

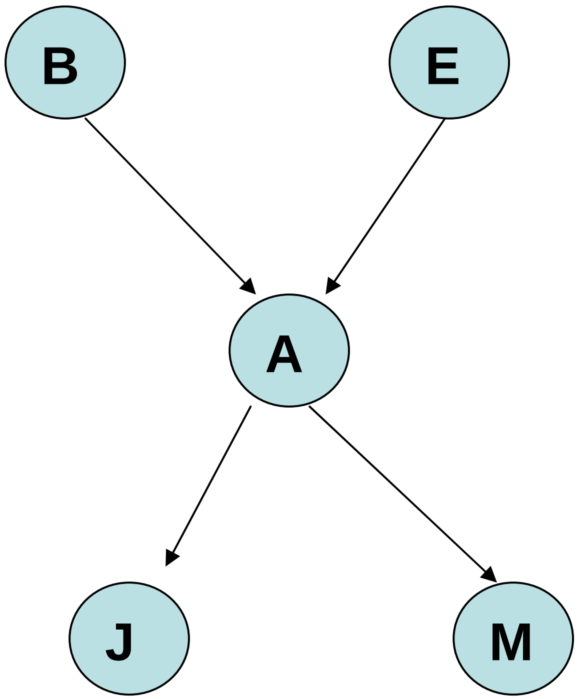
P(B | J,¬M) = P(B,J,¬M) / P(J,¬M) ≈ N_B / N_c

Problem: What if the condition rarely happens?

We would need lots and lots of samples, and most would be wasted

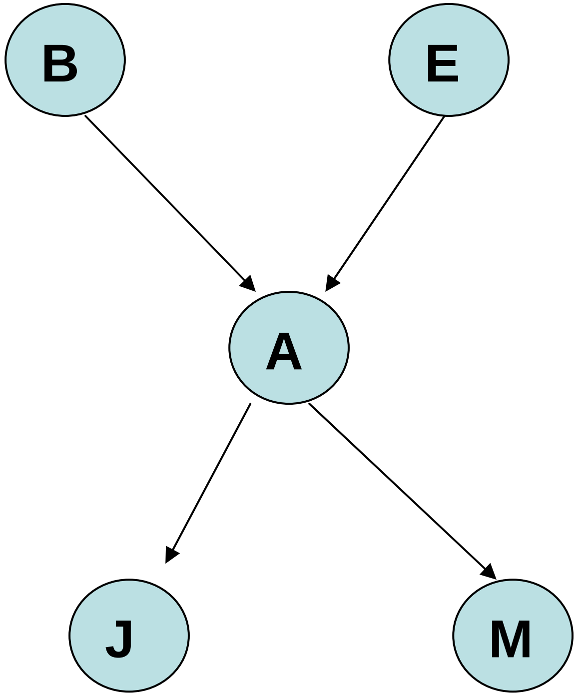
Weighted sampling

- Compute P(B | J,¬M)
- We can manually set the value of J to 1 and M to 0
- This way, all samples will contain the correct values for the conditional variables
- Problems?



Weighted sampling

- Compute P(B | J,¬M)
- Given an assignment to parents, we assign a value of 1 to J and 0 to M.
- We record the *probability* of this assignment ($w = p_1 p_2$) and we weight the new joint sample by *w*



