

Bayesian Networks

Fabrizio Riguzzi

- Acknowledgments: some slides from
 - Andrew Moore's tutorials
<http://www.autonlab.org/tutorials/>
 - Irina Rish and Moninder Singh's tutorial
<http://www.research.ibm.com/people/r/rish/>

Summary

- Conditional independence
- Definition of Bayesian networks
- Inference
- Learning
- Markov networks

Domain Modeling

- We use a set of random variables to describe the domain of interest
- Example: home intrusion detection system, variables:
 - Earthquake E , values $e_1=\text{no}$, $e_2=\text{moderate}$, $e_3=\text{severe}$
 - Burglary B , values: $b_1=\text{no}$, $b_2=\text{yes through door}$, $b_3=\text{yes through window}$
 - Alarm A , values $a_1=\text{no}$, $a_2=\text{yes}$
 - Neighbor call N , values $n_1=\text{no}$, $n_2=\text{yes}$

Inference

- We would like to answer the following questions
 - What is the probability of a burglary through the door? (compute $P(b_2)$, belief computation)
 - What is the probability of a burglary through the door given that the neighbor called ? (compute $P(b_2|n_2)$, belief updating)

Inference

- What is the probability of a burglary through the door given that there was a moderate earthquake and the neighbor called ? (compute $P(b_2|n_2,e_2)$, belief updating)
- What is the probability of a burglary through the door and of the alarm ringing given that there was a moderate earthquake and the neighbor called ? (compute $P(a_2,b_2|n_2,e_2)$, belief updating)
- What is the most likely value for burglary given that the neighbor called ($\text{argmax}_b P(b|n_2)$, belief revision)

Types of Problems

- Diagnosis: $P(\text{cause}|\text{symptom})=?$
- Prediction: $P(\text{symptom}|\text{cause})=?$
- Classification: $\operatorname{argmax}_{\text{class}} P(\text{class}|\text{data})$

Inference

- In general, we want to compute the probability $P(\mathbf{q}|\mathbf{e})$
 - of a query $\mathbf{q}$ (assignment of values to a set of variables $\mathbf{Q}$)
 - given the evidence $\mathbf{e}$ (assignment of values to a set of variables $\mathbf{E}$)

Joint Probability Distribution

- The **joint probability distribution** (jpd) of a set of variables $\mathbf{U}$ is given by $P(\mathbf{u})$ for all values $\mathbf{u}$
- For our example
 - $\mathbf{U}=\{E,B,A,N\}$
 - We have the jpd if we know $P(\mathbf{u})=P(e,b,a,n)$ for all the possible values e, b, a, n .

Inference

- If we know the jpd, we can answer all the possible queries:

$$\begin{aligned} P(\mathbf{q}|\mathbf{e}) &= \frac{P(\mathbf{q}, \mathbf{e})}{P(\mathbf{e})} \\ &= \frac{\sum_{\mathbf{x}, \mathbf{X} = \mathbf{U} \setminus \mathbf{Q} \setminus \mathbf{E}} P(\mathbf{x}, \mathbf{q}, \mathbf{e})}{\sum_{\mathbf{y}, \mathbf{Y} = \mathbf{U} \setminus \mathbf{E}} P(\mathbf{y}, \mathbf{e})} \end{aligned}$$

Computational Cost

- If we have n binary variables ($|\mathbf{U}|=n$), knowing the jpd requires storing $O(2^n)$ different values.
- Even if we had the space to store all the 2^n different values, computing $P(\mathbf{q}|\mathbf{e})$ would require $O(2^n)$ operations
- Impractical for real world domains
- How to avoid the space and time problems? Use conditional independence assertions

Conditional Independence

- $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ vectors of multivalued variables
- $\mathbf{X}$ and $\mathbf{Y}$ are **conditionally independent** given $\mathbf{Z}$ if

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z}: P(\mathbf{y}, \mathbf{z}) > 0 \rightarrow P(\mathbf{x} | \mathbf{y}, \mathbf{z}) = P(\mathbf{x} | \mathbf{z})$$

- We write $I\langle \mathbf{X}, \mathbf{Z}, \mathbf{Y} \rangle$
- Special case: $\mathbf{X}$ and $\mathbf{Y}$ are **independent** if

$$\forall \mathbf{x}, \mathbf{y}: P(\mathbf{y}) > 0 \rightarrow P(\mathbf{x} | \mathbf{y}) = P(\mathbf{x})$$

- We write $I\langle \mathbf{X}, \{\}, \mathbf{Y} \rangle$

Chain Rule

- n random variables $X_1, \dots, X_n$
- Let $\mathbf{U} = \{X_1, \dots, X_n\}$
- Joint event $\mathbf{u} = (x_1, \dots, x_n)$
- Chain rule:

$$\begin{aligned} P(\mathbf{u}) &= P(x_1, \dots, x_n) \\ &= P(x_n | x_{n-1} \dots, x_1) \dots P(x_2 | x_1) P(x_1) \\ &= \prod_{i=1}^n P(x_i | x_{i-1} \dots, x_1) \end{aligned}$$

Conditional Independence

- Π_i is a subset of $\{X_{i-1}, \dots, X_1\}$ such that
- X_i is conditionally independent of $\{X_{i-1}, \dots, X_1\} \setminus \Pi_i$ given Π_i
$$P(x_i | x_{i-1} \dots, x_1) = P(x_i | \pi_i) \quad \text{whenever } P(x_{i-1}, \dots, x_1) > 0$$
- where π_i is a set of values for Π_i
- Π_i **parents** of X_i

