



A Brief Introduction to Graphical Models

Presenter: Yijuan Lu
November 12, 2004



References

- “Introduction to Graphical Models”, Kevin Murphy, Technical Report, May 2001
- “Learning in Graphical Models”, Michael I. Jordan, MIT Press, 1999.



Outline

- Application
- Definition
- Representation
- Inference
- Learning
- Conclusion



Application

- Probabilistic expert system for medical diagnosis
- Widely adopted by Microsoft
 - e.g. the Answer Wizard of Office 95
 - the Office Assistant of Office 97
 - over 30 technical support troubleshooters
- Vista system used at NASA Mission Control Center to interpret live telemetry



Application

- Machine Learning
- Statistics
- Speech Recognition
- Natural Language Processing
- Computer Vision
- Image Processing
- Bio-informatics

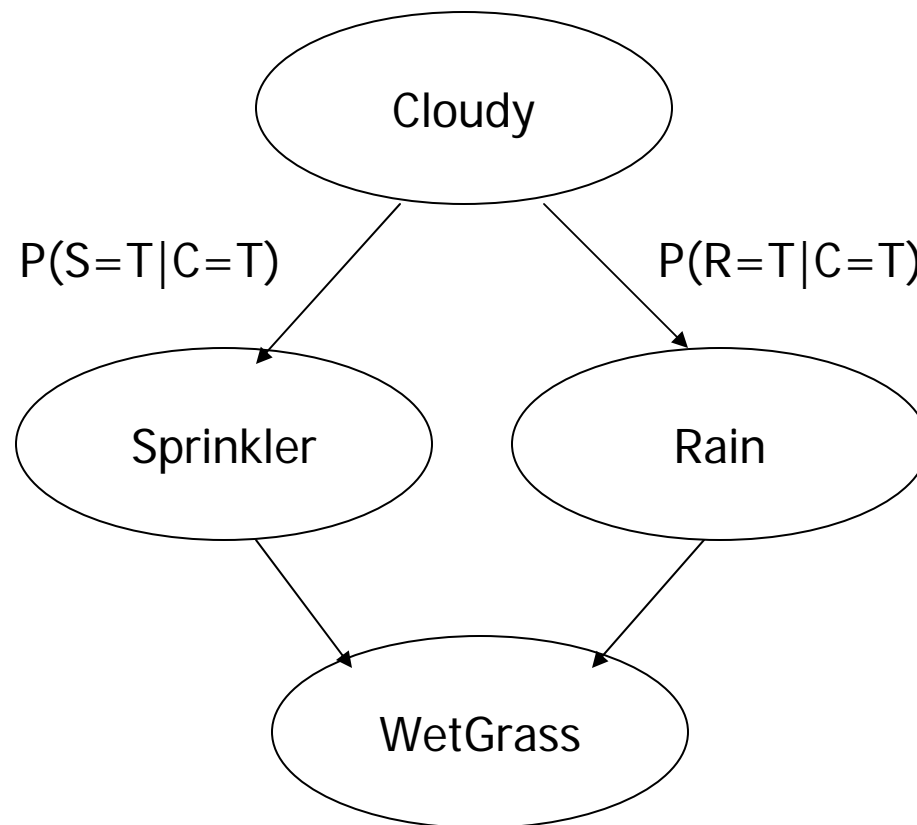
.....



What causes grass wet?

- Mr. Holmes leaves his house:
 - the grass is wet in front of his house.
 - several reasons are possible: either it rained or the sprinkler of Holmes has been on during the night.
- Then, Mr. Holmes looks at the sky and find it is cloudy:
 - Since when it is cloudy, usually it rains and the sprinkler is off.
 - He concludes it is more likely that rain causes grass wet.

What causes grass wet?

