

Probabilistic Graphical Models

Machine Learning 10-601B

Seyoung Kim

Many of these slides are derived from Tom
Mitchell, Ziv Bar-Joseph. Thanks!

Naïve Bayes Classifier Revisited

Full joint distribution:
exponential number of
parameters

$$P(X|Y) = P(X_1, \dots, X_n | Y)$$



Probabilistic
graphical models


Conditional independence
assumption for Naïve Bayes classifier:
linear number of parameters

$$P(X_1 \dots X_n | Y) = \prod_i P(X_i | Y)$$

Probabilistic Graphical Models

- Key Idea:
 - Conditional independence assumptions useful
 - but Naïve Bayes is extreme!
 - Probabilistic graphical models are a joint probability distribution defined over a graph
 - sets of conditional independence assumptions are expressed via graph structure
- Two types of graphical models:
 - Directed graphs (aka Bayesian Networks)
 - Undirected graphs (aka Markov Random Fields)

today





MORE ACM AWARDS

A.M. TURING AWARD



A.M. TURING AWARD WINNERS BY...

ALPHABETICAL LISTING

YEAR OF THE AWARD

RESEARCH SUBJECT



Photo-Essay

BIRTH:

September 4, 1936, Tel Aviv.

EDUCATION:

B.S., Electrical Engineering (Technion, 1960); M.S., Electronics (Newark College

JUDEA PEARL

United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.



SHORT ANNOTATED
BIBLIOGRAPHY



ACM DL
AUTHOR PROFILE



ACM TURING AWARD
LECTURE VIDEO



RESEARCH
SUBJECTS



ADDITIONAL
MATERIALS

Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for *causal inference* that has had significant impact in the social sciences.

Marginal Independence

Definition: X is marginally independent of Y if

$$(\forall i, j) P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

Equivalently, if

$$(\forall i, j) P(X = x_i | Y = y_j) = P(X = x_i)$$

Equivalently, if

$$(\forall i, j) P(Y = y_i | X = x_j) = P(Y = y_i)$$

Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

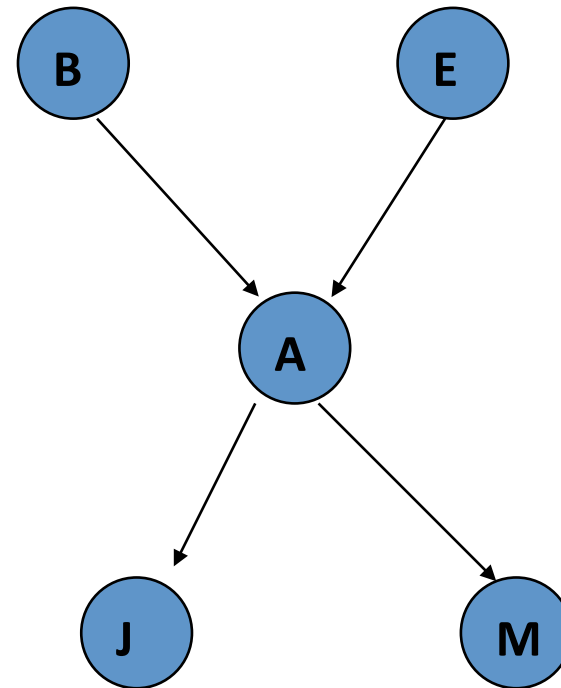
which we often write

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

E.g., $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

Directed Acyclic Graph

- A directed acyclic graph is defined by:
 - A set of nodes
 - A set of edges, such that directed edges do not form a cycle
- Terminology
 - **Parents** of node v
 - **Children** of node v
 - **Descendants** of node v