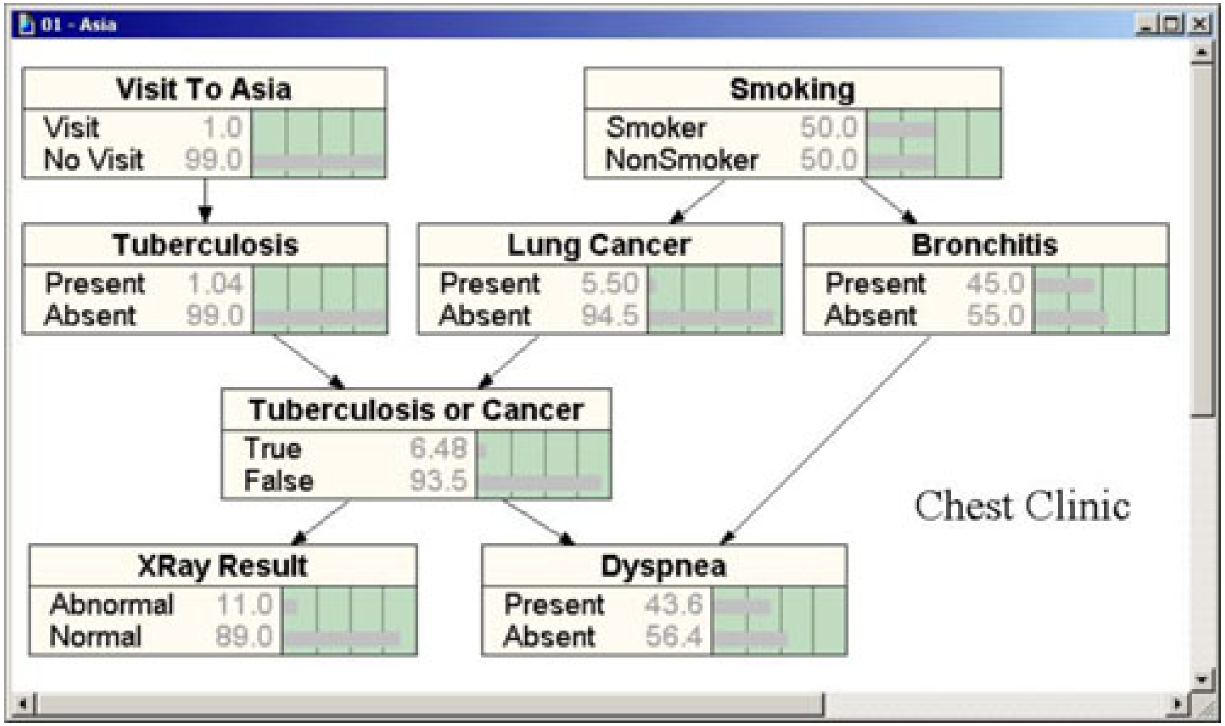
Weighted sampling algorithm for computing P(B | J,¬M)

- Set $N_B, N_c = 0$
- Sample the joint setting the values for *J* and *M*, compute the weight, *w,* of this sample
- $N_c = N_c + w$
- If $B = 1$, $N_B = N_B + w$

- After many iterations, set

$$P(B \mid J, \neg M) = N_B / N_c$$

Bayesian networks for cancer detection



Constructing networks

- So far we assumed that the network is derived from domain knowledge.
- That's not always easy to do
- Examples:

  - How are different regions in the brain related?

  - How are terrorists related (social networks)?

Inferring structure from data

- It is possible to infer structure if enough data is provided
- The goal would be to find a structure that leads to a *maximal likelihood**

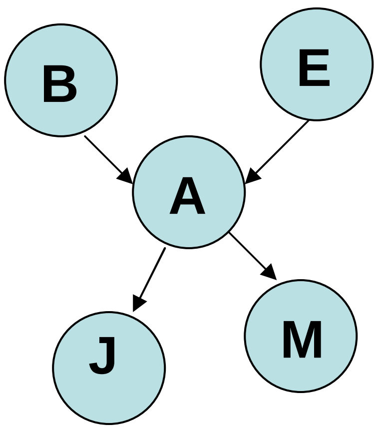
$$\mathrm{Max}_S P(D \mid S)$$

- Problems?

*More on this next week

Inferring structure using maximum likelihood principle

- The more edges we have, the higher the likelihood!

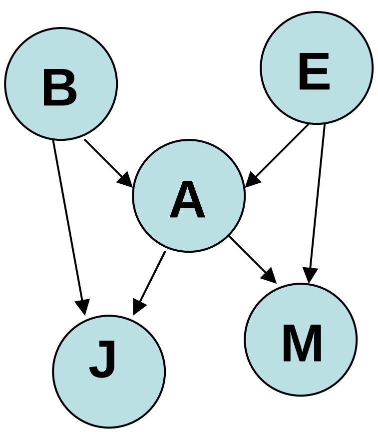


P(M | A, E) ≥ P(M | A)

Why?

