DSCI 552, Machine Learning for Data Science

University of Southern California

M. R. Rajati, PhD

Lesson 13
Graphical Models

http://www.computervisionblog.com/2015/04/deep-learning-vs-probabilistic.html

Graphical Models

- Graphical models: children of graph theory and probability theory
- Connect neural networks and models such as HMMs, MRFs, and Kalman Filters

Advantages of Graphical Models

- Handling inference and learning in a unified manner
- Providing a unified framework for supervised and unsupervised learning
- Handling missing data easily
- Modeling conditional independence
  Transparency and Explainability (if desired)

Graphs

A graph consists of a collection of nodes and edges.

– Nodes, or vertices, are usually associated with the variables distinction between discrete and continuous ignored in this initial discussion

– Edges connect nodes to one another.

Types of Graphical Models

- Two types:
  - undirected graphical models
  - and directed graphical models.
- Main focus: directed graphical models.

(Specific forms of) Graphical models are also known as:

- Belief networks,
- Bayesian networks,
- Markov random Fields (MRFs)

Learning and Inference

- Key concept of graphical models
    - What can be inferred should not be learned
- Weights make local assertions about the relationships between neighboring nodes

Learning and Inference

- Inference algorithms turn local assertions into global assertions about the relationships between nodes.
- Examples:
  - correlations between hidden units conditioned on a certain input and its corresponding output
  - the probability of an input vector given an output vector
- This is achieved by calculating joint probability distribution from the network

Preliminaries

A specific form of graphical model are Bayesian networks:

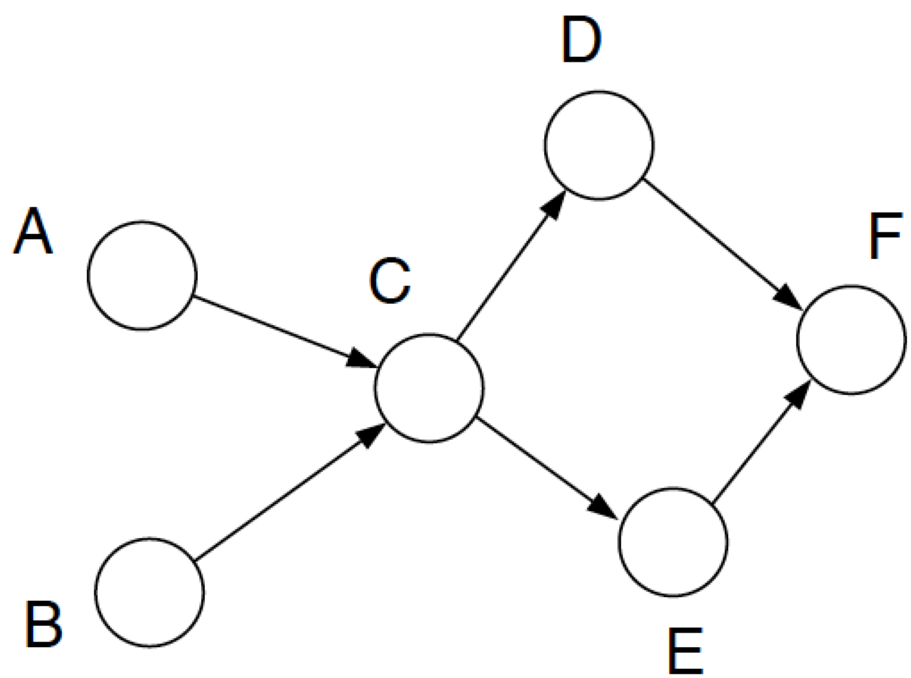
– directed acyclic graphs (DAGs)

– directed: all connections have arrows associated with them;

– acyclic: following the arrows around it is not possible to complete a loop

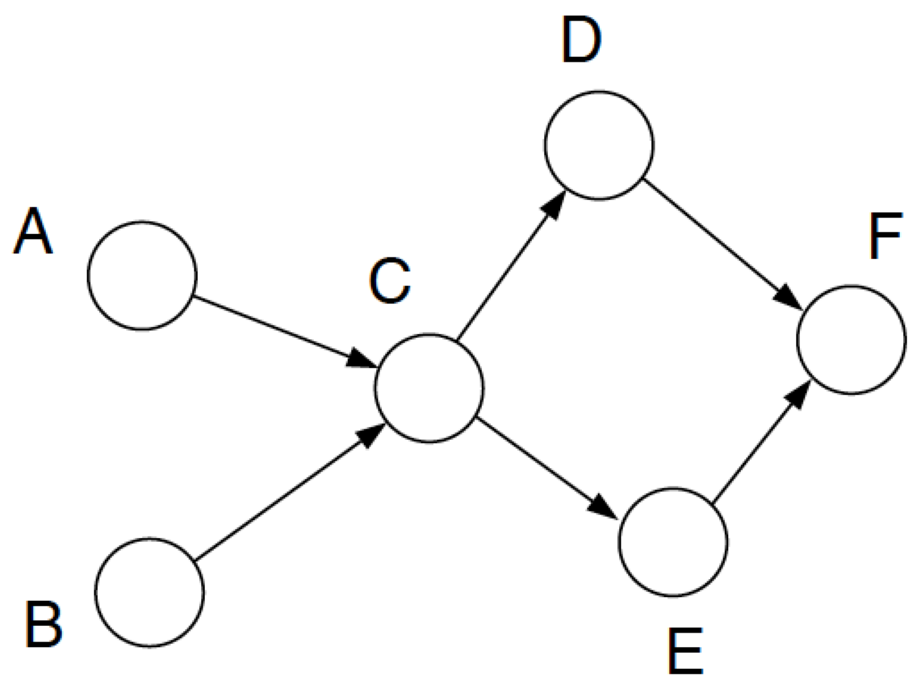
Preliminaries

- Consider an arbitrary directed (acyclic) graph, where each node in the graph corresponds to a random variable (scalar or vector):

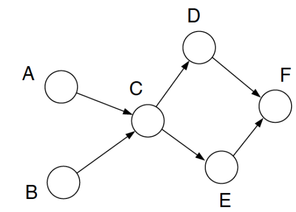


Preliminaries

- Edges represent statistical dependencies between the variables



Preliminaries



- No need to designate units as inputs, outputs or hidden
- We associate a probability distribution *P* (*A*, *B*, *C*, *D*, *E*, *F*) with this graph
- All of other calculations are consistent with this distribution
  - *Short hand notation for*
  - *P* (*A=a*, *B=b*, *C=c*, *D=d*, *E=e*, *F=f*)

Preliminaries

- Example:

$$P(E=e|C=c,D=d)=\frac{P(C=c,D=d,E=e)}{P(C=c,D=d)}$$

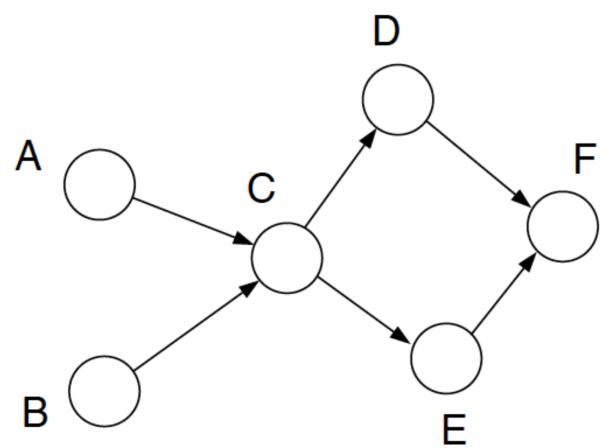
$$=\frac{\sum_a \sum_b \sum_f P(A=a,B=b,C=c,D=d,E=e,F=f)}{\sum_a \sum_b \sum_e \sum_f P(A=a,B=b,C=c,D=d,E=e,F=f)}$$

- We marginalize over a variable by wading it out via summing on all of its possible values

Marginals: simplified notation

- Example:

$$P(E|C,D) = \frac{P(C,D,E)}{P(C,D)} = \frac{\sum_a \sum_b \sum_f P(A,B,C,D,E)}{\sum_A \sum_B \sum_E \sum_F P(A,B,C,D,E)}$$



Problems in GMs

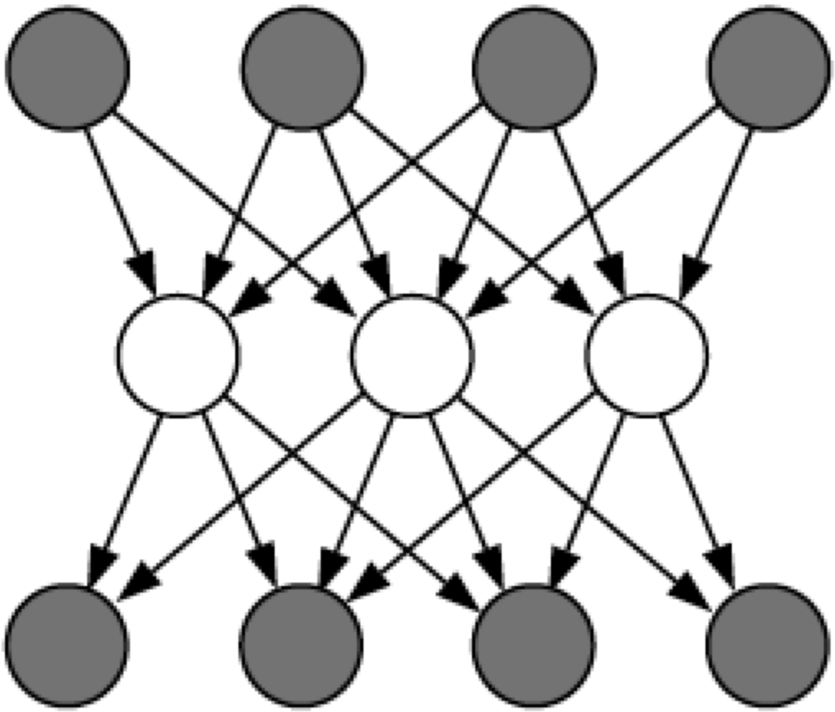
- The main problems that need to be addressed are:
  - inference (from observation it's cloudy infer probability of wet grass).
  - training the models;
  - determining the structure of the network (i.e. what is connected to what)

Notation

- In general the variables (nodes) may be split into two groups:
  - observed (shaded) variables are the ones we have knowledge about.
  - unobserved (unshaded) variables are ones we don't know about and therefore have to infer the probability.