Conditional Independence

- Knowing $\boldsymbol{\pi}_i$ for all i we could write

$$\begin{aligned} P(\mathbf{u}) &= P(x_1, \dots, x_n) \\ &= P(x_n | x_{n-1}, \dots, x_1) \dots P(x_2 | x_1) P(x_1) \\ &= P(x_n | \boldsymbol{\pi}_n) \dots P(x_2 | \boldsymbol{\pi}_2) P(x_1 | \boldsymbol{\pi}_1) \\ &= \prod_{i=1}^n P(x_i | \boldsymbol{\pi}_i) \end{aligned}$$

Conditional Independence

- In order to compute $P(\mathbf{u})$ we have to store

$$P(x_i | \boldsymbol{\pi}_i)$$

- for all values x_i and $\boldsymbol{\pi}_i$
- $P(x_i | \boldsymbol{\pi}_i)$: Conditional probability table
- If $\boldsymbol{\Pi}_i$ is much smaller than the set $\{X_{i-1}, \dots, X_1\}$, then we have huge savings
- If k is the maximum number of parents of a variable, then storage is $O(n2^k)$ instead of $O(2^n)$

Graphical Representation

- We can represent the conditional independence assertions using a directed graph with a node per variable
- Π_i is the set of parents of X_i
- The graph is acyclic

Example Network

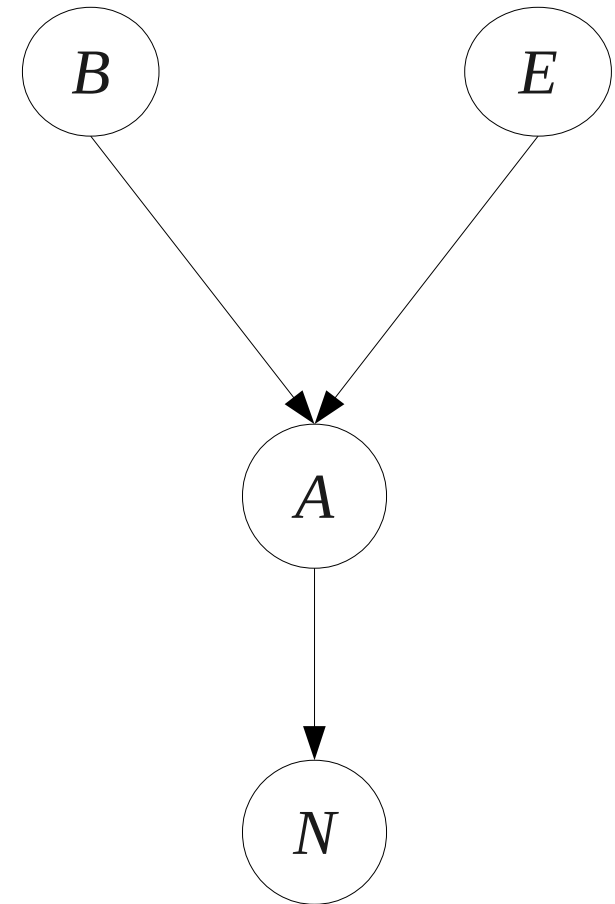
- Variable order: E,B,A,N
- Independencies

$$P(e)$$

$$P(b|e)=P(b)$$

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

$$P(n|a,b,e)=P(n|a)$$



Conditional Probability Tables

- Earthquake E, e_1 =no, e_2 =moderate, e_3 =severe
- Burglary B, : b_1 =no, b_2 =yes through door, b_3 =yes through window
- Alarm A, a_1 =no, a_2 =yes
- Neighbor call N, n_1 =no, n_2 =yes

P(B)	
B=no	0,7
B=door	0,1
B=windows	0,2

P(E)	
E=no	0,6
E=moderate	0,2
E=severe	0,2

P(A EB)	no,no	no,do	no,wi	mo,no	mo,do	mo,wi	se,no	se,do	se,wi
no	0,99	0,1	0,2	0,8	0,08	0,1	0,7	0,05	0,07
yes	0,01	0,9	0,8	0,2	0,92	0,9	0,3	0,95	0,93

P(N A)	A=no	A=yes
N=no	0,9	0,05
N=yes	0,1	0,95

Bayesian Network

- A **Bayesian network** [Pearl 85] (BN) B is a couple (G, Θ) where
 - G is a directed acyclic graph (DAG) (V, E) where
 - V is a set of vertices $\{X_1, \dots, X_n\}$
 - E is a set of edges, i.e. A set of couples (X_i, X_j)
 - $\langle X_1, \dots, X_n \rangle$ is a topological sort of G , i.e. $(X_i, X_j) \in E \Rightarrow i < j$
 - Θ is a set of conditional probability tables (cpt's)
$$\Theta = \left\{ \left\{ \theta_{x_i | \pi_i} \mid x_i \in X_i, \pi_i \in \Pi_i \right\} \mid i = 1, \dots, n \right\}$$
 - where Π_i is the set of parents of X_i