Bayesian Networks Definition

A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

- Each node denotes a random variable
- Edges denote dependencies

Bayesian Networks Definition

A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

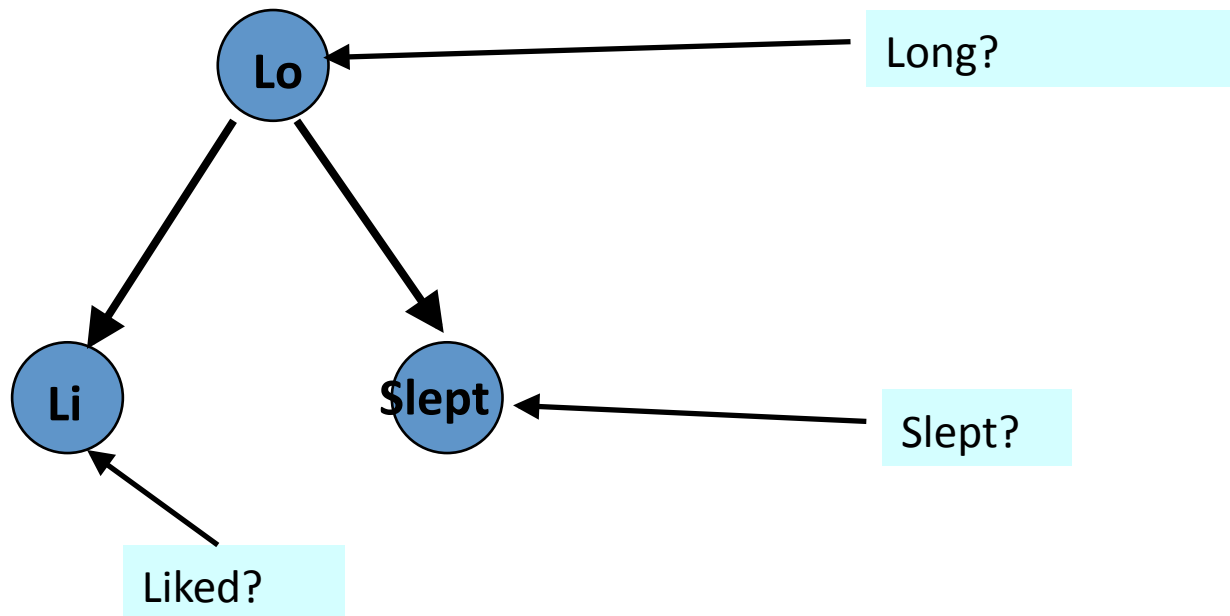
- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