Using Graphical Models

- Supervised Learning



Input

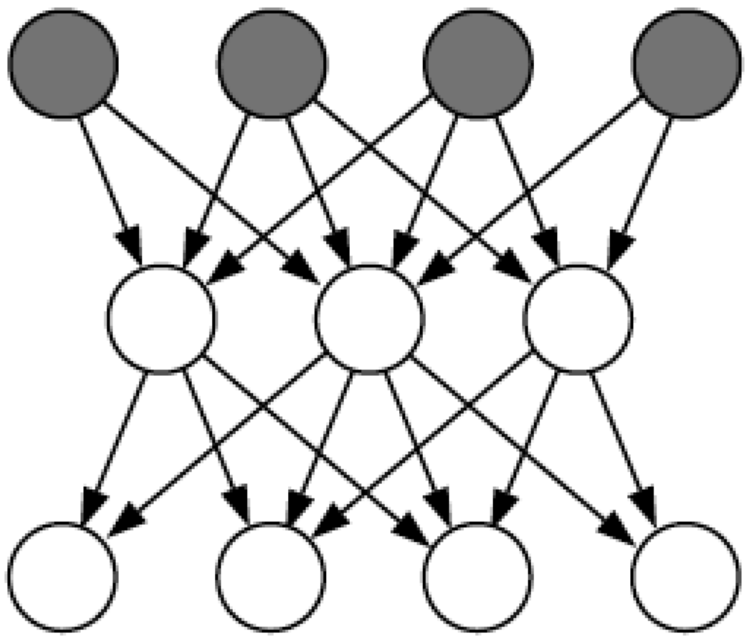
Model

Output

- Wade unshaded units out

Using Graphical Models

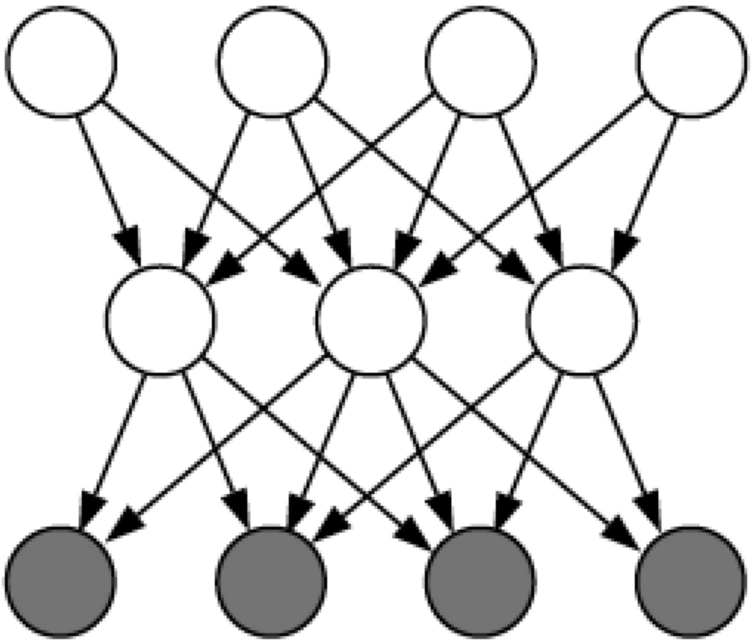
- Prediction



- Wade unshaded units out

Using Graphical Models

- Control and Optimization



- Wade unshaded units out

Building A Graphical Model

- There are two ways to build a graphical model:
  - Quantitative
  - Qualitative

Building GMs Qualitatively

- What is used to qualitatively build GMs?
  - prior knowledge of causal relationships
  - assessment from experts
  - learning from data
  - Application-specific architectures are preferred (e.g., layered graphs)

Conditional Independence

- A fundamental concept in graphical models is conditional independence.
- Consider three random variables, A, B and C:

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

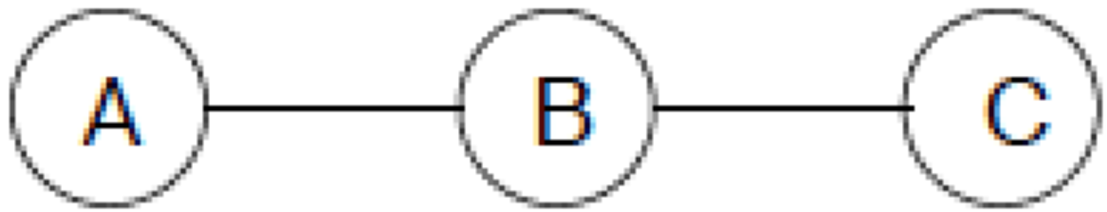
- If C is conditionally independent of A given B, then we can write

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

  - The value of A does not affect the distribution of C if B is known.

Conditional Independence: Graph

- Graphically this can be described as



- Conditional independence is important when modelling highly complex systems.

Building GMs Qualitatively: Structures

- C not observed:

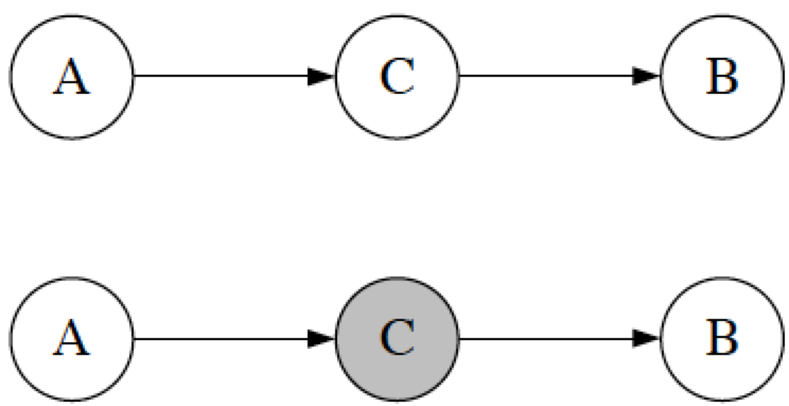
$$P(A,B) = \Sigma_C P(A,B,C) = P(A)\Sigma_C P(C|A)P(B|C)$$

then A and B are dependent on each other.

- C = c observed:

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

A and B are then independent. The path is sometimes called blocked.



Building GMs Qualitatively: Structures

- C not observed:

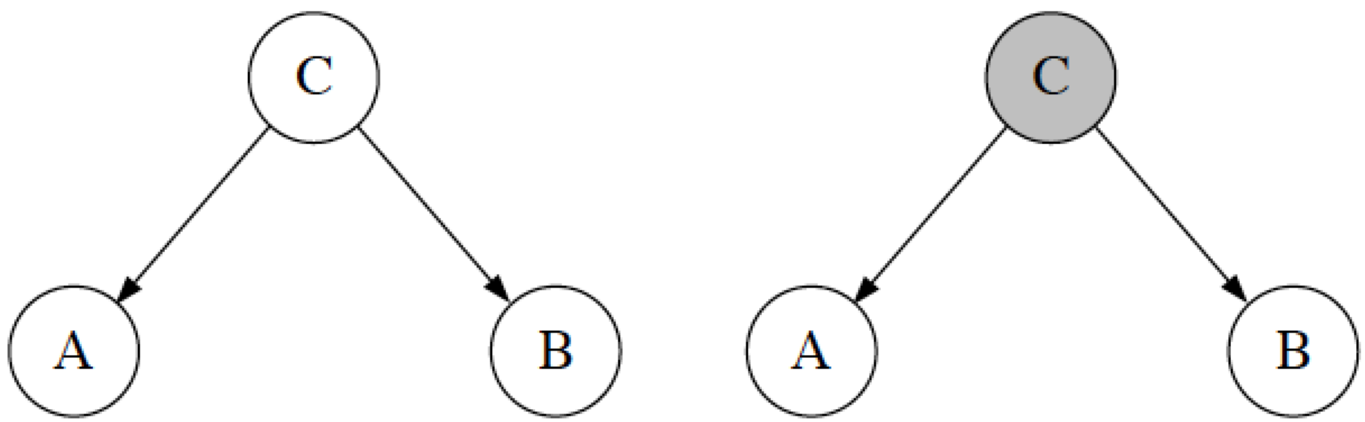
$$P(A,B) = \Sigma_C P(A,B,C) = \Sigma_C P(C)P(A|C)P(B|C)$$

then A and B are dependent on each other.

- C = c observed:

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

A and B are then independent.



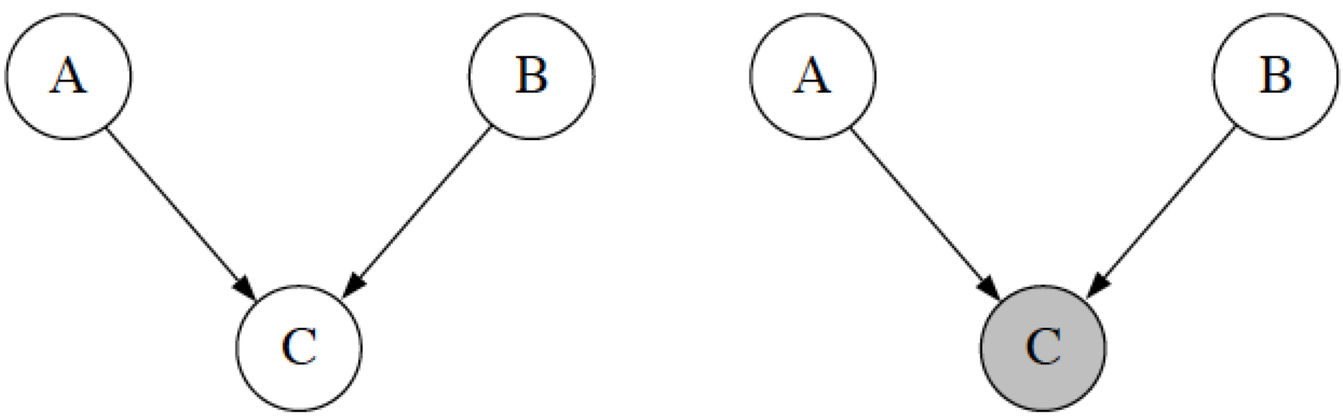
Building GMs Qualitatively

C not observed:

$P(A,B) = \Sigma_C P(A,B,C)$

$= P(A)P(B)\Sigma_C P(C|A,B) = P(A)P(B)$

A and B are independent of each other.