Bayesian Network

- BNs are also called belief networks or directed acyclic graphical models

Bayesian Network

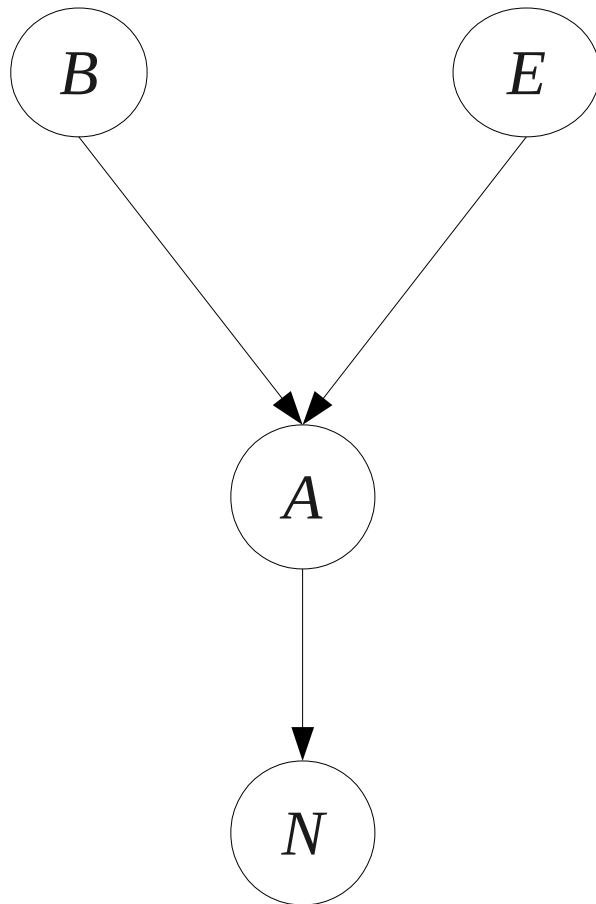
- A BN (G, Θ) **represents** a jpd P iff
 - given its parents in G , each variable is independent of its other predecessors

$$P(x_i | x_{i-1} \dots, x_1) = P(x_i | \pi_i)$$

- $\theta_{x_i | \pi_i} = P(x_i | \pi_i)$ for all i and π_i
- In this case

$$\begin{aligned} P(x_1, \dots, x_n) &= \prod_{i=1}^n P(x_i | \pi_i) \\ &= \prod_{i=1}^n \theta_{x_i | \pi_i} \end{aligned}$$

Example: Alarm with Boolean variables



$P(B)$	
$B=\text{yes}$	0,1
$B=\text{no}$	0,9

$P(E)$	
$E=\text{yes}$	0,05
$E=\text{no}$	0,95

$P(A BE)$	no,no	no,yes	yes,no	yes,yes
$A=\text{yes}$	0,1	0,85	0,9	0,99
$A=\text{no}$	0,9	0,15	0,1	0,01

$P(N A)$	$A=\text{no}$	$A=\text{yes}$
$N=\text{yes}$	0,1	0,95
$N=\text{no}$	0,9	0,05

Example

- Compute $P(B=\text{yes}|N=\text{yes},E=\text{no})=P(B|N,\sim E)$ using the joint probability
- $P(B|N,\sim E)=P(B,N,\sim E)/P(N,\sim E)$
- $P(B,N,\sim E)=P(B,\sim E,\sim A,N)+P(B,\sim E,A,N)$
- $P(N,\sim E)=P(B,N,\sim E)+P(\sim B,N,\sim E)=$
 $P(B,\sim E,\sim A,N)+P(B,\sim E,A,N) + P(\sim B,\sim E,\sim A,N)$
 $+P(\sim B,\sim E,A,N)$

Example

$$P(B, \sim E, \sim A, N) = P(B)P(\sim E) \\ P(\sim A|B, \sim E)P(N|\sim A) = \\ 0.1 * 0.95 * 0.1 * 0.1 = 0.00095$$

P(B)	
B=yes	0,1
B=no	0,9

P(E)	
E=yes	0,05
E=no	0,95

$$P(B, \sim E, A, N) = P(B)P(\sim E) \\ P(A|B, \sim E)P(N|A) = \\ 0.1 * 0.95 * 0.9 * 0.95 = 0.081225$$

P(A BE)	no,no	no,yes	yes,no	yes,yes
A=yes	0,1	0,85	0,9	0,99
A=no	0,9	0,15	0,1	0,01

P(N A)	A=no	A=yes
N=yes	0,1	0,95
N=no	0,9	0,05

Example

$$P(\sim B, \sim E, \sim A, N) = P(\sim B)P(\sim E)P(\sim A | \sim B, \sim E)P(N | \sim A) = 0.9 * 0.95 * 0.9 * 0.1 = 0.07695$$

P(B)	
B=yes	0,1
B=no	0,9

P(E)	
E=yes	0,05
E=no	0,95

$$P(\sim B, \sim E, A, N) = P(\sim B)P(\sim E)P(A | \sim B, \sim E)P(N | A) = 0.9 * 0.95 * 0.1 * 0.95 = 0.081225$$

P(A BE)	no,no	no,yes	yes,no	yes,yes
A=yes	0,1	0,85	0,9	0,99
A=no	0,9	0,15	0,1	0,01

P(N A)	A=no	A=yes
N=yes	0,1	0,95
N=no	0,9	0,05

Example

- $P(B, \sim E, \sim A, N) = 0.00095$
- $P(B, \sim E, A, N) = 0.081225$
- $P(\sim B, \sim E, \sim A, N) = 0.07695$
- $P(\sim B, \sim E, A, N) = 0.081225$