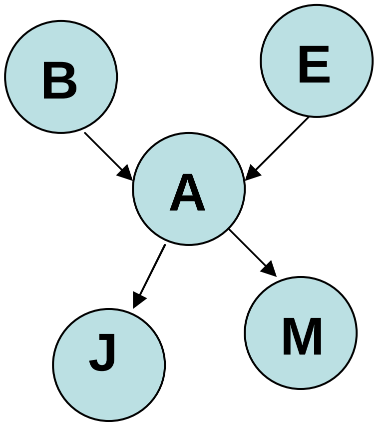
- If the two are independent and we have perfect data, trivially holds

- We have more parameters to fit. If there is some noise in the measurements, we would likely overfit the data



Inferring structure using maximum likelihood principle

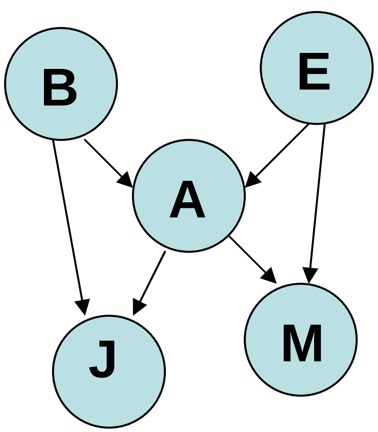
- The more edges we have, the higher the likelihood!



P(M | A, E) ≥ P(M | A)

Solutions:

- Statistical tests
- Penalty functions



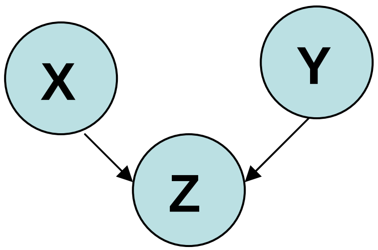
Likelihood ratio test

- Given two competing models we can compute their likelihood ratio

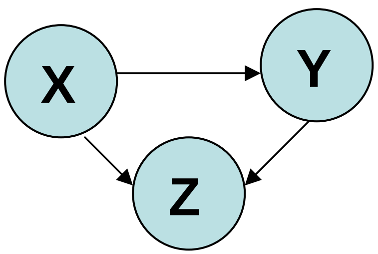
$$T(D) = 2\log\frac{P(D\mid B)}{P(D\mid A)}$$

Always ≥ 0, but by how much?

Model A



Model B



Likelihood ratio test

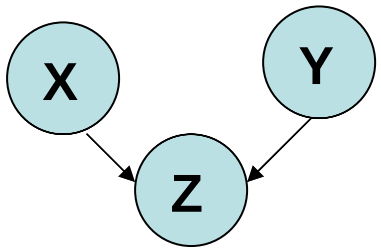
- Given two competing models we can compute their likelihood ratio

$$T(D) = 2\log\frac{P(D \mid B)}{P(D \mid A)} \sim \chi^2$$

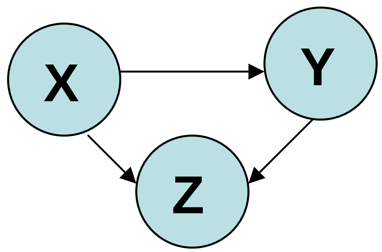
Always ≥ 0, but by how much?

The result is distributed according to χ^2, which is a distribution defined by the number of free parameters (the difference in complexity of the two models)

Model A



Model B



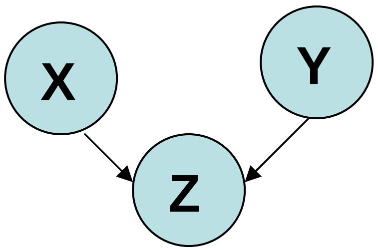
Likelihood ratio test

- Given two competing models we can compute their likelihood ratio

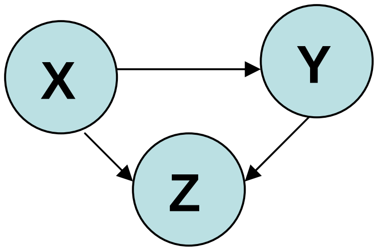
$$T(D) = 2\log\frac{P(D \mid B)}{P(D \mid A)} \sim \chi^2$$

Reject the more complicated model, unless the ratio is high enough (can use, for example, the Matlab function CHI2PDF to compute the probability of seeing this ratio as a result of noise).