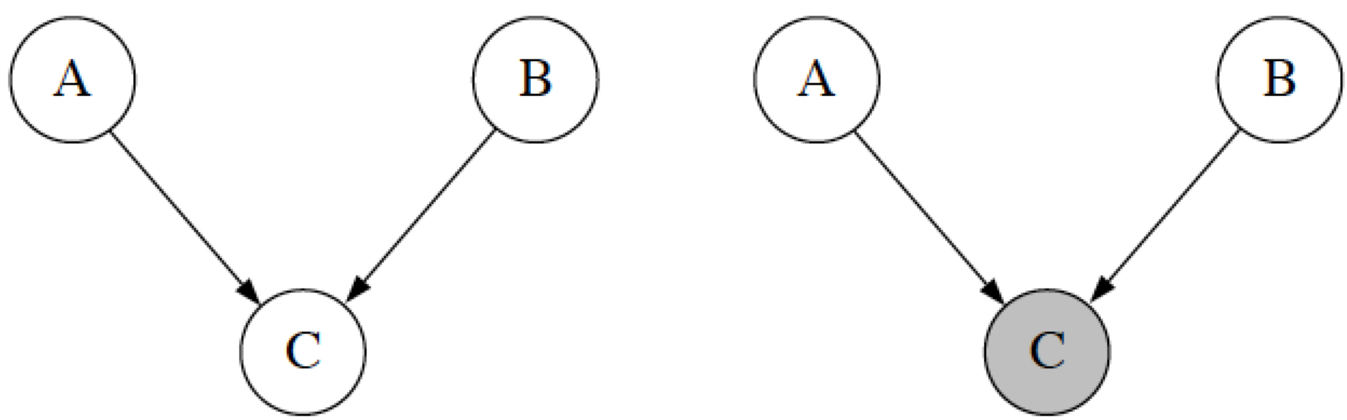
Building GMs Qualitatively

C=c observed:

$P(A,B|C=c) = P(A,B,C=c)/P(C=c)$

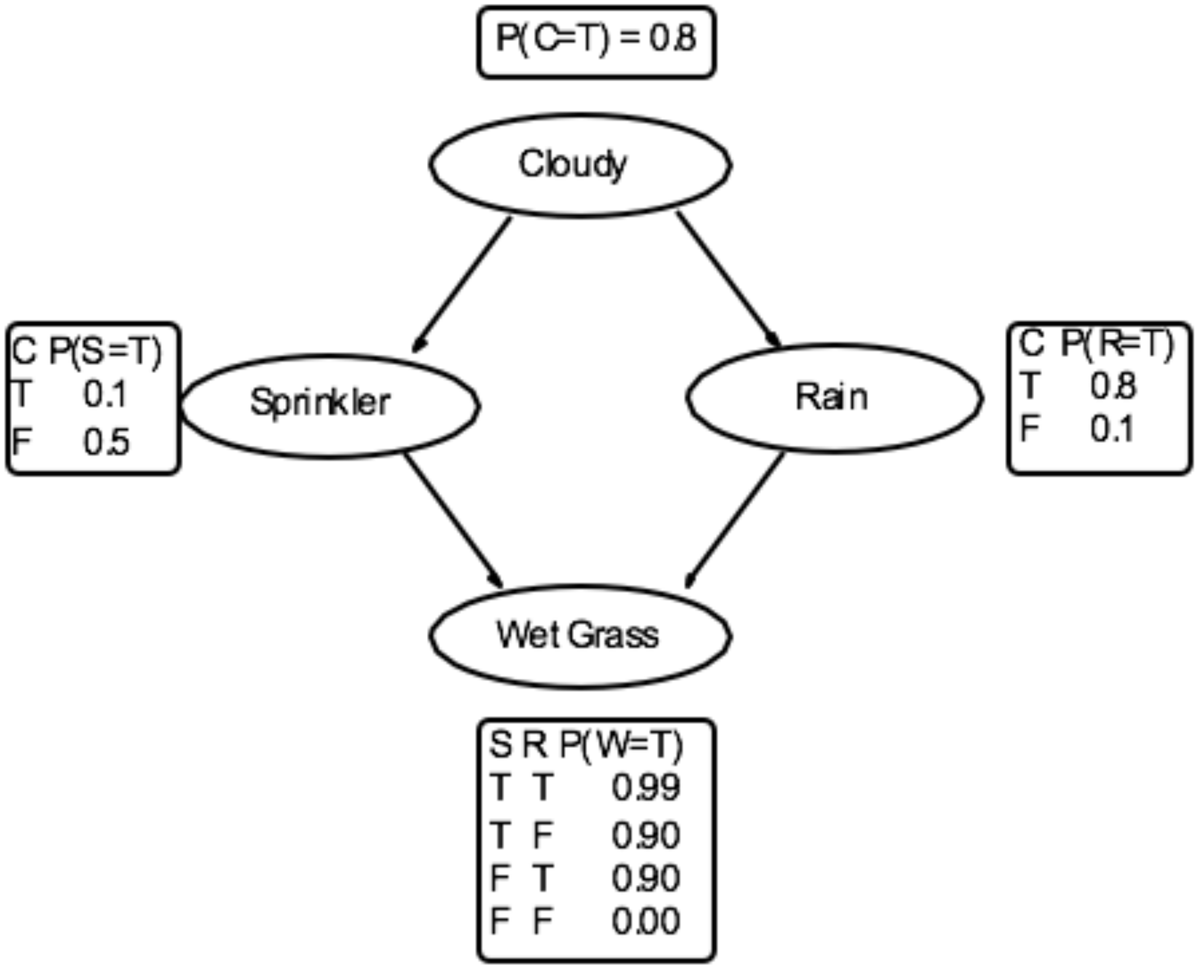
$=P(C=c|A,B)P(A)P(B)/P(C=c)$

A and B are not independent of each other if C is observed.



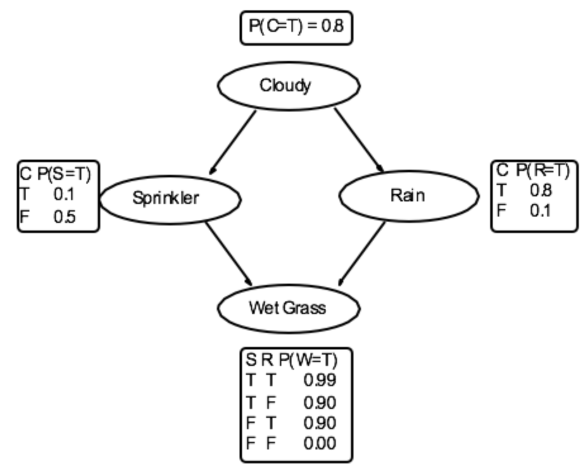
Simple Example

- Consider the following Bayesian network



Simple Example

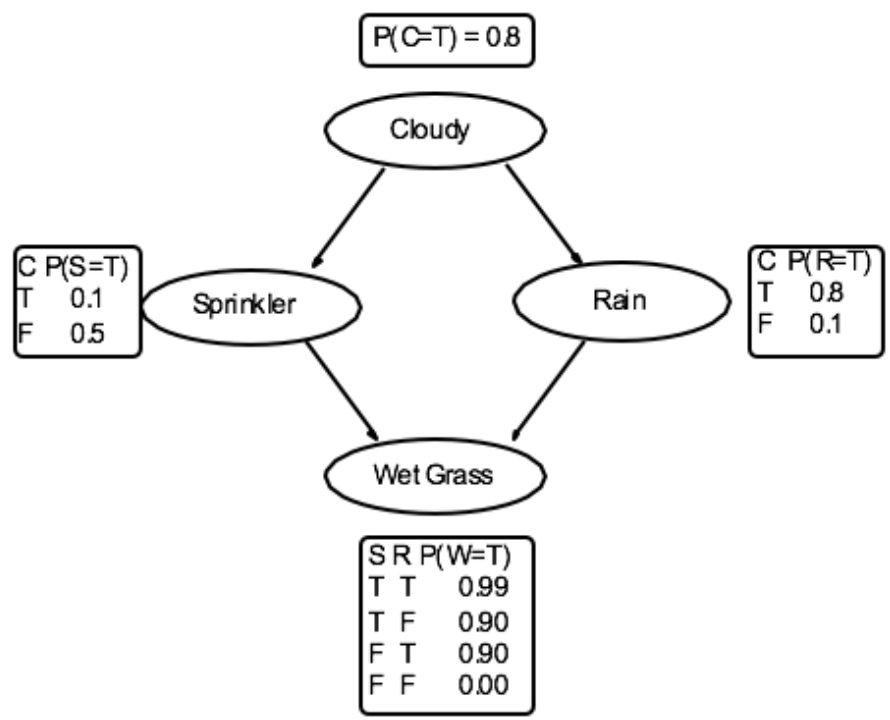
- Whether the grass is wet, *W*
- Whether the sprinkler has been used, *S*
- Whether it has rained, *R*
- Whether the it is cloudy *C*
  - Associated with each node
  - conditional probability table (CPT)



Simple Example

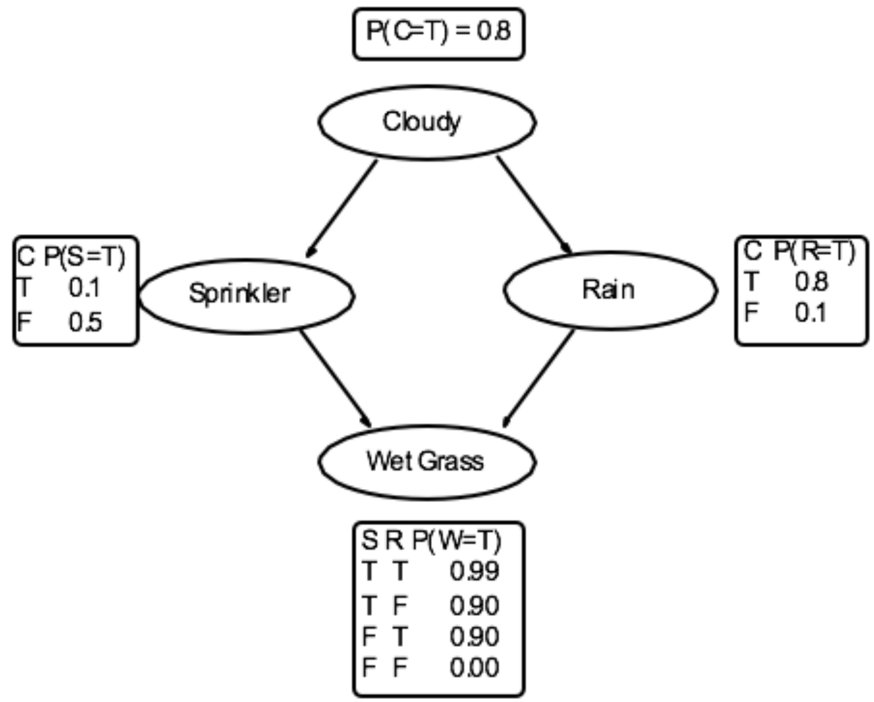
- The Model yields a set of conditional independence assumptions so that:

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



Simple Example

- Possible to use CPTs for inference: Given that it is cloudy, what is the probability that the grass is wet:
- $P(W = T|C = T)$



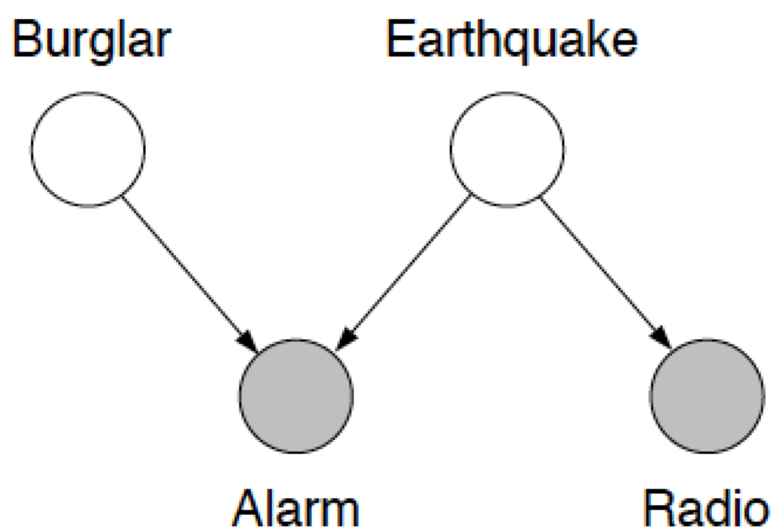
Simple Example

- Possible to use CPTs for inference: Given that it is cloudy, what is the probability that the grass is wet:
- $P(W = T|C = T)$

$$P(W = \mathrm{T}|C = \mathrm{T}) = \sum_{S=\{\mathrm{T},\mathrm{F}\}} \sum_{R=\{\mathrm{T},\mathrm{F}\}} \frac{P(C = \mathrm{T}, S, R, W = \mathrm{T})}{P(C = \mathrm{T})} = 0.7452$$

Example: Alarm

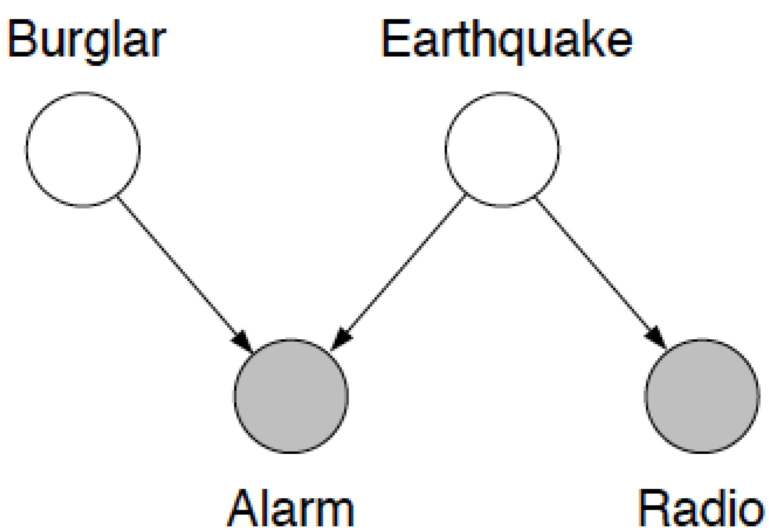
- Question: What is P(E|B)?



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Alarm

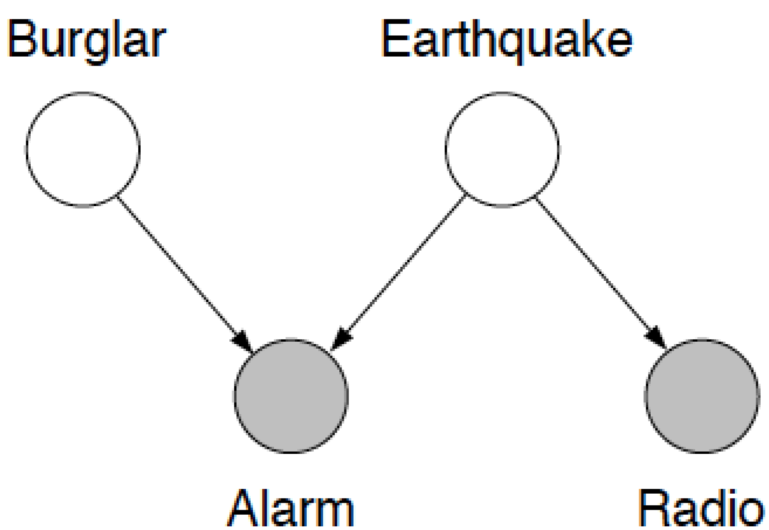
- Question: What is P(E|B)?
- P(E|B)=P(E)=0.01, because they are independent



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Explaining Away

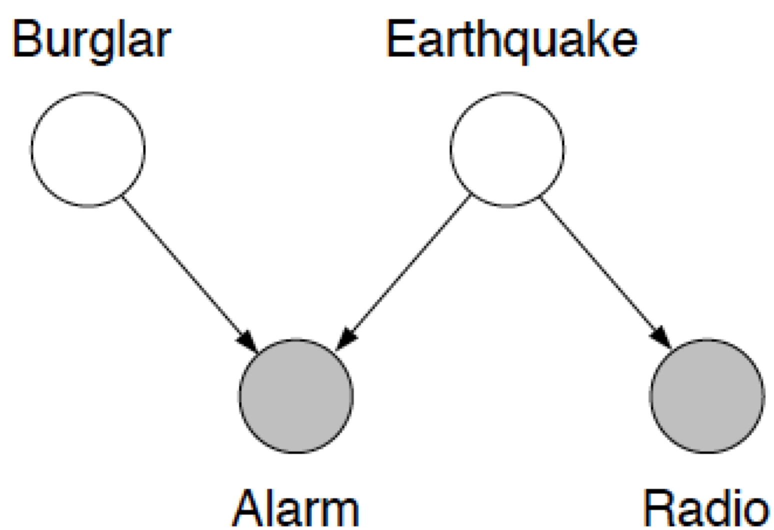
- If something has multiple causes, observing one of the causes reduces the probability of the other cause, i.e. it explains the other cause away.



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Explaining Away

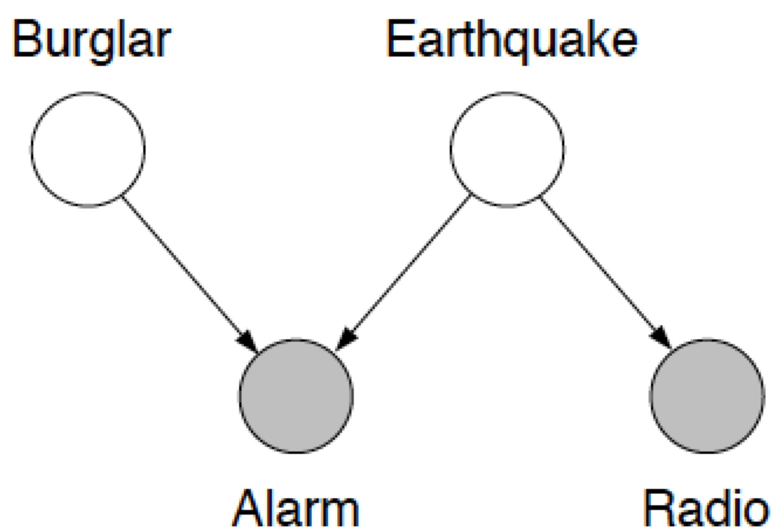
- We have to compare P(E|A,B) with P(E|A)



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Explaining Away

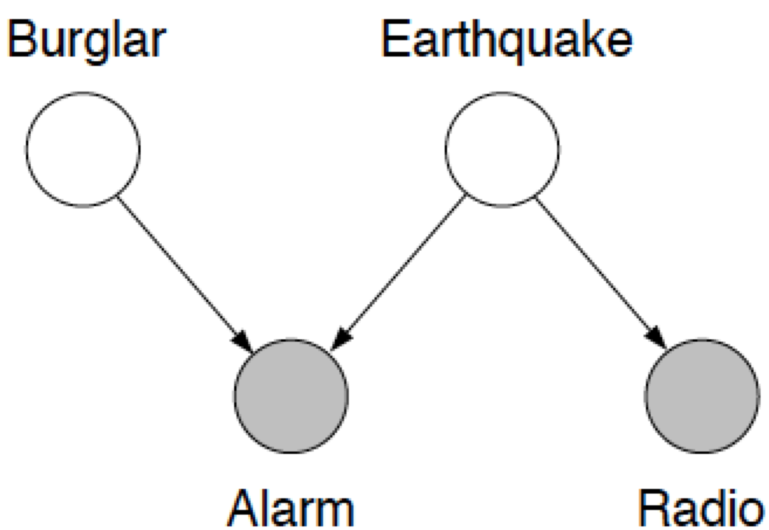
- What is P(E|A)?
- Bayes' Rule:
- P(E|A)=P(A|E)P(E)/P(A)



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Alarm

- First we need to calculate P(A).