- $P(B, N, \sim E) = P(B, \sim E, \sim A, N) + P(B, \sim E, A, N) = 0.00095 + 0.081225 = 0.082175$
- $P(N, \sim E) = P(B, N, \sim E) + P(\sim B, N, \sim E) = 0.082175 + 0.07695 + 0.081225 = 0.24035$
- $P(B|N, \sim E) = P(B, N, \sim E) / P(N, \sim E) = 0.082175 / 0.24035 = 0.3418972$

How to Build a Bayesian Network

- Choose an ordering $X_1 \dots X_n$ for the variables
- For $i = 1$ to n :
 - Add X_i node to the network
 - Set Π_i to be a minimal subset of $\{X_1 \dots X_{i-1}\}$ such that we have conditional independence of X_i and all other members of $\{X_1 \dots X_{i-1}\}$ given Π_i
 - Assign a value to $P(x_i | \pi_i)$ for all the values of x_i and π_i

Building a Bayesian Network

- Usually the expert considers a variable X as a child of Y if Y is a **direct cause** of X
- Correlation and causality are related but are **not** the same thing
 - See the book [Pearl 00]

Pathfinder system [Suermondt et al. 90]

- Diagnostic system for lymph-node diseases.
- 60 diseases and 100 symptoms and test-results.
- 14,000 probabilities
- Expert consulted to make net.
- 8 hours to determine variables.
- 35 hours for net topology.
- 40 hours for probability table values.

Pathfinder system [Suermondt et al. 90]

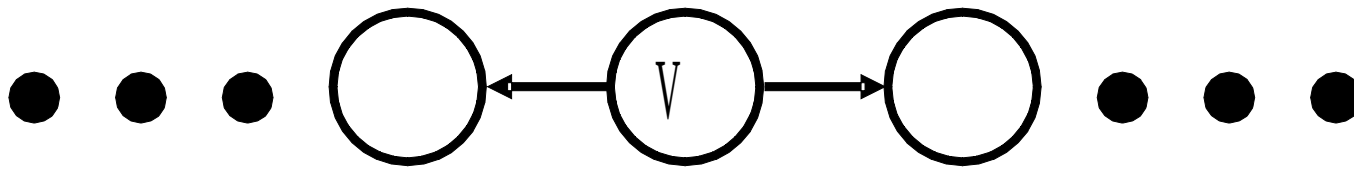
- Pathfinder is now outperforming the world experts in diagnosis.
- Being extended to several dozen other medical domains.

How to Tell Independence

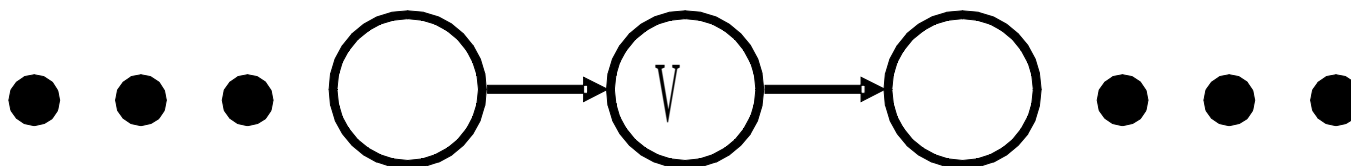
- There is a relatively simple algorithm for determining whether two variables in a Bayesian network are conditionally independent:
d-separation.
- Definition: X and Y are **d-separated** by a set of evidence variables Z iff every undirected path from X to Y is “blocked”, where a path is “blocked” iff one or more of the following conditions is true: ...

Blocked Path

There exists a variable V on the path such that
it **is** in the evidence set Z
the arcs putting V in the path are “tail-to-tail”

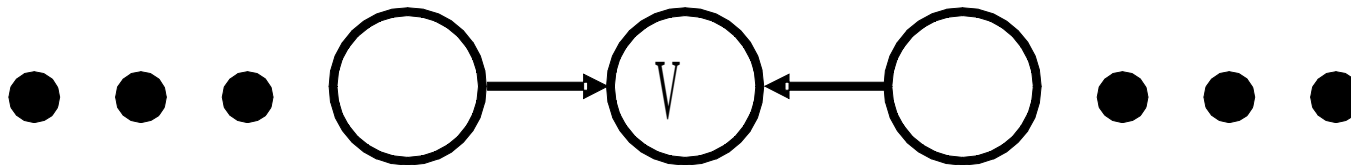


Or, there exists a variable V on the path such that
it **is** in the evidence set Z
the arcs putting V in the path are “tail-to-head”

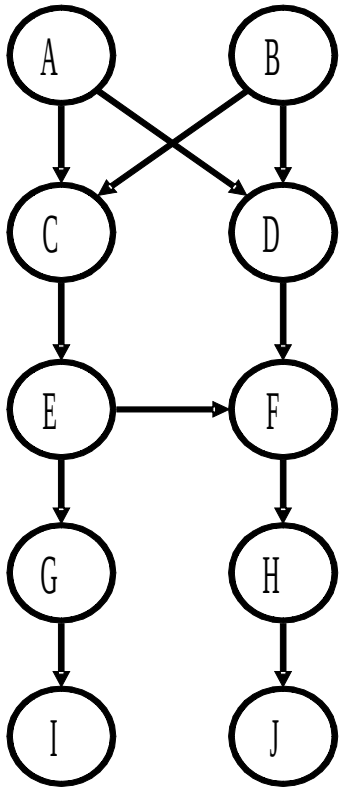


Blocked Path

- ... Or, there exists a variable V on the path such that it **is NOT** in the evidence set Z
neither are any of its descendants
the arcs putting V on the path are “head-to-head”