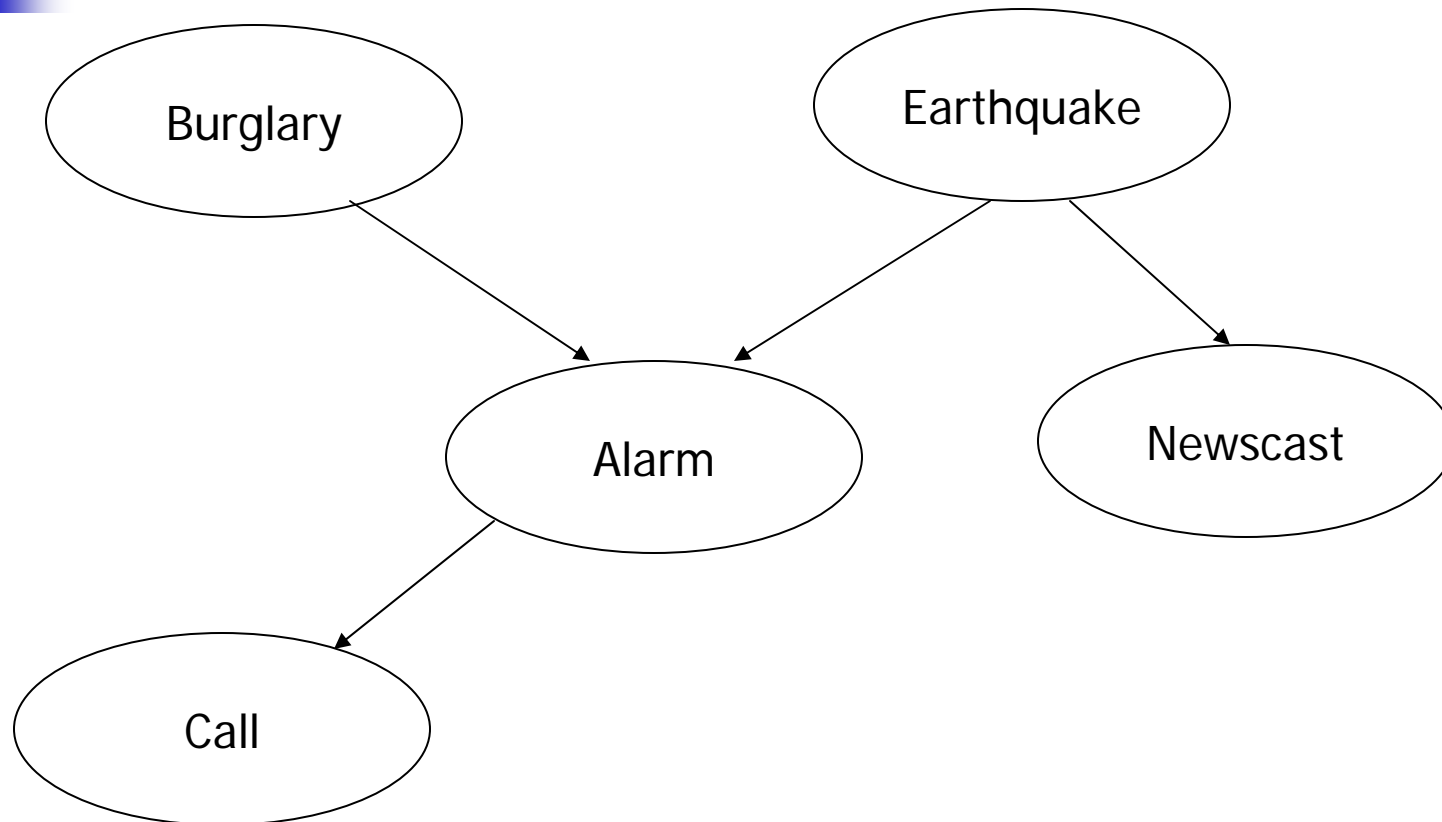




Earthquake or burglary?

- Mr. Holmes is in his office
 - He receives a call from his neighbor that the alarm of his house went off.
 - He thinks that somebody broke into his house.
- Afterwards he hears an announcement from radio that a small earthquake just happened
 - Since the alarm has been going off during an earthquake.
 - He concludes it is more likely that earthquake causes the alarm.

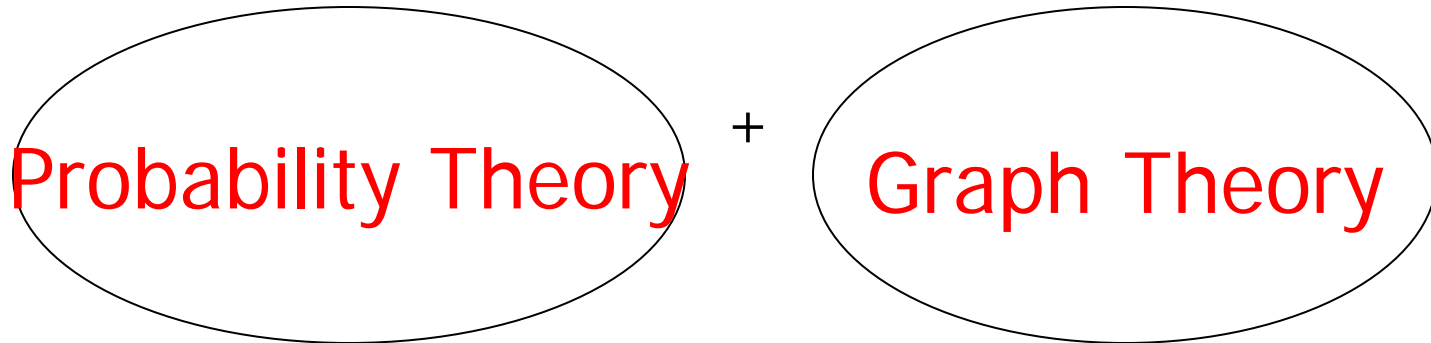
Earthquake or burglary?