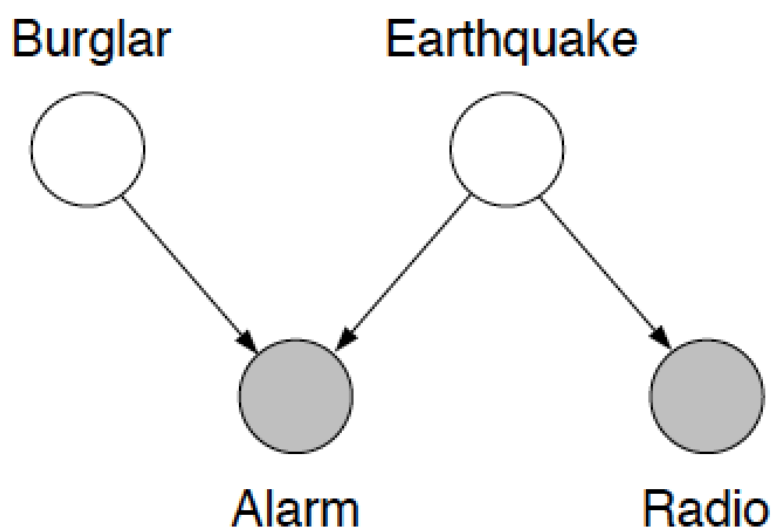


P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Alarm

- First we need to calculate P(A).
- Answer:

P(A)=P(A|B,E)P(B,E)+P(A|B,~E)P(B,~E)+P(A|~B,E)P(~B,E) +P(A|~B,~E) P(~B,~E)=1(0.7)(0.01)+0.9(0.7)(1-0.01)+0.7(1-0.7)(0.01)+0.1(1-0.7)(1-0.01)=0.6625

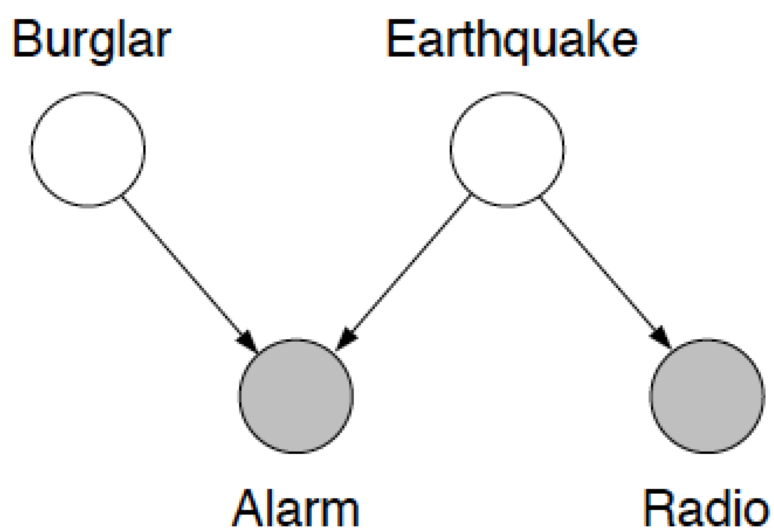


P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Alarm Given Earthquake

- Second, we need to calculate P(A|E).
- Answer:

P(A|E)=P(A|B,E)P(B)+P(A|~B,E)P(~B)=1(0.7)+0.7(1-0.7)=.91

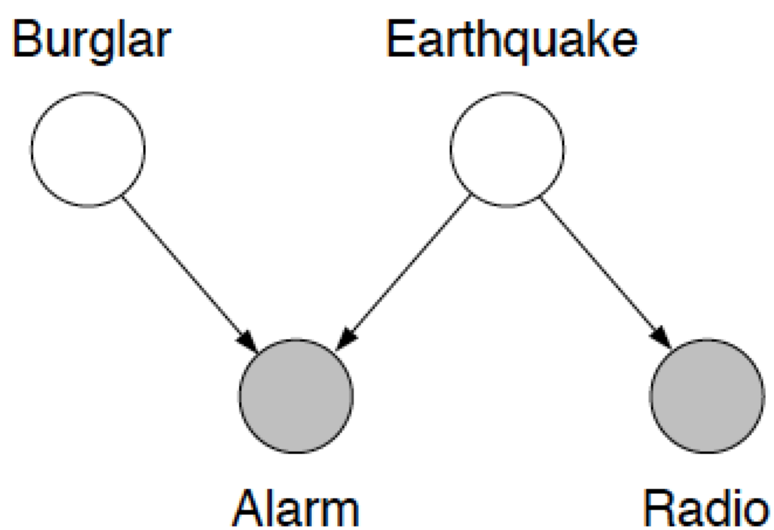


P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Earthquake Given Alarm

- P(E|A)=P(A|E)P(E)/P(A)
- Answer:

P(E|A)=(0.91)(0.01)/0.6625=0.013735



P(B)=0.7
P(E)=0.01
P(A|B,E)=1
P(A|B,~E)=.9
P(A|~B,E)=.7
P(A|~B,~E)=.1

Example: Earthquake | Alarm , Burglar

- Now let's calculate $P(E|A,B)$
- Bayes' Rule:

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

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

But $P(B,E)=P(E|B)P(B)$

So

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

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

Example: Earthquake | Alarm , Burglar

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

- But $P(E|B)=P(E)$ because they are independent

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

- $P(A|B)=P(A|B,E)P(E)+P(A|B,\sim E)P(\sim E)$
- $P(A|B)=1(.01)+0.9(1-0.01)$
- $=0.901$

$P(B)=0.7$
$P(E)=0.01$
$P(A|B,E)=1$
$P(A|B,\sim E)=.9$
$P(A|\sim B,E)=.7$
$P(A|\sim B,\sim E)=.1$

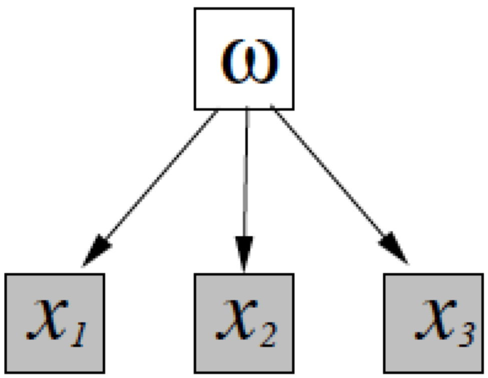
Example: Earthquake | Alarm , Burglar

$$P(E|A,B)=P(A|B,E)P(E)/P(A|B)$$
$$=1(0.01)/0.901=$$
$$0.01109877913<0.013735= P(E|A)$$

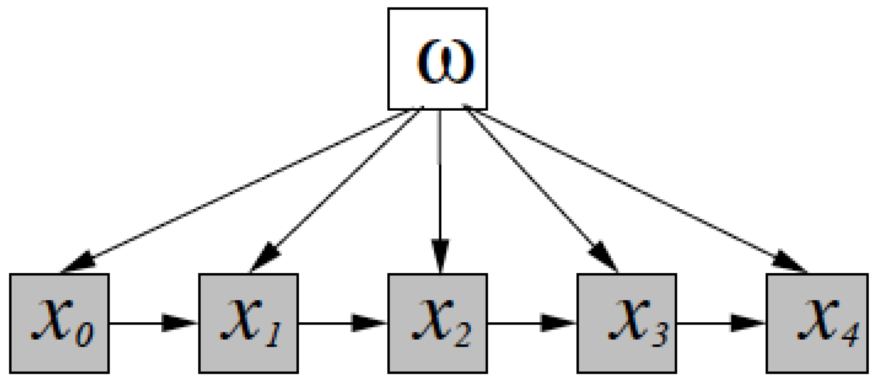
$P(B)=0.7$
$P(E)=0.01$
$P(A|B,E)=1$
$P(A|B,\sim E)=.9$
$P(A|\sim B,E)=.7$
$P(A|\sim B,\sim E)=.1$

Beyond Naive Bayes' Classifier

- Consider classifiers for the class given sequence: $x_1,\ x_2,\ x_3$



$P(\omega_j)\prod_{i=1}^{3} P(x_i|\omega_j)$



$P(\omega_j)P(x_0|\omega_j)\prod_{i=1}^{4} P(x_i|x_{i-1},\omega_j)$

Beyond Naive Bayes' Classifier

- Consider the simple generative classifiers above (with joint distribution)
  - naive-Bayes' classifier on left (conditional independent features given class)
  - for the classifier on the right - a bigram model
    - addition of sequence `start` feature x_0 (note $P(x_0|\omega_j) = 1$)
    - addition of sequence `end` feature x_{d+1} (variable length sequence)
- Decision now based on a more complex model
  - this is the approach used for generating (class-specific) language models

Exercise

Calculate the following probabilities. Give both the formula and calculations with values. These questions are designed so that they can be answered with a minimum of computation.

1. P(a,¬b,c,¬d)

P (a)P (¬b|a)P (c|a)P (¬d|¬b)

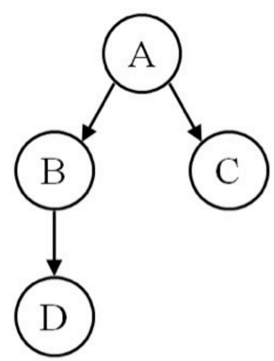
= 0.1 × 0.5 × 0.4 × 0.8 = 0.016

$P(A)$

+a	0.1
¬a	0.9

$P(B|A)$

+a	+b	0.5
+a	¬b	0.5
¬a	+b	0.8
¬a	¬b	0.2



$P(C|A)$

+a	+c	0.4
+a	¬c	0.6
¬a	+c	0.7
¬a	¬c	0.3

$P(D|B)$

+b	+d	0.9
+b	¬d	0.1
¬b	+d	0.2
¬b	¬d	0.8

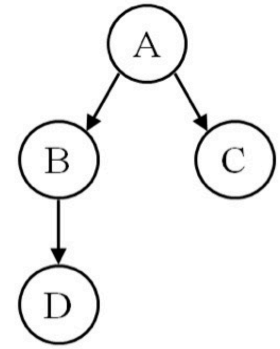
Exercise

$P(A)$

+a	0.1
¬a	0.9

$P(B|A)$

+a	+b	0.5
	¬b	0.5
¬a	+b	0.8
	¬b	0.2



$P(C|A)$

+a	+c	0.4
	¬c	0.6
¬a	+c	0.7
	¬c	0.3

$P(D|B)$

+b	+d	0.9
	¬d	0.1
¬b	+d	0.2
	¬d	0.8

2. P(b)

$P(b) = \sum_{A=\{a,\neg a\}} P(A)P(b|A)$

$= 0.1 \times 0.5 + 0.9 \times 0.8 = 0.77$

3. P(a|b)

$P(a|b) = P(a,b) / P(b) = P(a)P(b|a) / P(b)$

$= 0.1 \times 0.5 / .77 = 0.064935$

Exercise

4. P(d|a)

$P(d|a) = \sum_{B=\{b,\neg b\}} P(d|B)p(B|a)$

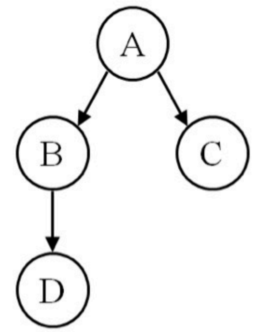
$= 0.9 \times 0.5 + 0.2 \times 0.5 = 0.55$

$P(A)$

+a	0.1
¬a	0.9

$P(B|A)$

+a	+b	0.5
	¬b	0.5
¬a	+b	0.8
	¬b	0.2



$P(C|A)$

+a	+c	0.4
	¬c	0.6
¬a	+c	0.7
	¬c	0.3

$P(D|B)$

+b	+d	0.9
	¬d	0.1
¬b	+d	0.2
	¬d	0.8

5. P(d|a,c)

From the conditional independence properties of the graph, $D \perp C|\{A\}$. Hence, $P(d|a,c) = p(d|a)$
$= 0.55$