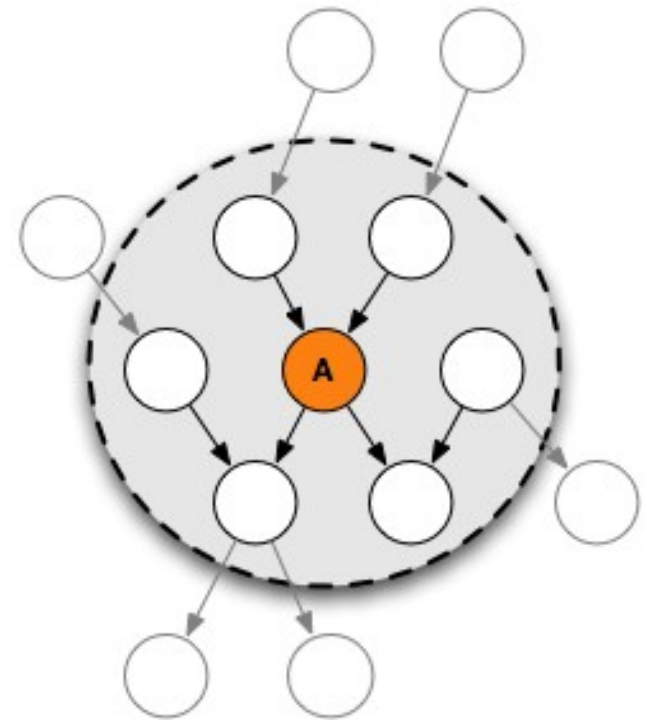
Example



- $I\langle C, \{\}, D\rangle?$ No
- $I\langle C, \{A\}, D\rangle?$ No
- $I\langle C, \{A, B\}, D\rangle?$ Yes
- $I\langle C, \{A, B, J\}, D\rangle?$ No
- $I\langle C, \{A, B, E, J\}, D\rangle?$ Yes

Markov Blanket

- The **Markov blanket** of a variable V is any subset of variables S such that $I\langle\{V\}, S, U - S - \{V\}\rangle$
- Given a variable V , the set of its parents, its children, and its children's other parents is a Markov blanket for V
- Can be checked using d-separation



Inference with Bayesian Networks

- With a Bayesian Network we save space, do we also save time?

- Do we have to use the formula

$$P(\mathbf{q}|\mathbf{e}) = \frac{\sum_{\mathbf{x}, \mathbf{X} = U \setminus Q \setminus E} P(\mathbf{x}, \mathbf{q}, \mathbf{e})}{\sum_{\mathbf{y}, \mathbf{Y} = U \setminus E} P(\mathbf{y}, \mathbf{e})}$$

- to compute $P(\mathbf{q}|\mathbf{e})$?

Inference with Bayesian Networks

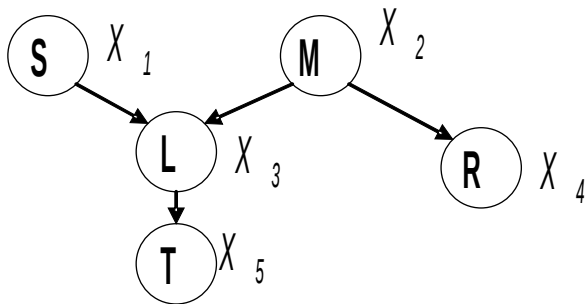
- There are quicker algorithms
 - Exact methods for polytrees
 - Belief propagation
 - Exact methods for general networks
 - Junction tree
 - Variable elimination
 - Approximate methods for general networks:
 - Stochastic simulation
 - Loopy belief propagation
 - Variational methods

Complexity of Inference

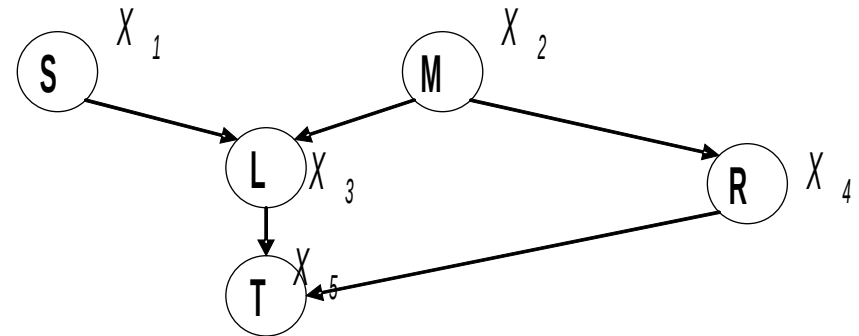
- Exact inference with BN is #P-complete
- #P-complete: a special case of NP-complete problems
 - The answer to a #P-complete problem is the number of solutions to a NP-complete problem

Polytrees

A polytree is a directed acyclic graph in which no two nodes have more than one path between them.