$$P(X_1 \dots X_n) = \prod_i P(X_i \mid Pa(X_i))$$

$Pa(X)$ = immediate parents of X in the graph

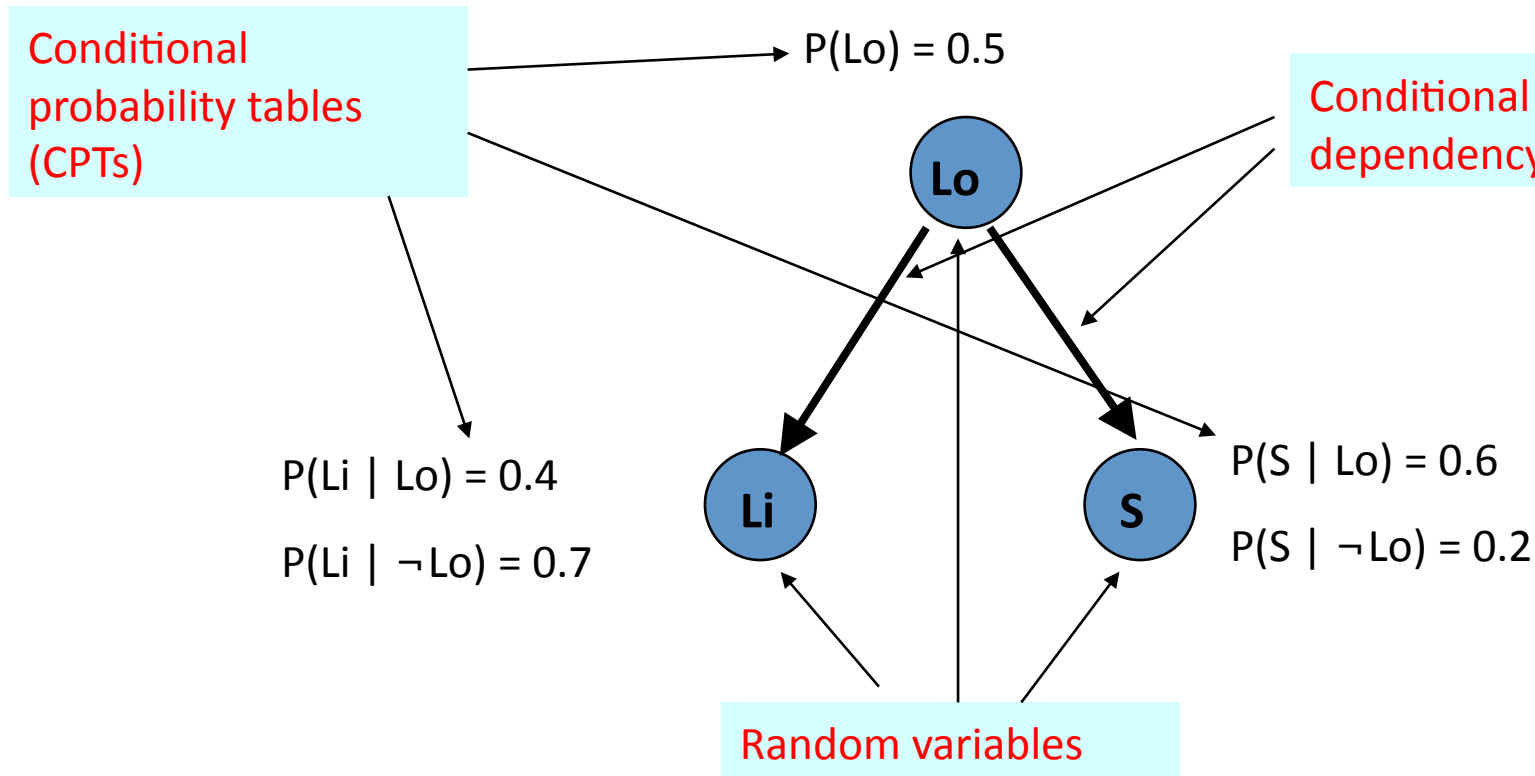
Bayesian networks

- Let's use a movie example: We would like to determine the joint probability for length, liked and slept in a movie



Bayesian networks: Notations

Bayesian networks are directed acyclic graphs.



Bayesian networks: Notations

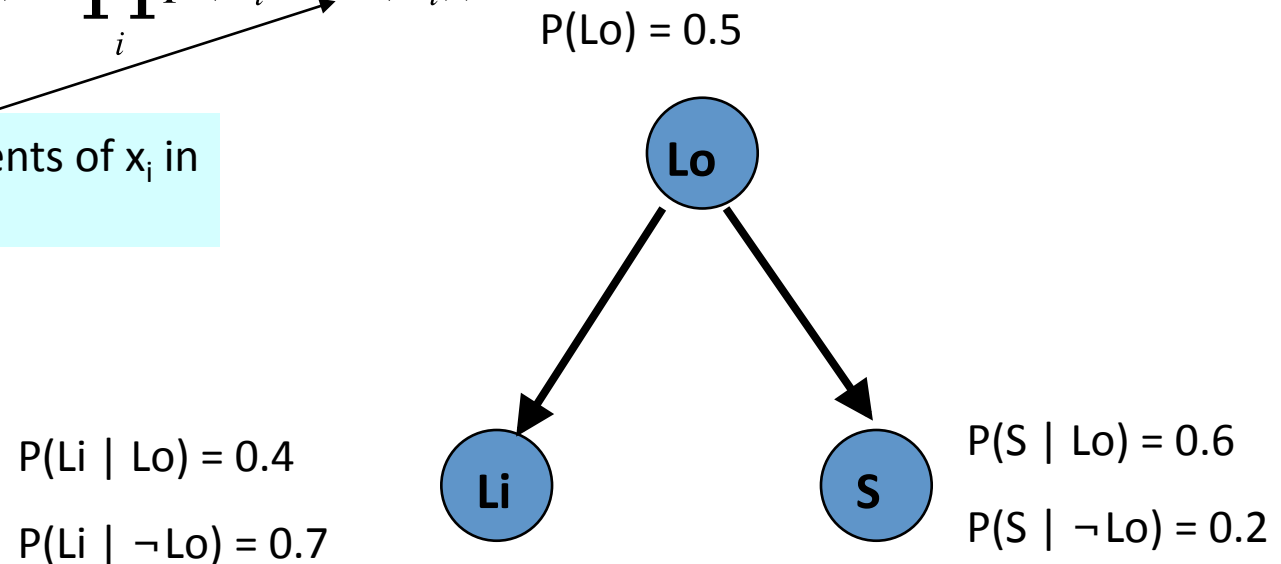
The Bayesian network below represents the following joint probability distribution:

$$p(Lo, Li, S) = P(Lo)P(Li | Lo)P(S | Lo)$$

More generally Bayesian network represent the following joint probability distribution:

$$p(x_1 \dots x_n) = \prod_i p(x_i | Pa(x_i))$$

The set of parents of x_i in the graph



Network construction and structural interpretation

Constructing a Bayesian network

- How do we go about constructing a network for a specific problem?