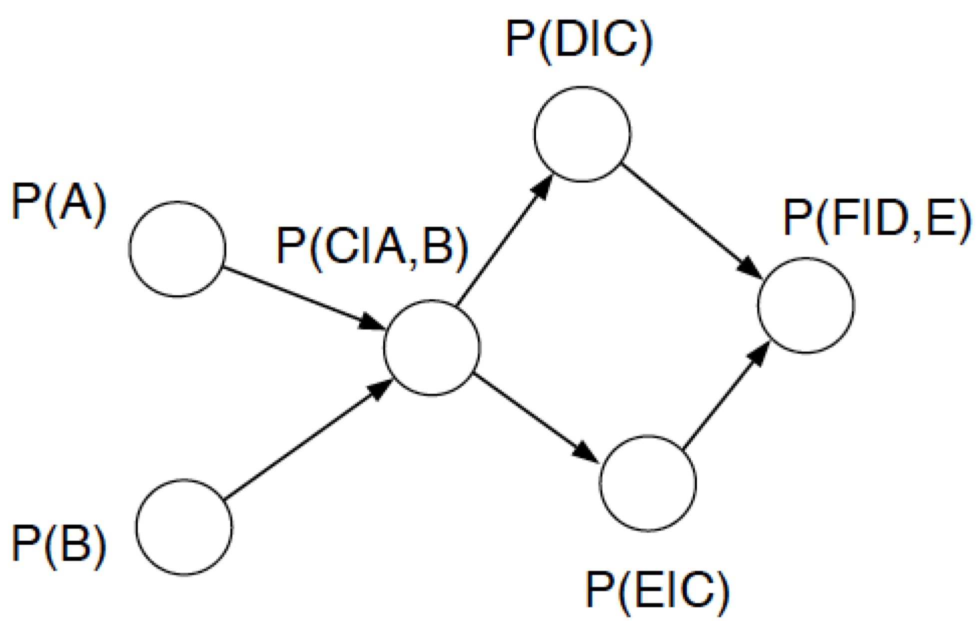
Appendix:
More on Graphical Models

Building GMs Quantitatively

- Question: how do we specify a joint distribution over the nodes in the graph?
- Answer:
  - associate a conditional probability with each node
  - take the product of the local probabilities to yield the global probabilities

Building GMs Quantitatively

- Associate a conditional probability with each node
- Take the product of the local probabilities to yield the global probabilities



Building GMs Quantitatively

- Let $S=\{S_1,\ldots,S_N\}$ represent the set of random variables corresponding to the N nodes of the graph
- For any node S_i , let $\text{pa}(S_i)$ represent the set of parents of node S_i
- Then

$$P(S_1,S_2,\ldots S_N)=\prod_i P\big(S_i\,|\,\text{pa}(S_i)\big)$$

General Inference

- A general approach for inference with Bayesian Networks is message passing
- We present a very brief overview here

General Inference

- Process involves identifying:
- Cliques C: fully connected (every node is connected to every other node) subset of all the nodes.
- Separators S: the subset of the nodes of a clique that are connected to nodes outside the clique.

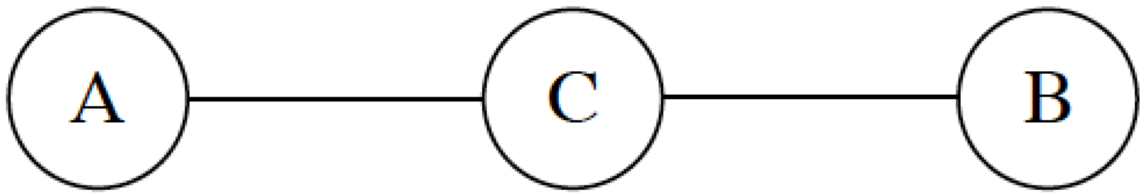
General Inference

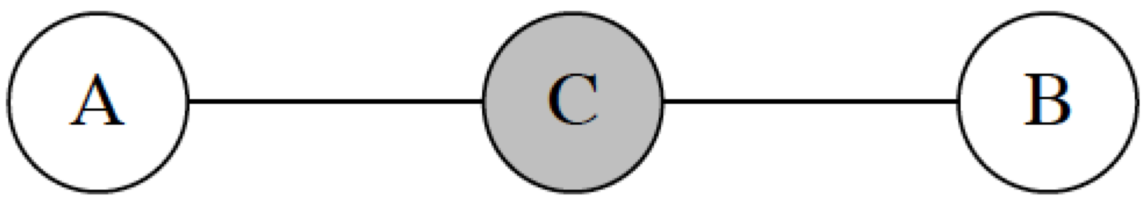
- Thus given the value of the separators for a clique it is conditionally independent of all other variables.

General Inference

- To understand General Inference, we need to understand the semantics of undirected Graphical Models.

Semantics of undirected graphs

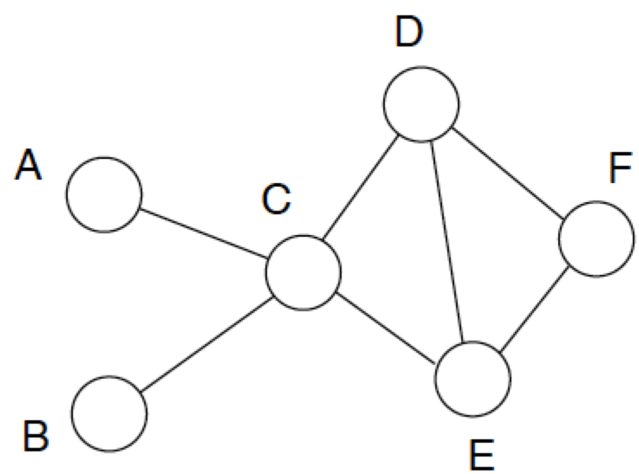




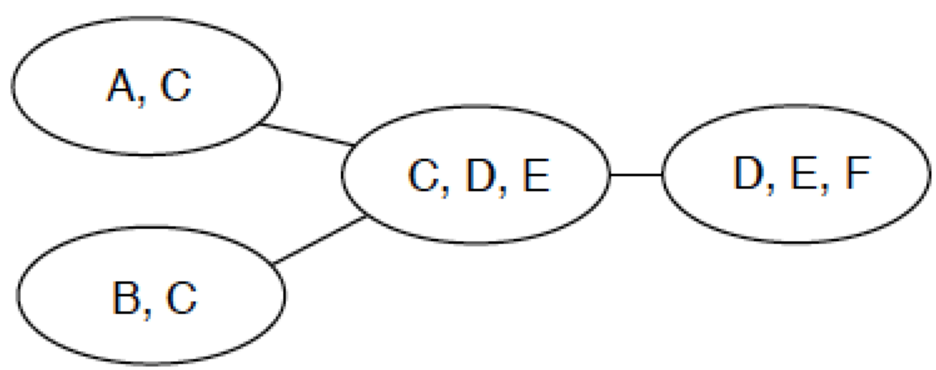
A and B are marginally dependent

A and B are conditionally independent

Quantitative specifcation of undirected models

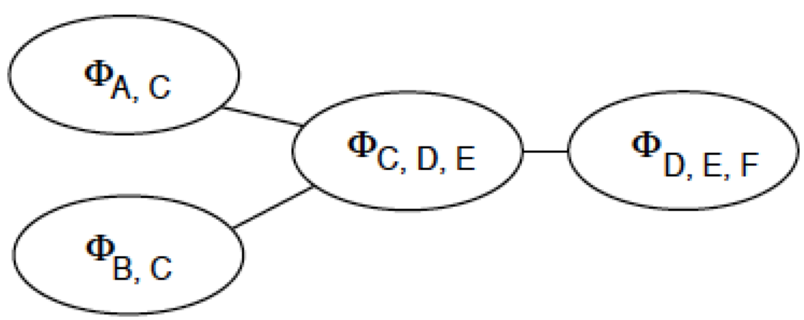


Identify the cliques in the graph:



Quantitative specifcation of undirected models

- Define a configuration of a clique as a specification of values for each node in the clique
- Define a potential of a clique as a function that associates a real number with each configuration of the clique



Quantitative specifcation of undirected models

Consider the example of a graph with binary nodes

A potential is a table with entries for each combination of nodes in a clique

		B	
		0	1
A	0	1.5	.4
	1	.7	1.2

Quantitative specifcation of undirected models

“Marginalizing” over a potential table simply means collapsing (summing) the table along one or more dimensions

- marginalizing over B

A	
0	1.9
1	1.9

- marginalizing over A

B	0	1
	2.2	1.6

Quantitative specifcation of undirected models

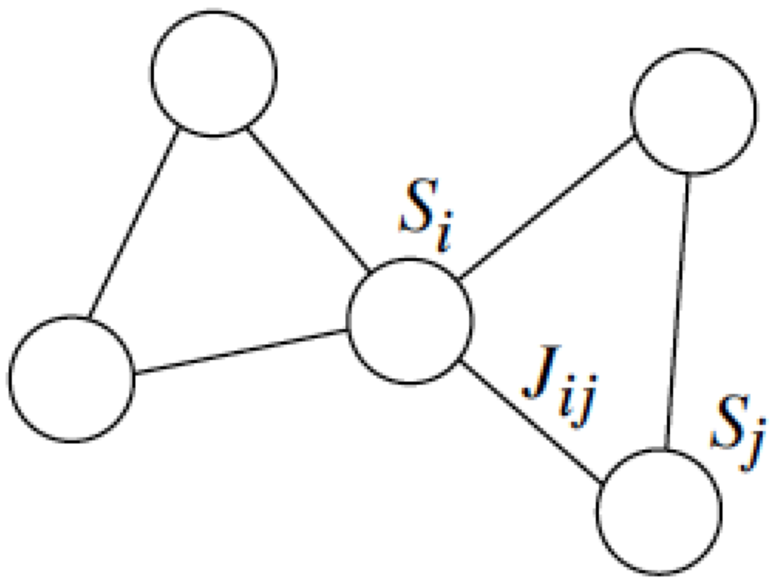
Finally, define the probability of a global configuration of the nodes as the product of the local potentials on the cliques:

$$P(A, B, C, D, E, F) = \phi_{(A,B)}\phi_{(B,C)}\phi_{(C,D,E)}\phi_{(D,E,F)}$$

where, without loss of generality, we assume that the normalization constant (if any) has been absorbed into one of the potentials