A polytree



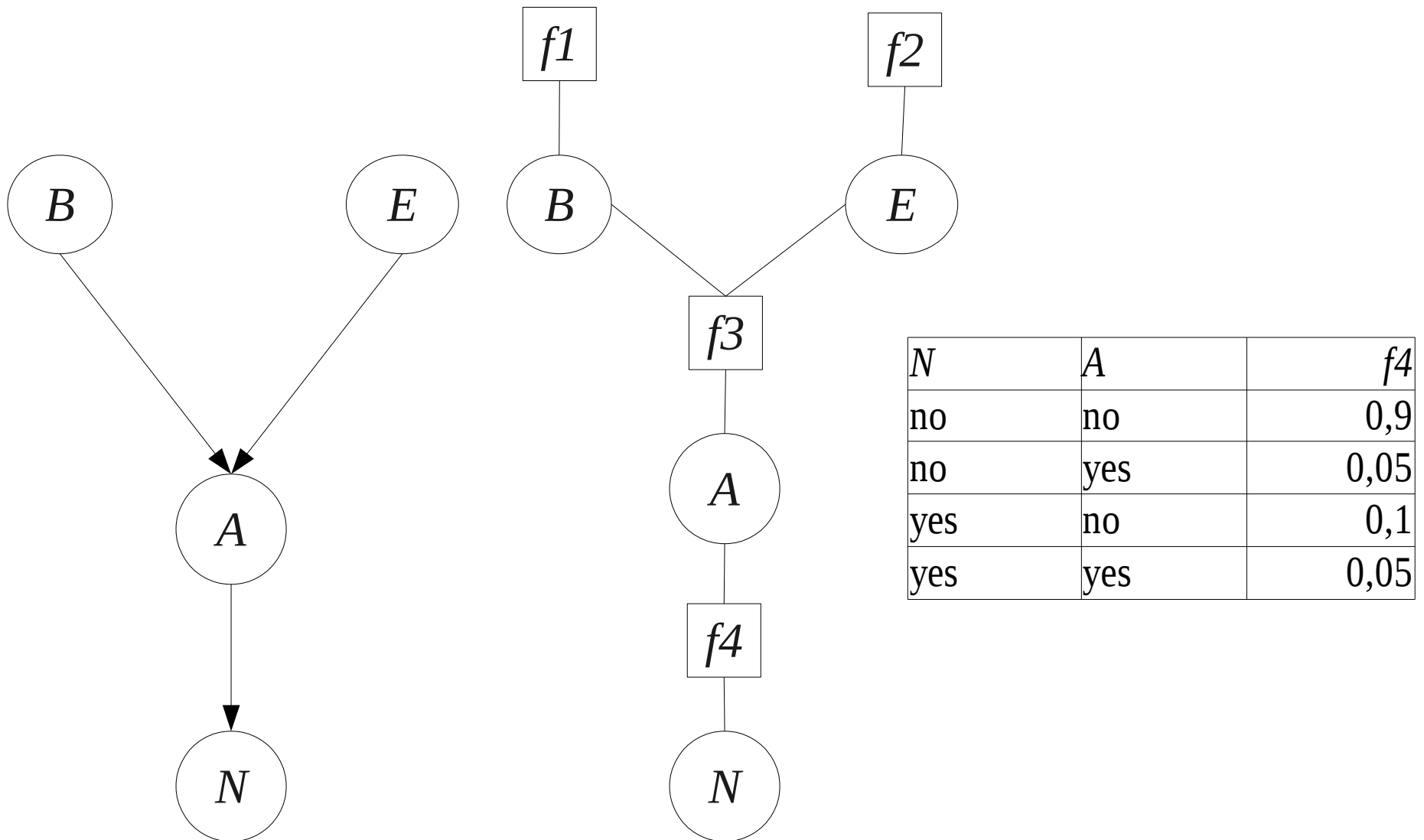
Not a polytree

- i.e. There are no cycles in the corresponding undirected graph

Belief Propagation [Pearl 88]

- Best presented over Factor Graphs
- A **Factor Graph** is a bipartite graph (V, F, E) where vertices V index the variables, the vertices F index the families (factors), and edges E connect V and F
- A factor, given the values of the variables involved in the factor, returns a non-negative number.
- A family in a BN can be seen as a factor

Example Network



Messages

- The message from a variable node X to a neighbor factor node f is

$$\mu_{X \rightarrow f}(x) = \prod_{h \in \text{nb}(X) \setminus f} \mu_{h \rightarrow X}(x)$$

- where $\text{nb}(X)$ is the set of neighbor of X , the set of factors X appears in
- The message from a factor to a variable is

$$\mu_{f \rightarrow X}(x) = \sum_{\neg \{X\}} (f(\mathbf{x}) \prod_{Y \in \text{nb}(f) \setminus X} \mu_{Y \rightarrow f}(y))$$

- Where $\text{nb}(f)$ is the set of arguments of f and the sum is over all of these except X

Belief

- The un-normalized belief of each variable X_i in iteration k can be computed from the equation

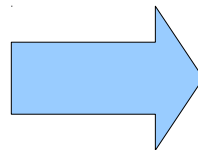
$$b_i(x_i) = \prod_{f \in nb(X_i)} \mu_{f \rightarrow X_i}(x_i)$$

- For example, if X_1 has 3 values x_{11} , x_{12} , x_{13} , their probabilities are
- $B = b_1(x_{11}) + b_1(x_{12}) + b_1(x_{13})$
- $P(x_{11}) = b_1(x_{11})/B$ $P(x_{12}) = b_1(x_{12})/B$ $P(x_{13}) = b_1(x_{13})/B$