Graphical Models

- Graphical Model:



- Provides a natural tool for two problems:
Uncertainty and **Complexity**
- Plays an important role in the design and analysis of machine learning algorithms



Graphical Model

- **Modularity**: a complex system is built by combining simpler parts.
- **Probability theory**: ensures consistency, provides interface models to data.
- **Graph theory**: intuitively appealing interface, efficient general purpose algorithms.

Representation

Graphical representation of probabilistic relationship between a set of random variables.

Variables are represented by nodes.

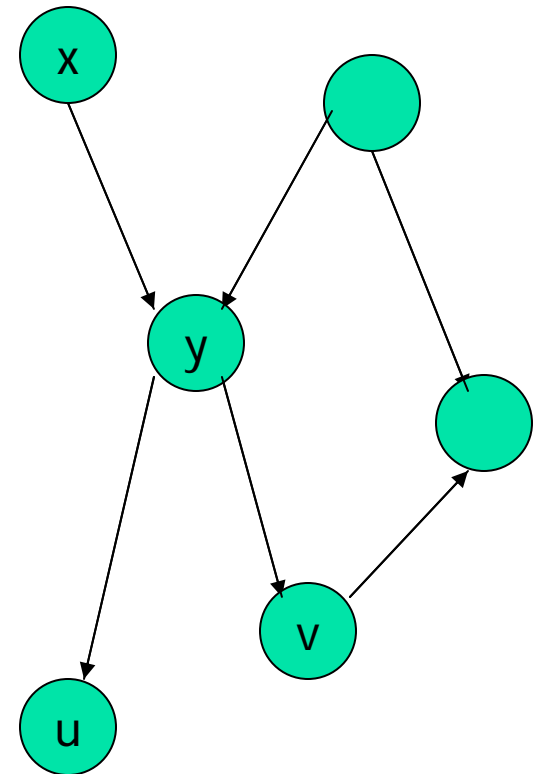
- Binary events
- Discrete variables
- Continuous variables

Conditional (in)dependency is represented by (missing) edges.

Directed Graphical Model: (Bayesian network)

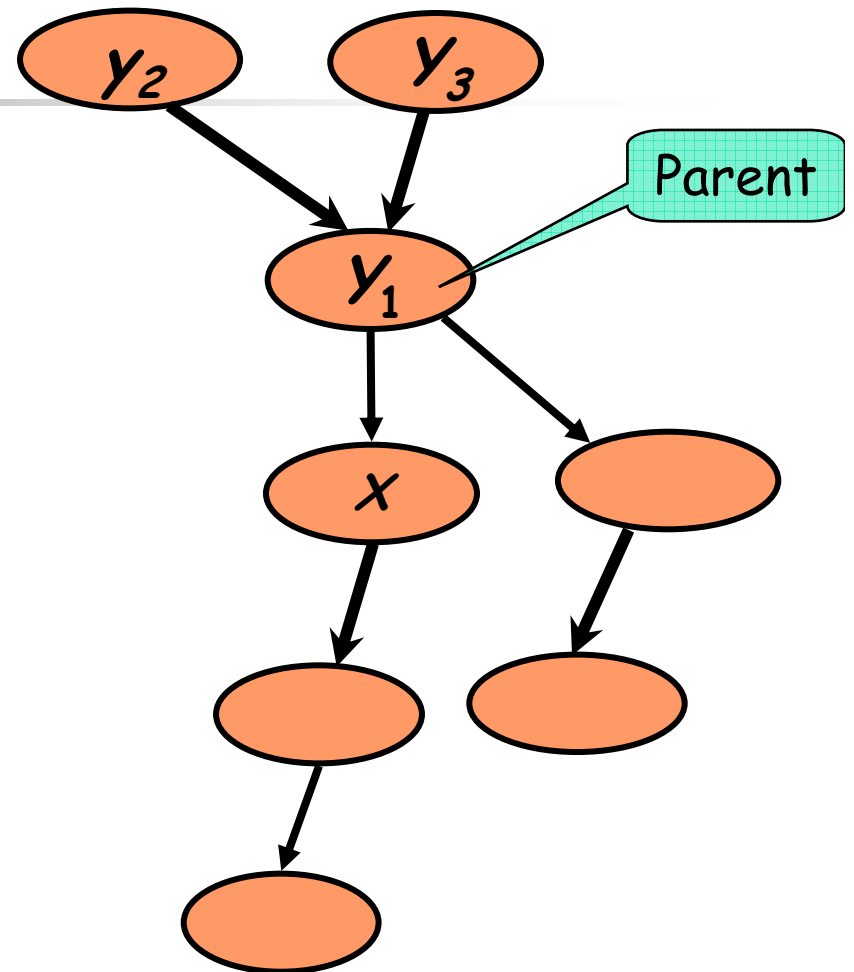
Undirected Graphical Model: (Markov Random Field)