Step 1: Identify the random variables

Step 2: Determine the conditional dependencies

Step 3: Populate the CPTs



Can be learned from observation data!

A example problem

- An alarm system
 - B – Did a burglary occur?
 - E – Did an earthquake occur?
 - A – Did the alarm go off?
 - M – Mary calls
 - J – John calls
- How do we reconstruct the network for this problem?

Factoring joint distributions

- Using the chain rule we can always factor a joint distribution as follows:

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

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

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

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

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

- This type of conditional dependencies can also be represented graphically.

A Bayesian network

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

Number of parameters:

M: 2^4 Each parent contributes to 2 parameters

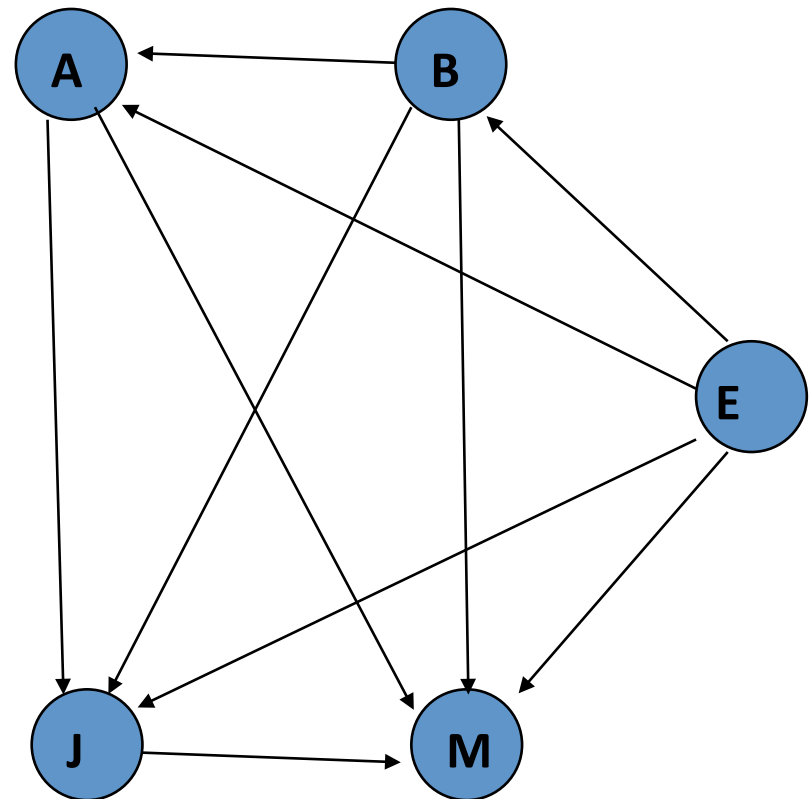
J: 2^3

A: 4

B: 2

E: 1

A total of 31 parameters



A better approach

- An alarm system
 - B – Did a burglary occur?
 - E – Did an earthquake occur?
 - A – Did the alarm go off?
 - M – Mary calls
 - J – John calls
- Let's use our knowledge of the domain!

Reconstructing a network

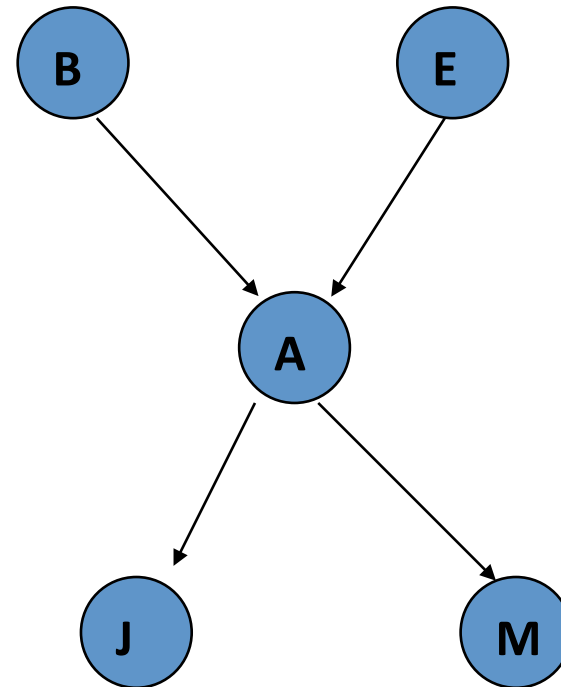
B – Did a burglary occur?