Incorporation of Evidence

- For each factor f , for each combination of values of the arguments that is incompatible with the evidence, $f(\mathbf{x})$ is set to 0
- Example: evidence $N=\text{yes}$, factor f_4 becomes

N	A	f_4
no	no	0,9
no	yes	0,05
yes	no	0,1
yes	yes	0,05



N	A	f_4
no	no	0
no	yes	0
yes	no	0,1
yes	yes	0,05

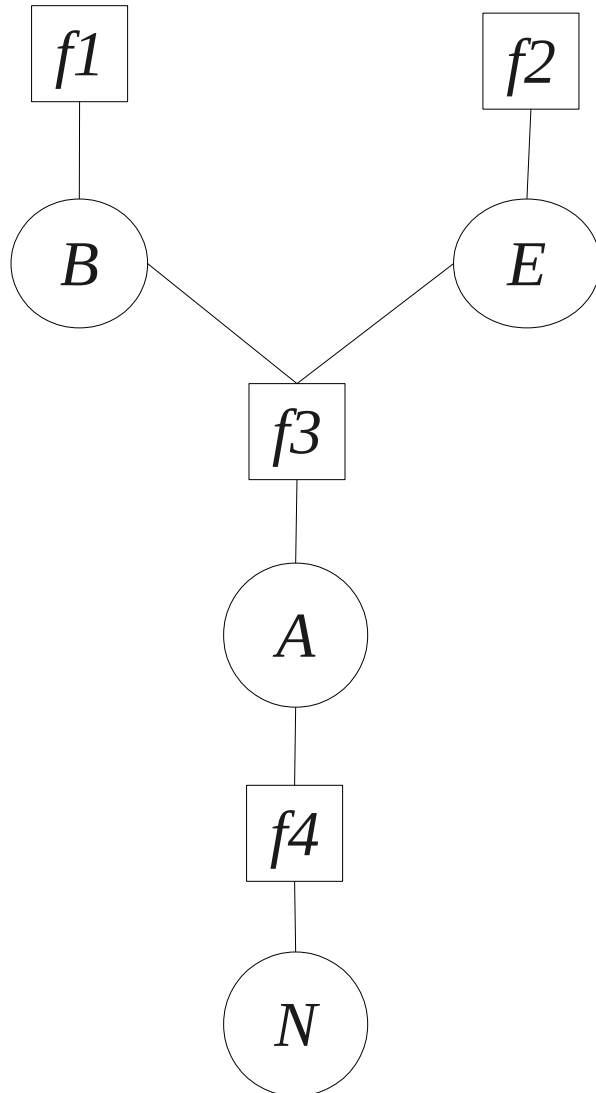
Algorithm

- Initialize all messages to 1 or randomly
- Loop
 - Select an arc
 - Compute the value of the message on the arc
- Until the messages do not change anymore
- If the network is a polytree, this algorithm converges
- Various strategies for selecting the arc to update

Message schedules

- The order in which messages are updated
- Asynchronous schedules: messages are updated sequentially, one arc at a time
- Synchronous schedules: all messages are updated in parallel.
- Flooding (asynchronous): messages are passed from the roots to the sinks and back at each step
- The most widely used and generally best-performing method

Example: Alarm with Binary Variables



B	$f1$
no	0,9
yes	0,1

E	$f2$
no	0,95
yes	0,05

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

N	A	$f4$
no	no	0,9
no	yes	0,05
yes	no	0,1
yes	yes	0,95

Example

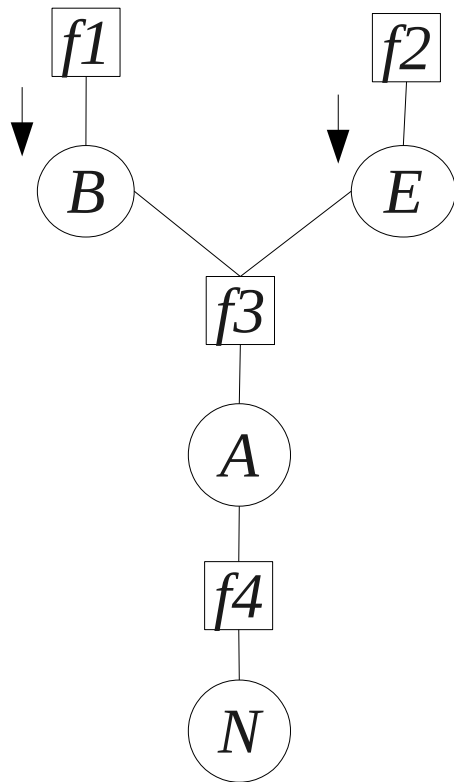
- We want to compute $P(B=\text{yes}|N=\text{yes},E=\text{no})=P(b_T|n_T,e_F)$
- Incorporation of Evidence: we have to modify f2 and f4