Combined: chain graph



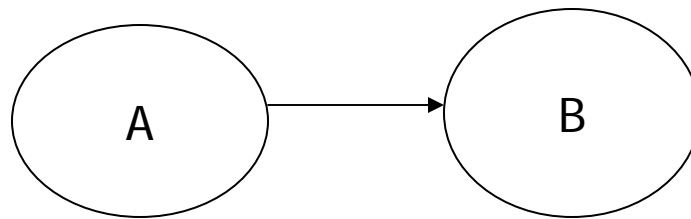
Bayesian Network

- Directed acyclic graphs (DAG).
- Directed edge means causal dependencies.
- For each variable X and parents $pa(X)$ exists a conditional probability
-- $P(X|pa(X))$
- Joint distribution





Simple Case



- That means: the value of B depends on A
- Dependency is described by the **conditional probability** $P(B|A)$
- Suppose already know the joint probability of the A and B : $P(A,B)$



Simple Case

- From the joint probability, we can derive all other probabilities:

- Marginalization: (sum rule)

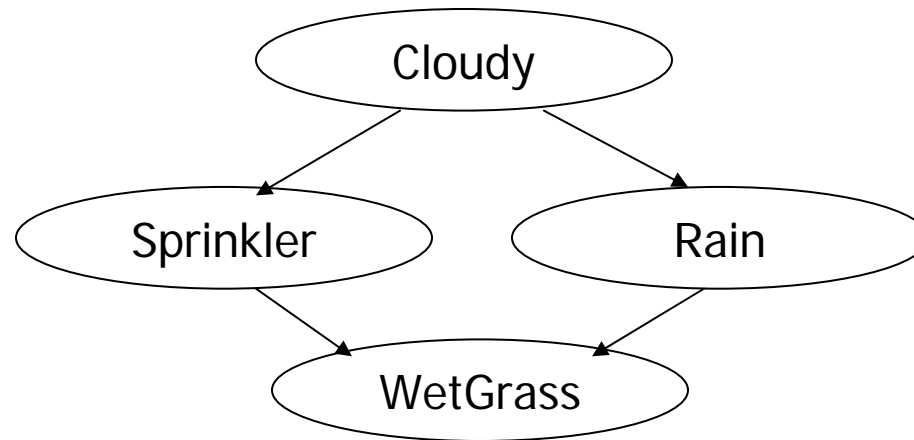
$$P(A) = \sum_B P(A, B) \quad P(B) = \sum_A P(A, B)$$

- Conditional probabilities: (Bayesian Rule)

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



Simple Example



$$P(W = T) = \sum_{c,s,r} P(C = c, S = s, R = r, W = T)$$

$$\begin{aligned} P(S = T | W = T) &= \frac{P(S = T, W = T)}{P(W = T)} \\ &= \frac{\sum_{c,r} P(C = c, S = T, R = r, W = T)}{P(W = T)} \end{aligned}$$



Bayesian Network

- Variables: $U = \{X_1, X_2, \dots, X_n\}$
- The joint probability of $P(U)$ is given by
$$P(U) = P(X_1 \dots X_n) = P(X_1)P(X_2 | X_1) \dots P(X_n | X_{n-1}, \dots, X_1)$$
- If the variables are binary,
we need $O(2^n)$ parameters to describe P
- Can we do better?
- Key idea: use properties of independence.



Independent Random Variables

- X is independent of Y iif
$$P(X=x|Y=y) = P(X=x) \text{ for all values } x,y$$
- If X and Y are independent then
$$P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$$
- Unfortunately, most of random variables of interest are not independent of each other



Conditional Independence

- A more suitable notion is that of conditional independence.
- X and Y are conditional independent given Z iff $P(X=x|Y=y,Z=z)=P(X=x|Z=z)$ for all values x,y,z
- notion: $I(X,Y|Z)$
- $P(X,Y,Z)=P(X|Y,Z)P(Y|Z)P(Z)=P(X|Z)P(Y|Z)P(Z)$