E – Did an earthquake occur?

A – Did the alarm go off?

M – Mary calls

J – John calls



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

Reconstructing a network

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

Number of parameters:

A: 4

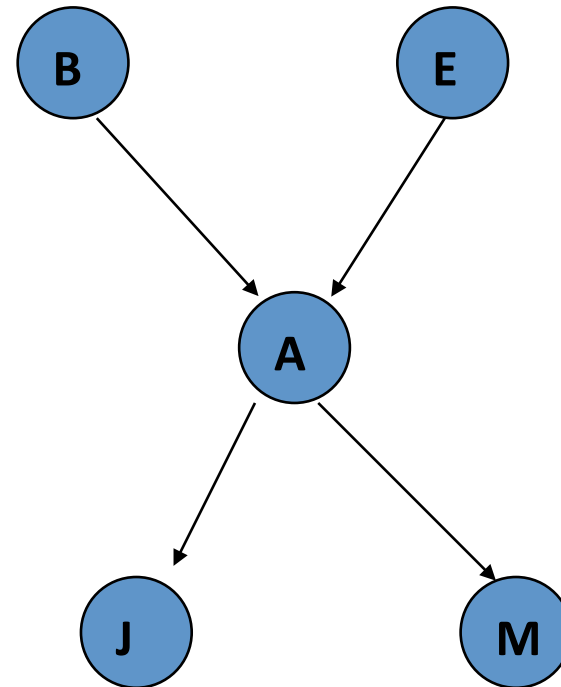
B: 1

E: 1

J: 2