Boltzmann machine

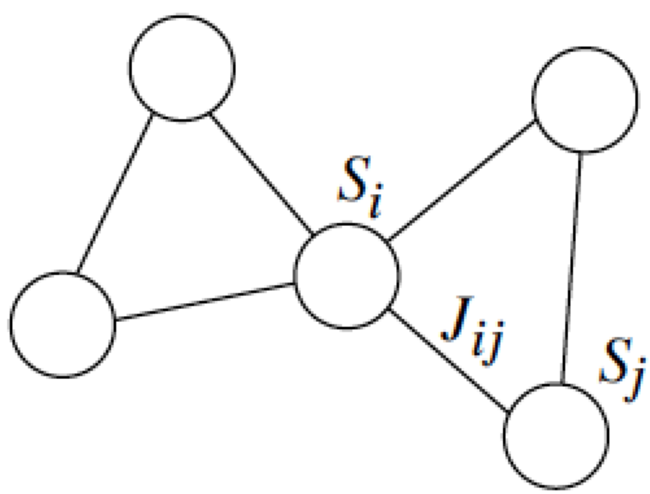
- The Boltzmann machine is a special case of an undirected graphical model
- For a Boltzmann machine all of the potentials are formed by taking products of factors of the form $\exp(S_i S_j J_{ij})$



Boltzmann machine

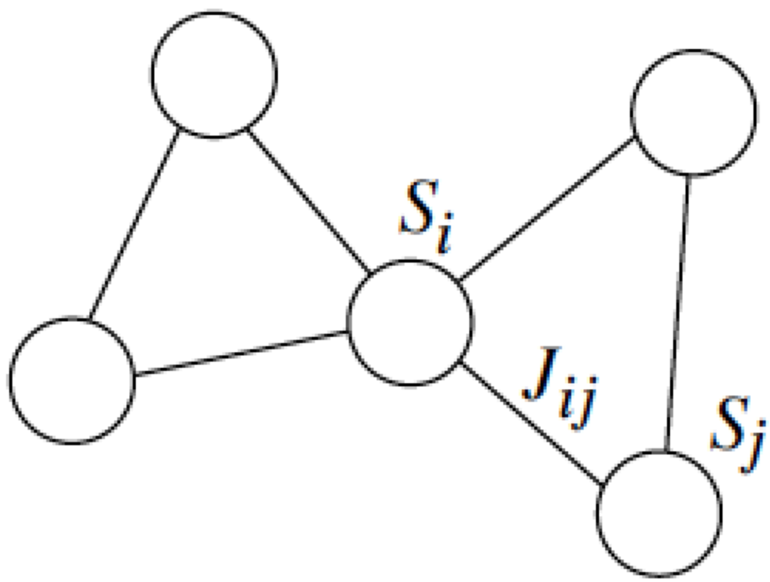
Setting J_{ij} equal to zero for non-neighboring nodes guarantees that we respect the clique boundaries

But we don't get the full conditional probability semantics with the Boltzmann machine parameterization



Boltzmann machine

The family of distributions parameterized by a Boltzmann machine on a graph is a proper subset of the family characterized by the conditional independencies



Inference algorithms for directed graphs

Several inference algorithms; some operate directly on the directed graph

The most popular inference algorithm, known as the junction tree algorithm (which we'll discuss here), operates on an undirected graph

Inference algorithms for directed graphs

It also has the advantage of clarifying some of the relationships between the various algorithms

To understand the junction tree algorithm, we need to understand how to "compile" a directed graph into an undirected graph

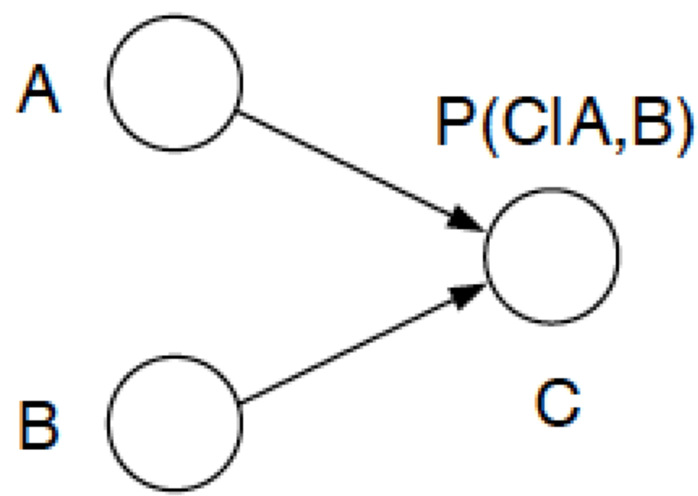
Moral graphs

Note that for both directed graphs and undirected graphs, the joint probability is in a product form

So let's convert local conditional probabilities into potentials; then the products of potentials will give the right answer

Moral graphs

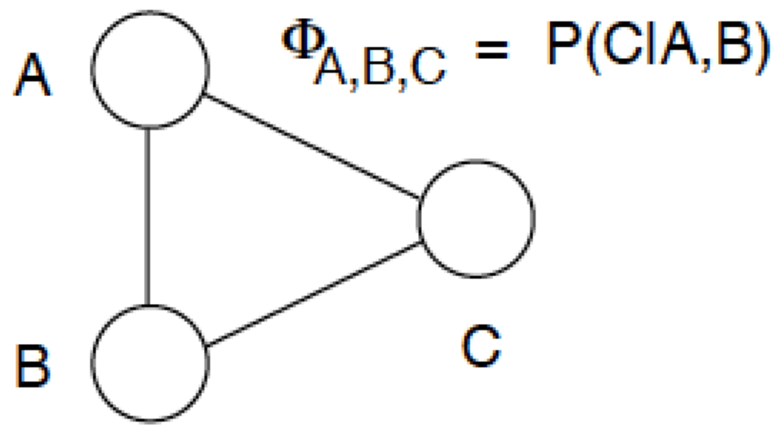
Indeed we can think of a conditional probability, e.g., P (C|A, B) as a function of the three variables A, B, and C (we get a real number for each configuration):



Moral graphs

Problem: A node and its parents are not generally in the same clique

Solution: Marry the parents to obtain the "moral graph"



Moral graphs

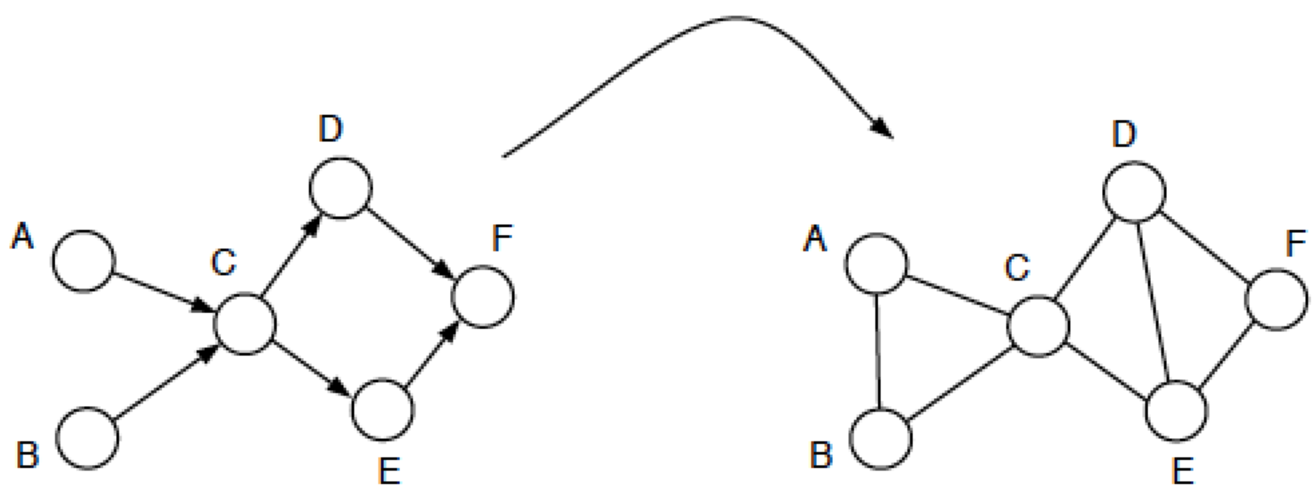
Define the potential on a clique as the product over all conditional probabilities contained within the clique

Moral graphs

Now the products of potentials gives the right answer:

$$P(A, B, C, D, E, F) = P(A)P(B)$$

$$P(C|A, B)P(D|C)P(E|C)P(F|D, E)$$



Moral graphs

$P(A, B, C, D, E, F) = P(A)P(B)$
$P(C|A, B)P(D|C)P(E|C)P(F|D, E)$
$=\varphi(A,B,C)\ \varphi(C,D,E)\ \varphi(D,E,F)$

Where:

- $\varphi(A,B,C) = P(A)P(B)\ P(C|A, B)$
- $\varphi(C,D,E) = P(D|C)P(E|C)$
- $\varphi(D,E,F) = P(F|D, E)$

Evidence and Inference

"Absorbing evidence" means observing the values of certain of the nodes

Absorbing evidence divides the units of the network into two groups:

visible units $\{V\}$	those for which we have instantiated values ("evidence nodes").
hidden units $\{H\}$	those for which we do not have instantiated values.

Evidence and Inference

"Inference" means calculating the conditional distribution

visible units $\{V\}$	those for which we have instantiated values ("evidence nodes").
hidden units $\{H\}$	those for which we do not have instantiated values.

$$P(H|V) = \frac{P(H,V)}{\Sigma_{\{H\}} P(H,V)}$$

- Prediction and diagnosis are special cases

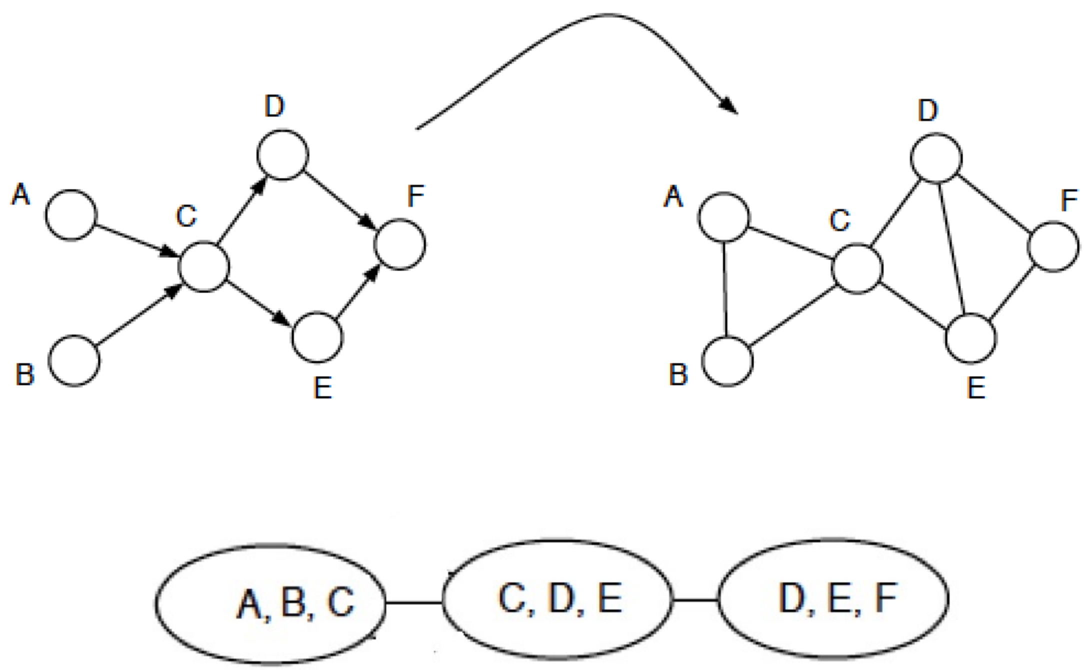
Propagation of probabilities

Now suppose that some evidence has been absorbed.