Bayesian Network

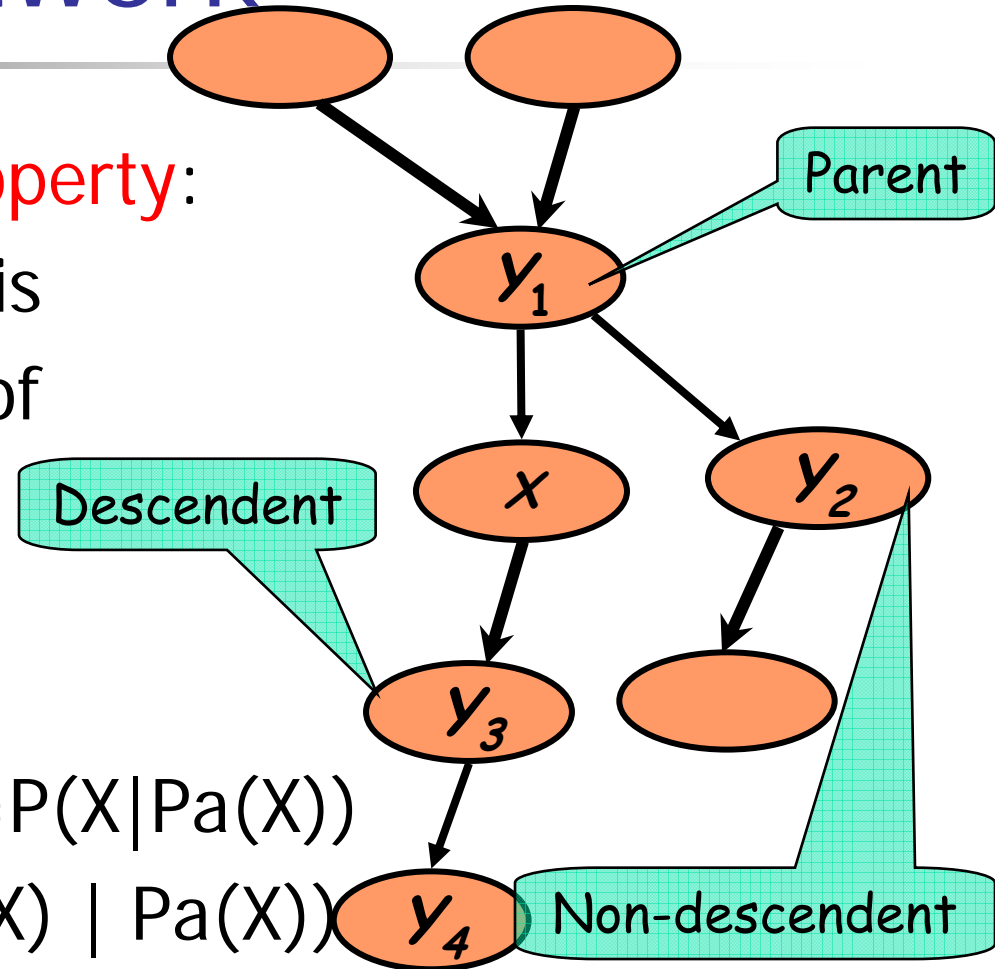
- **Directed local Markov Property:**

Each random variable X , is conditional independent of its non-descendants, given its parents $\text{Pa}(X)$

- Formally,

$$P(X | \text{NonDesc}(X), \text{Pa}(X)) = P(X | \text{Pa}(X))$$

- Notation: $I(X, \text{NonDesc}(X) | \text{Pa}(X))$





Bayesian Network

- Factored representation of joint probability
 - Variables: $U = \{ X_1, X_2, \dots, X_n \}$
 - The joint probability of $P(U)$ is given by

$$P(U) = P(X_1 \dots X_n) = P(X_1)P(X_2 | X_1) \dots P(X_n | X_{n-1}, \dots, X_1)$$

$$= \prod_{i=1}^n P(X_i | pa(X_i))$$

- the joint probability is product of all conditional probabilities



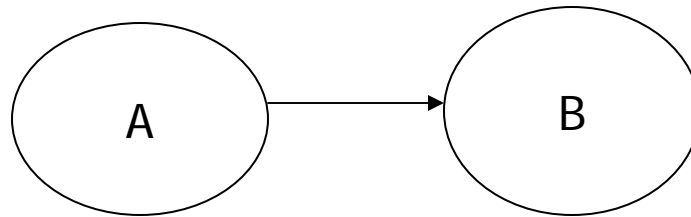
Bayesian Network

- Complexity reduction
 - Joint probability of n binary variables
 $O(2^n)$
 - Factorized form
 $O(n * 2^k)$
- K : maximal number of parents of a node



Simple Case

$$(P(U) = \prod_{i=1}^n P(X_i | pa(X_i)))$$



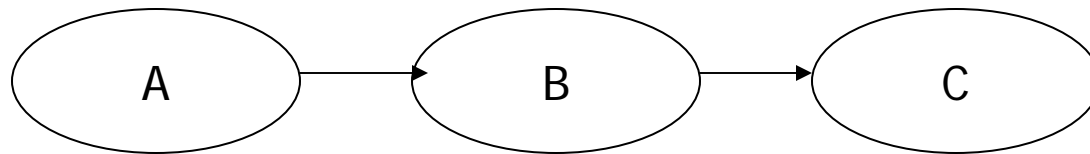
- Dependency is described by the **conditional probability** $P(B|A)$
- Knowledge about A: priori probability $P(A)$
- Calculate the joint probability of the A and B

$$P(A,B) = P(B|A)P(A)$$



Serial Connection

$$(P(U) = \prod_{i=1}^n P(X_i | pa(X_i)))$$



- Calculate as before:
 - $P(A, B) = P(B|A)P(A)$
 - $P(A, B, C) = P(C|A, B)P(A, B)$
 $= P(C|B)P(B|A)P(A)$
- $I(A, C|B)$.