M: 2

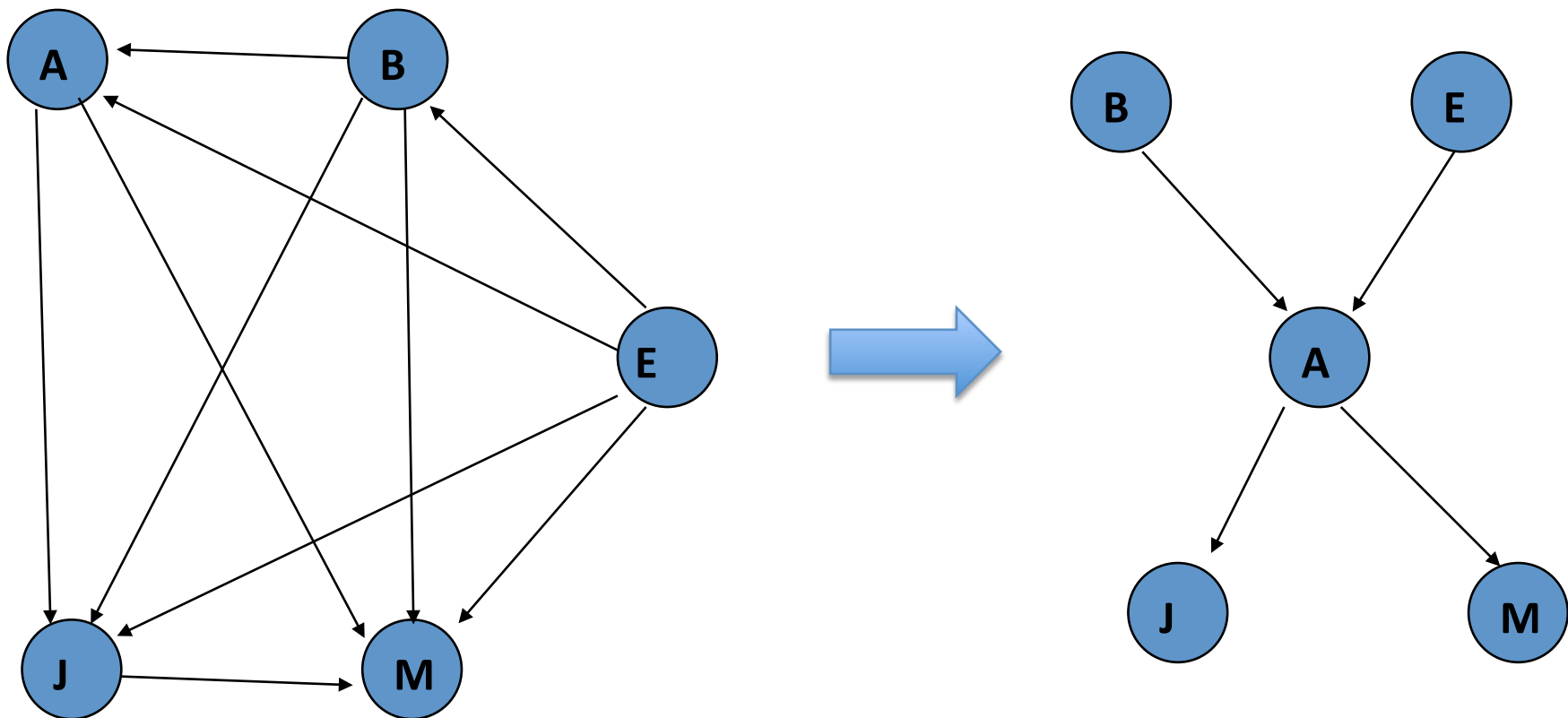
A total of 10 parameters



By relying on domain knowledge we saved 21 parameters!

A Bayesian network

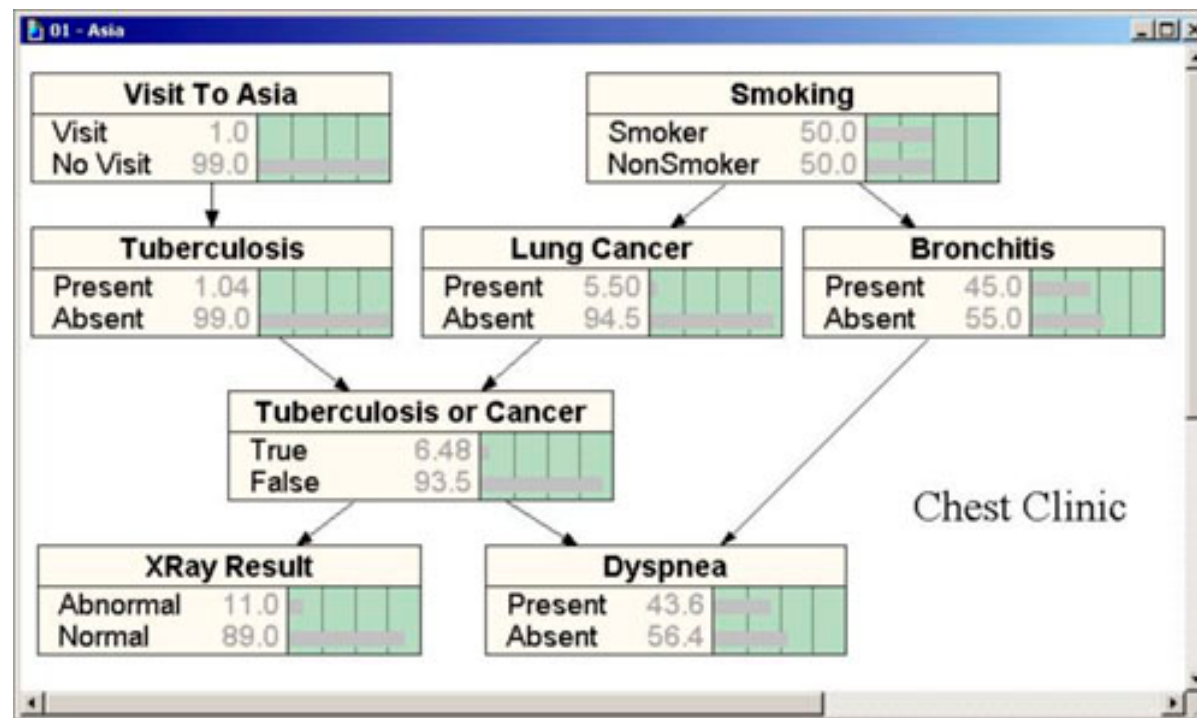
$P(M \mid J, A, B, E)$	$P(J \mid A, B, E)$	$P(A \mid B, E)$	$P(B \mid E)$	$P(E)$
$P(M A)$	$P(J A)$	$P(A B, E)$	$P(B)$	$P(E)$