Model A



Model B



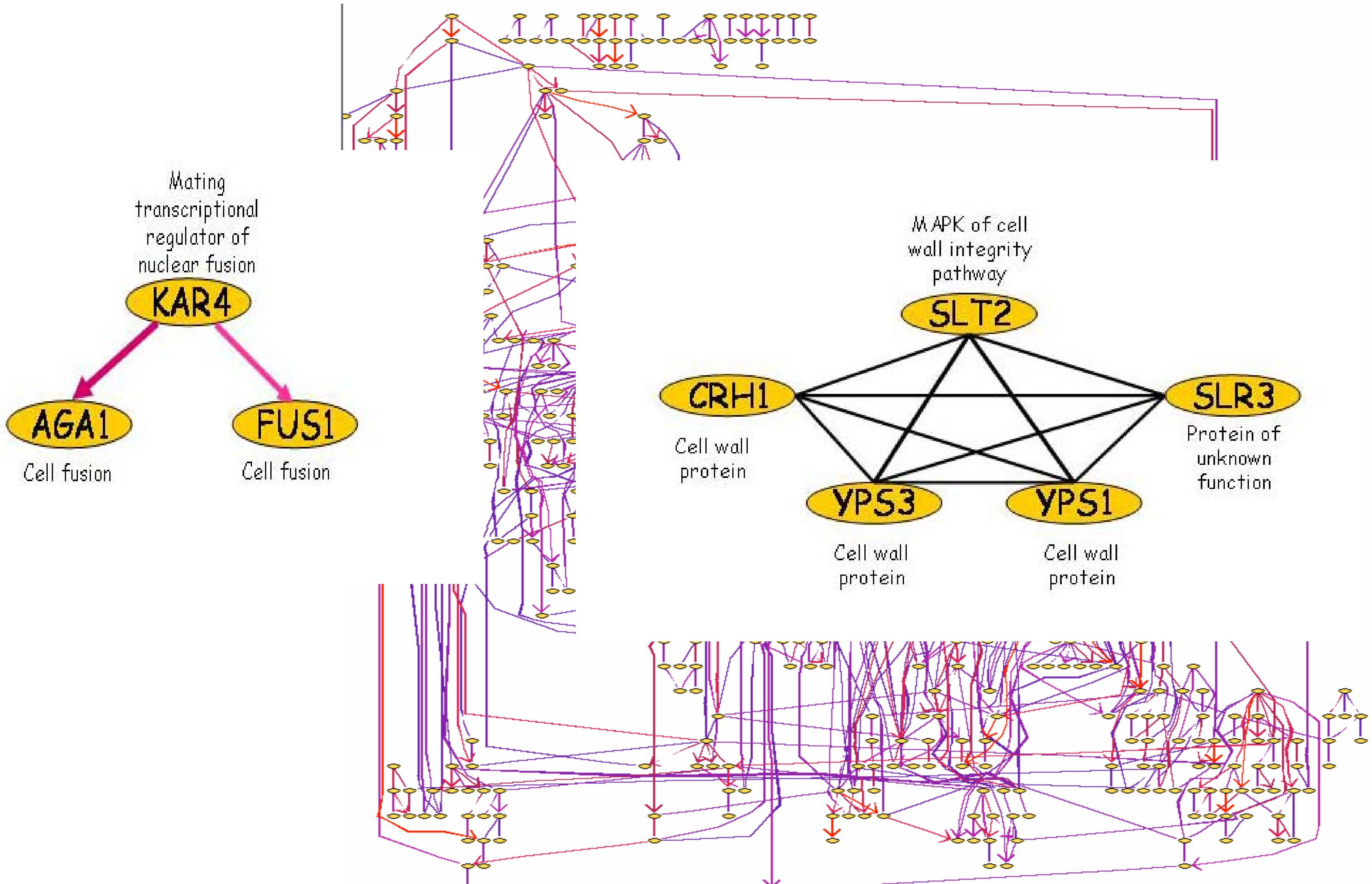
Penalty functions

- Likelihood ratio tests are appropriate for relatively small problems (few variables)
- For larger problems we usually use a penalty function
- This function penalizes the likelihood based on the complexity of the model

$$L(D \mid M) = P(D \mid M) - f(n)$$

  where n is related to the number of parameters
- Most commonly used penalty function:
  - AIC: Akaike's Information Criterion
  - BIC: Bayesian information criterion

Structure learning for biology



Important points

- Bayes rule
- Joint distribution, independence, conditional independence
- Definition Bayesian networks
- Inference in Bayesian networks
- Constructing a Bayesian network