Converging Connection

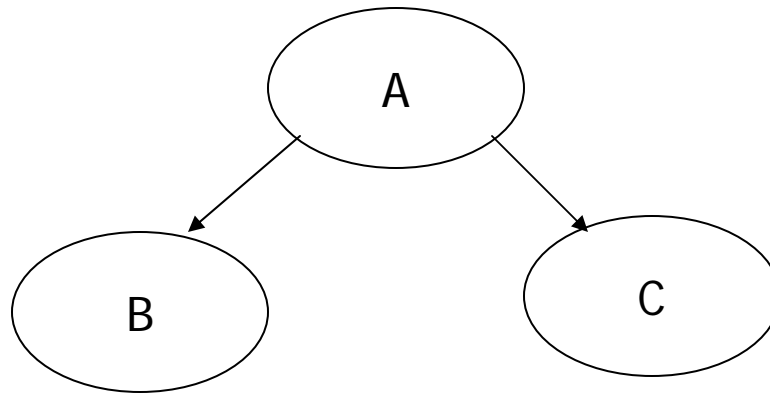
$$(P(U) = \prod_{i=1}^n P(X_i | pa(X_i)))$$



- Value of A depends on B and C:
 $P(A|B,C)$
- $P(A,B,C) = P(A|B,C)P(B)P(C)$

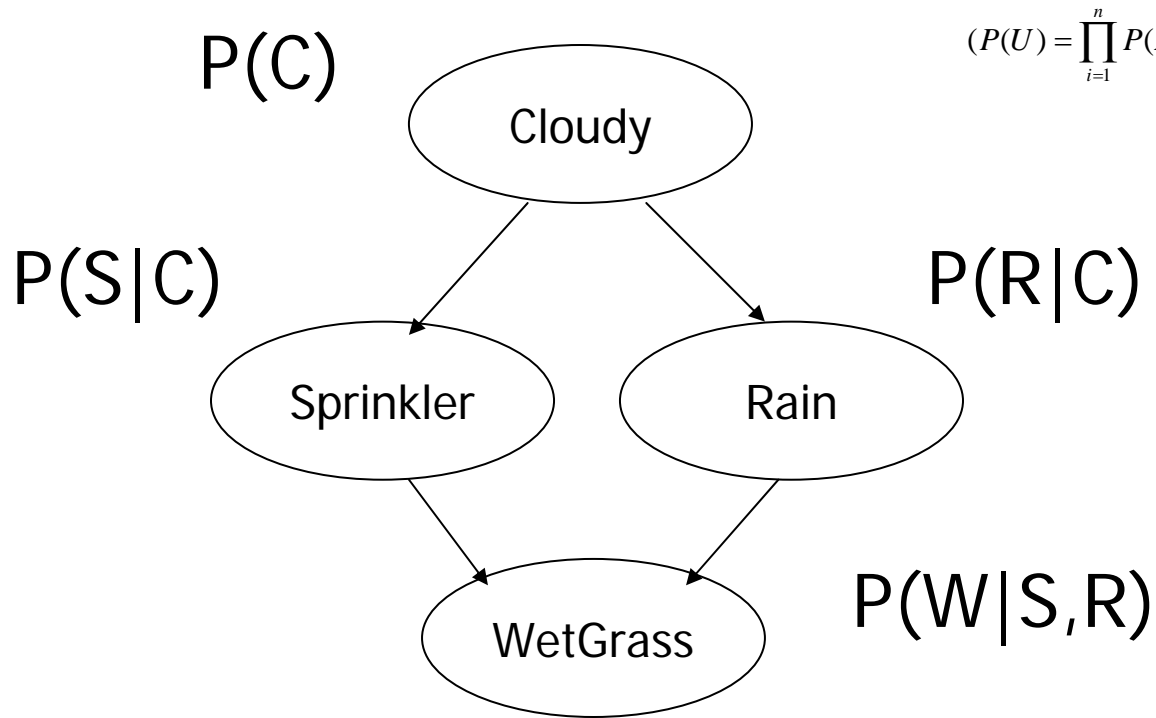
Diverging Connection

$$(P(U) = \prod_{i=1}^n P(X_i | pa(X_i)))$$



- B and C depend on A: $P(B|A)$ and $P(C|A)$
- $P(A,B,C) = P(B|A)P(C|A)P(A)$
- $I(B,C|A)$