Constructing a Bayesian network: Revisited

- Step 1: Identify the random variables
- Step 2: Determine the conditional dependencies
 - Select an 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
 - From examples using density estimation

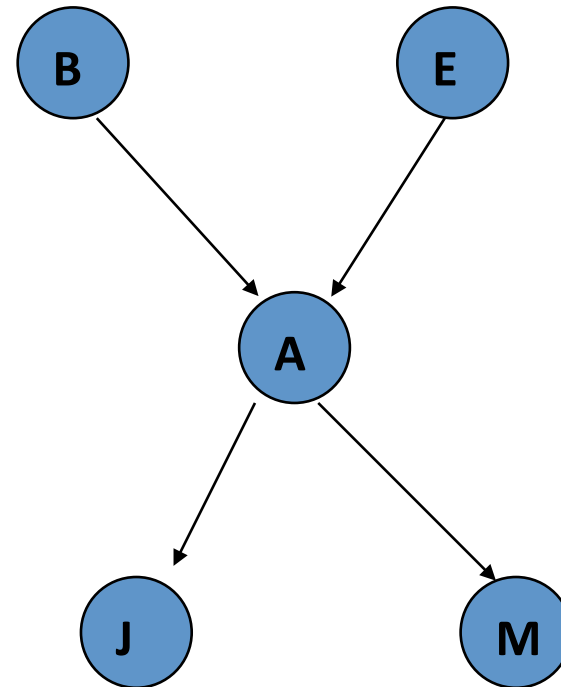
Example: Bayesian networks for cancer detection



Conditional independence

- Two variables x, y are said to be conditionally independent given a third variable z if $p(x, y | z) = p(x | z)p(y | z)$
- In a Bayesian network a variable is conditionally independent of all other variables given its **Markov blanket**

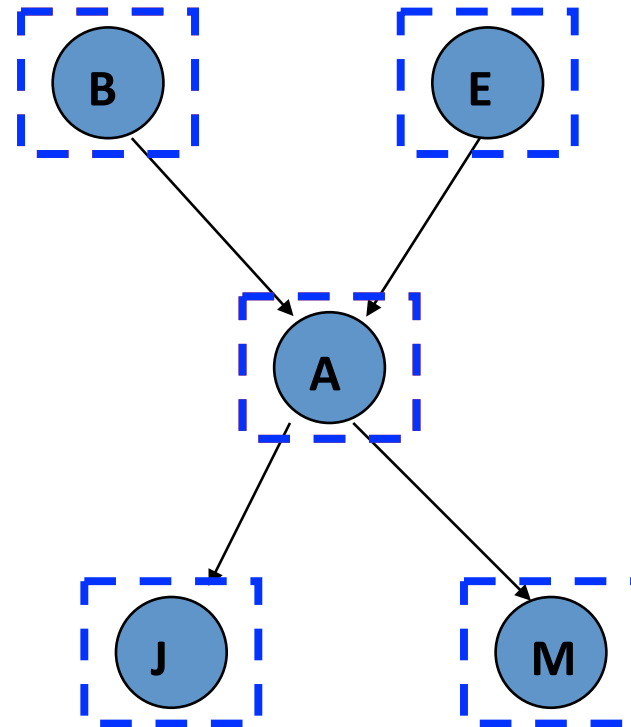
Markov blanket: All parent, children and co-parents of children