How do we propagate this effect to the rest of the graph?

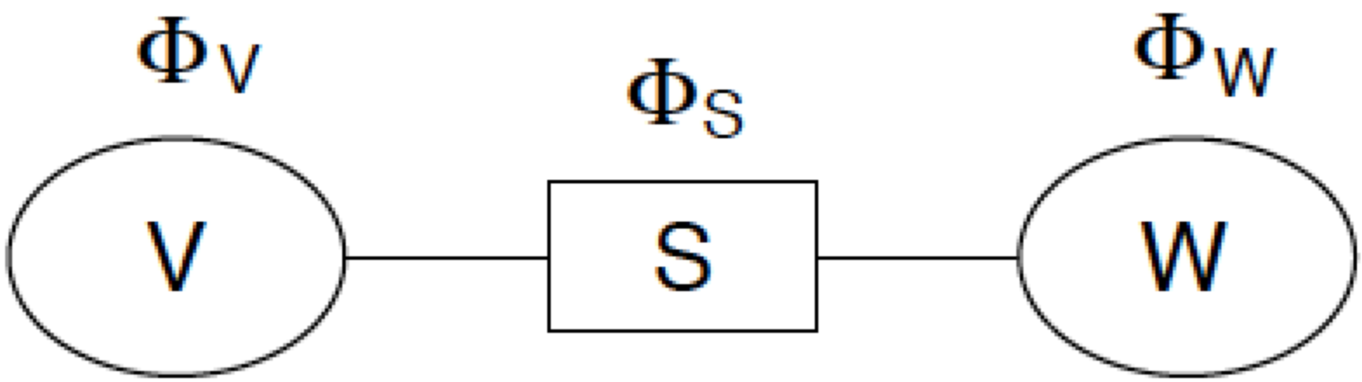
Clique trees

A clique tree is an (undirected) tree of cliques



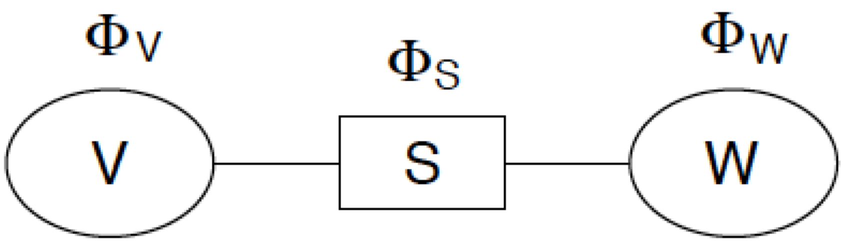
Clique trees

Consider cases in which two neighboring cliques V and W have an overlap S (e.g., (A, C) overlaps with (C, D, E)).



Clique trees

The cliques need to “agree" on the probability of nodes in the overlap; this is achieved by marginalizing and rescaling:



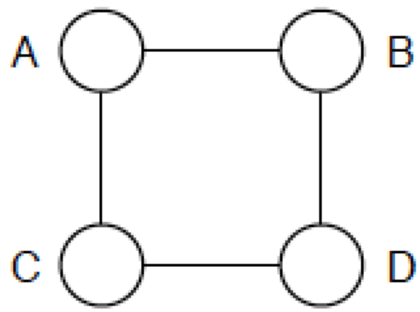
$$\phi_S^* = \sum_{V \setminus S} \phi_V$$

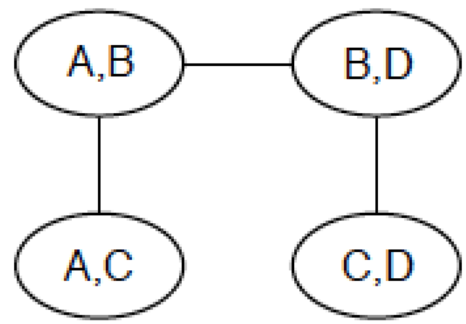
$$\phi_W^* = \phi_W \frac{\phi_S^*}{\phi_S}$$

This occurs in parallel, distributed fashion throughout the clique tree

A problem

Consider the following graph and a corresponding clique tree:



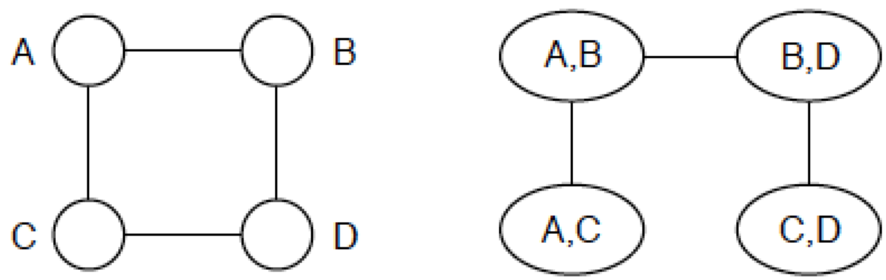


Note that C appears in two non-neighboring cliques.

Question: What guarantee do we have that the probability associated with C in these two cliques will be the same?

A problem

Question: What guarantee do we have that the probability associated with C in these two cliques will be the same?

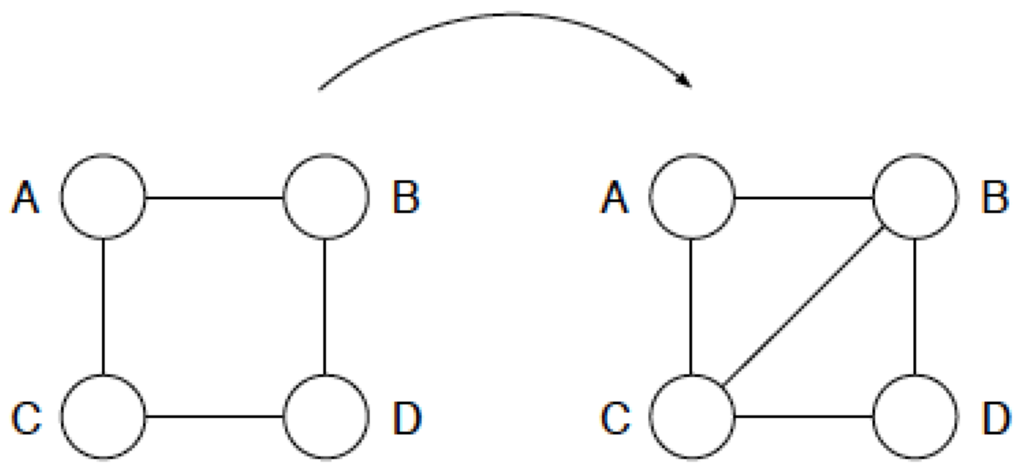


Answer: Nothing. In fact this is a problem with the algorithm as described so far. It is not true that in general local consistency implies global consistency.

Triangulation (last idea, hang in there)

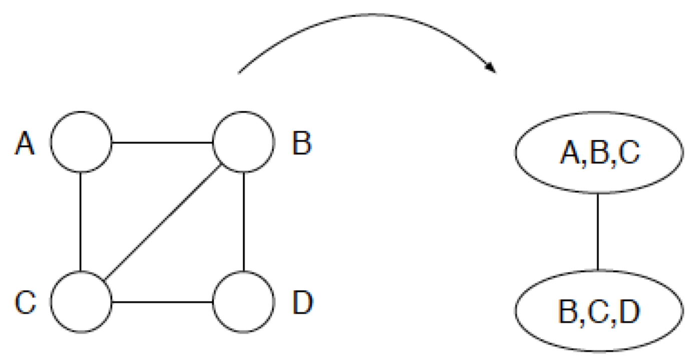
A triangulated graph is one in which no cycles with four or more nodes exist in which there is no chord.

We triangulate a graph by adding chords:



Triangulation

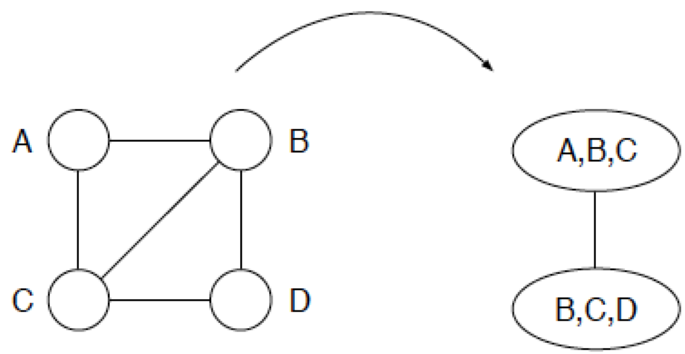
Now we no longer have our problem:



A clique tree for a triangulated graph has the running intersection property: if a node appears in two cliques, it appears everywhere on the path between the cliques

Thus local consistency implies global consistency for such clique trees

Triangulation



Thus local consistency implies global consistency for such clique trees

Junction trees

A clique tree for a triangulated graph is referred to as a junction tree

In junction trees, local consistency implies global consistency. Thus the local message-passing algorithm is (provably) correct.

Junction trees

It's also possible to show that only triangulated graphs have the property that their clique trees are junction trees. Thus, if we want local algorithms, we must triangulate.

Summary of the junction tree algorithm

1. Moralize the graph
2. Triangulate the graph
3. Propagate by local message-passing in the junction tree

- Note that the first two steps are “offline"
- Note also that these steps provide a bound of the complexity of the propagation step