Wetgrass



$$(P(U) = \prod_{i=1}^n P(X_i | pa(X_i)))$$

$$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$$

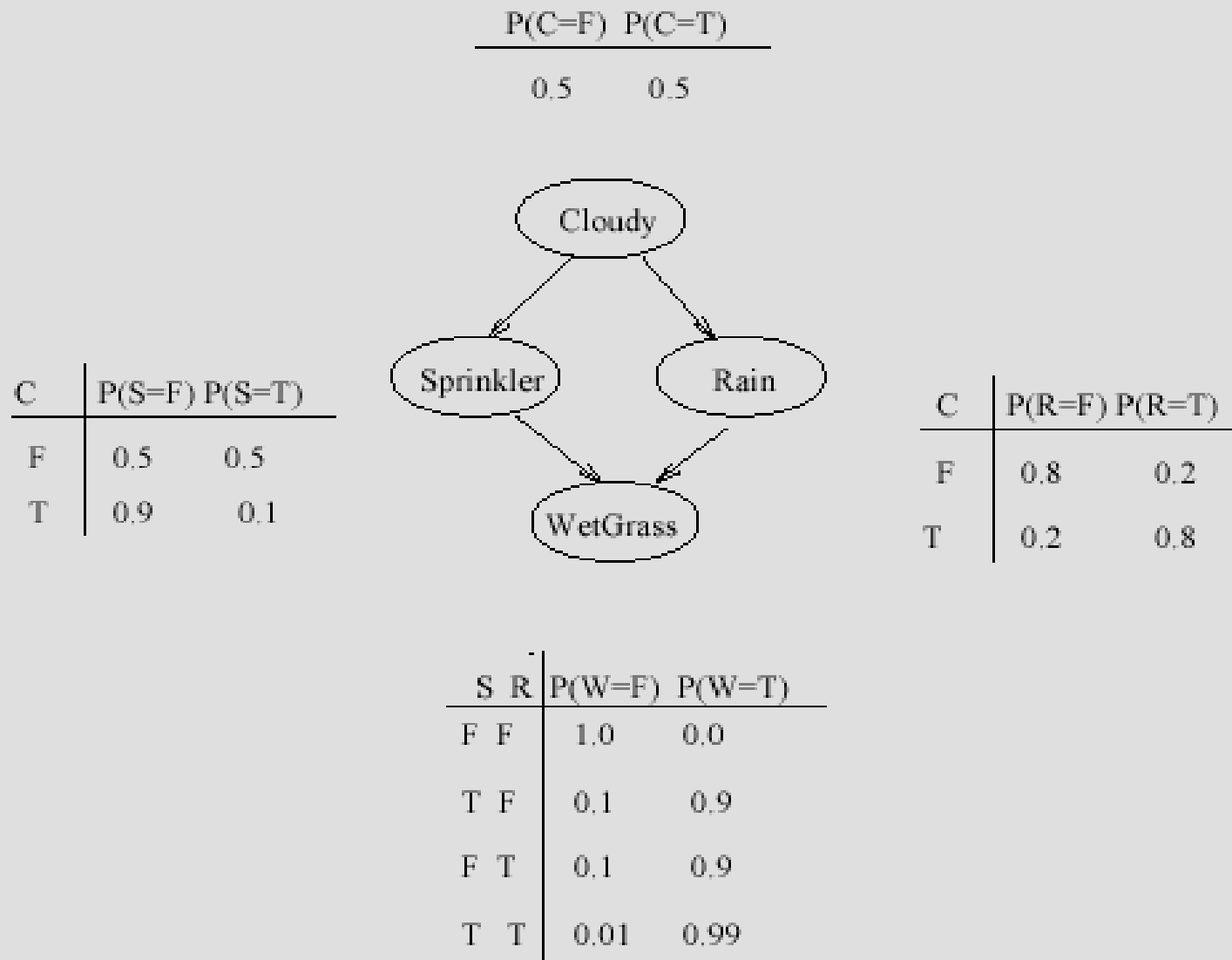


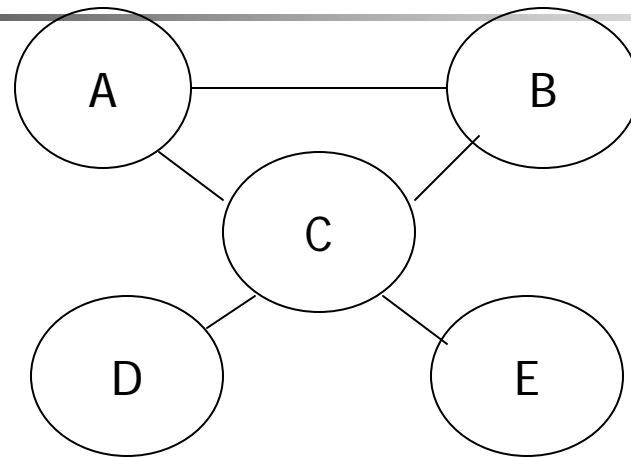
Figure 1: A simple Bayesian network, adapted from [RN95].