Markov blankets: Examples

Markov blanket for B:
E, A

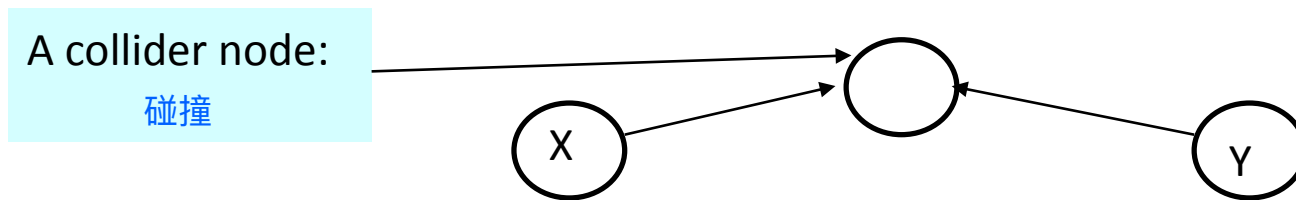
Markov blanket for A:
B, E, J, M



d-separation

- In some cases it would be useful for us to know under which conditions two variables are independent of each other
 - Helps when trying to do inference
 - Can help determine causality from structure
- Two variables x and y are d-separated given a set of variables Z (which could be empty) if x and y are conditionally independent given Z
- We denote such conditional independence as $I(x,y|Z)$

Collider Node

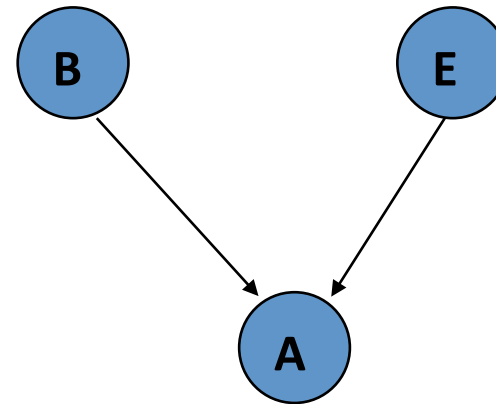


Collider Node

B – Did a burglary occur?

E – Did an earthquake occur?

A – Did the alarm go off?



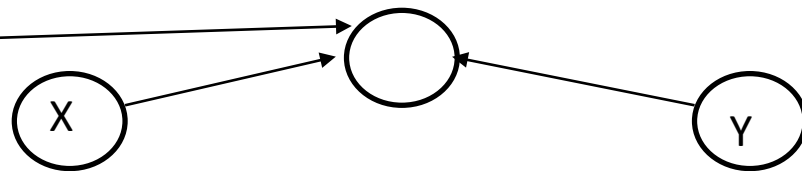
B and E are independent, before observing A

B and E become dependent, after observing A

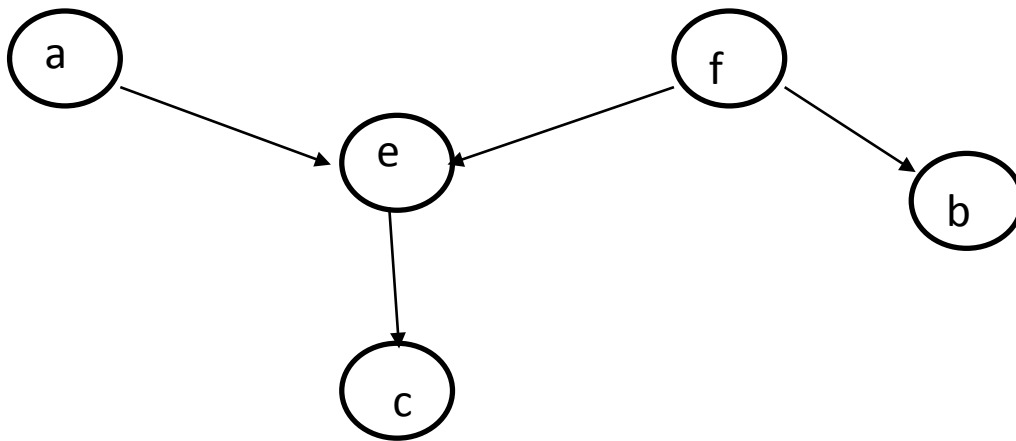
d-separation

- x and y are d-separated given Z , if all possible paths between x and y are blocked.
- A path between x and y is blocked if the path includes a node such that either
 1. It is not a collider node and is in the set Z , or
 2. It is a collider node and neither the node nor any of its descendants is in Z

A collider node:



D-separation



$a \perp b \mid c?$ False

$a \perp b \mid f?$ True

Important points

- Probabilistic graphical models as a representation of joint probability distribution
- Bayesian networks represent Joint distribution, independence, conditional independence
- Attributes of Bayesian networks
- Constructing a Bayesian network
- Markov blanket, d-separation