Markov Random Fields

- Links represent symmetrical probabilistic dependencies
- Direct link between A and B: conditional dependency.
- Weakness of MRF: inability to represent induced dependencies.

Markov Random Fields



- Global Markov property: x is independent of Y given Z iff all paths between X and Y are blocked by Z .
- (here: A is independent of E , given C)
- Local Markov property: X is independent of all other nodes given its neighbors.
- (here: A is independent of D and E , given C and B)



Inference

- Computation of the conditional probability distribution of one set of nodes, given a model and another set of nodes.
- Any conditional probability can be calculated from the joint probability distribution of the variables.
- Bottom-up
 - observation of an effect: e.g. wet grass
 - The probabilities of the reasons (rain, sprinkler) can be calculated accordingly
 - “diagnosis” from effects to reasons
- Top-down
 - Knowledge, that it is cloudy influences the probability that the grass is wet
 - Predict the effects



Inference

Observe: wet grass (denoted by $W=1$)

- Two possible causes: rain or sprinkler
Which is more likely?
- Using Bayes' rule to compute the posterior probabilities of the reasons (rain, sprinkler)

$$P(X|y) = \frac{P(y|X)P(X)}{P(y)} \quad \text{posterior} = \frac{\text{conditional likelihood} \times \text{prior}}{\text{likelihood}}$$



Inference

In the current example, we have

$$P(S = 1|W = 1) = \frac{P(S = 1, W = 1)}{P(W = 1)} = \frac{\sum_{c,r} P(C = c, S = 1, R = r, W = 1)}{P(W = 1)} = \frac{0.2781}{0.6471} = 0.430$$

and

$$P(R = 1|W = 1) = \frac{P(R = 1, W = 1)}{P(W = 1)} = \frac{\sum_{c,s} P(C = c, S = s, R = 1, W = 1)}{P(W = 1)} = \frac{0.4581}{0.6471} = 0.708$$

where

$$P(W = 1) = \sum_{c,s,r} P(C = c, S = s, R = r, W = 1) = 0.6471$$

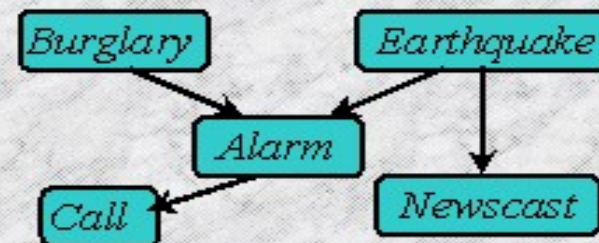
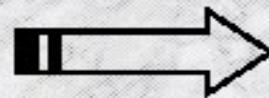


Inference algorithms and methods

- Junction tree algorithm
- Variable elimination
- Pearl's algorithm: generalization of forward-backward algorithm from HMMs
- Monte Carlo methods
- Variational methods
- Loopy belief propagation

Learning

The learning task



Input: training data

Output: BN modeling data

- Input: fully or partially observable data cases?
- Output: parameters or also structure?



Learning

- Learn parameters or structure from data
- Parameter learning: find maximum likelihood estimates of parameters of each conditional probability distribution
- Structure learning: find correct connectivity between existing nodes



Learning

Structure	Observation	Method
Known	Full	ML or MAP estimation
Known	Partial	EM algorithm
Unknown	Full	Model selection
Unknown	Partial	EM + model selection



Conclusion

- A graphical representation of the probabilistic structure of a set of random variables, along with functions that can be used to derive the joint probability distribution.
- Intuitive interface for modeling.
- Modular: Useful tool for managing complexity
- Common formalism for many models
 - Facilitates transfer of ideas between communities



EM Algorithm

Expectation (E) step

Use current parameters to estimate the unobserved data

Maximization (M) step

Use estimated filled in data to do ML/MAP estimation of the parameter

Set the parameter as the estimated value

repeat EM steps, until convergence

In the E step the expected values of all the nodes are computed using an inference algorithm. These expected values are treated as though they were observed (distributions), in the subsequent M step



Model Selection Method

- Select a 'good' model from among all possible models and use it as if it were the correct model
- A 'good' model satisfies some criterion used to measure the degree to which a network structure fits the prior knowledge and data
- A search algorithm is then used to find a network structure that receives a high score by the chosen criterion