E	$f2$
no	0.95
yes	0

N	A	$f4$
no	no	0
no	yes	0
yes	no	0.1
yes	yes	0.95

Example of Belief Propagation

- All messages initially equal to 1



$$\mu_{f1 \rightarrow B}(b) = \sum_{\neg\{B\}} (f1(b) \prod_{Y \in nb(f1) \setminus B} \mu_{Y \rightarrow f1}(y))$$

$$= f1(b)$$

$$\mu_{f1 \rightarrow B}(b_T) = 0,1$$

$$\mu_{f1 \rightarrow B}(b_F) = 0,9$$

$$\mu_{f1 \rightarrow B} = [0,1,0,9]$$

B	f1
no	0,9
yes	0,1

$$\mu_{f2 \rightarrow E}(e) = \sum_{\neg\{E\}} (f2(e) \prod_{Y \in nb(f2) \setminus E} \mu_{Y \rightarrow f2}(y))$$

$$= f2(e)$$

$$\mu_{f2 \rightarrow E} = [0,0,95]$$

E	f2
no	0,95
yes	0

Example

$$\mu_{f1 \rightarrow B} = [0, 1, 0, 9]$$

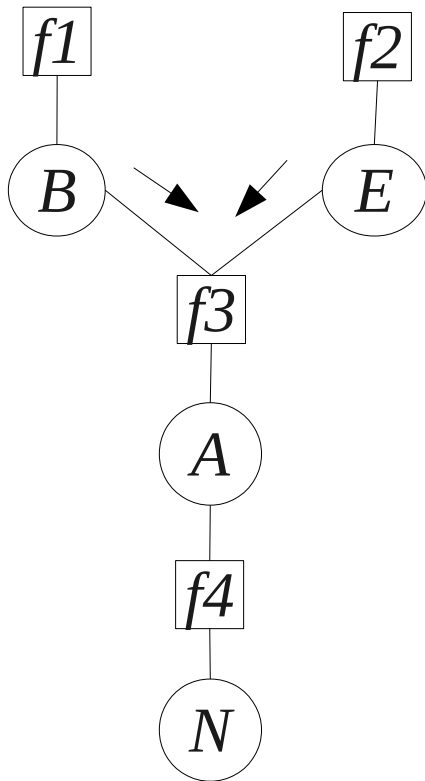
$$\mu_{B \rightarrow f3}(b) = \prod_{h \in nb(B) \setminus f3} \mu_{h \rightarrow B}(b) = \mu_{f1 \rightarrow B}(b)$$

$$\mu_{B \rightarrow f3} = [0, 1, 0, 9]$$

$$\mu_{f2 \rightarrow E} = [0, 0, 95]$$

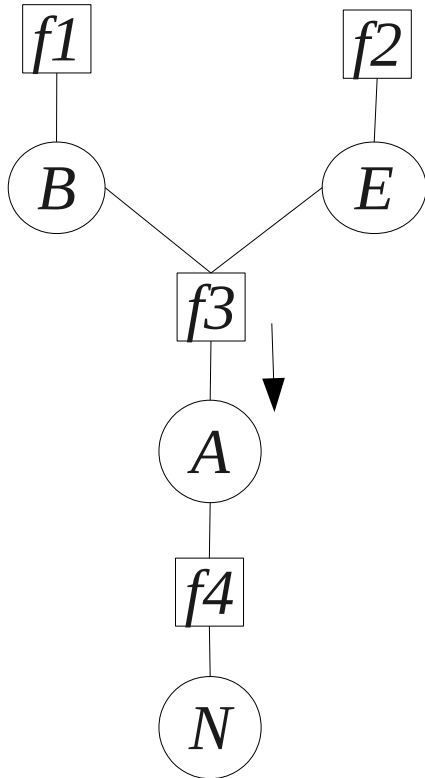
$$\mu_{E \rightarrow f3}(e) = \prod_{h \in nb(E) \setminus f3} \mu_{h \rightarrow E}(e) = \mu_{f2 \rightarrow E}(e)$$

$$\mu_{E \rightarrow f3} = [0, 0, 95]$$



Example

$$\begin{aligned}\mu_{f3 \rightarrow A}(a) &= \sum_{\neg\{A\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus A} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{b, e} f3(b, e, a) \mu_{E \rightarrow f3}(e) \mu_{B \rightarrow f3}(b)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow A}(a_T) &= \\ &= f3(b_T, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_T) \\ &+ f3(b_F, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_F) \\ &+ f3(b_T, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_T) \\ &+ f3(b_F, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_F)\end{aligned}$$

Example

$$\mu_{B \rightarrow f3} = [0,1,0,9]$$

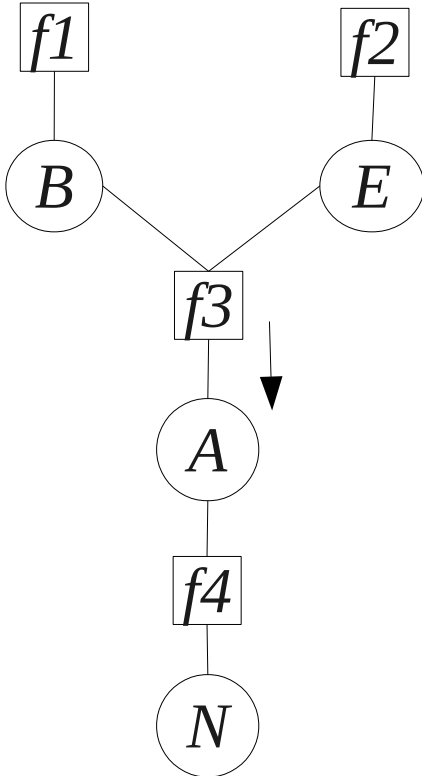
$$\mu_{E \rightarrow f3} = [0,0,95]$$

<i>B</i>	<i>E</i>	<i>A</i>	<i>f3</i>
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

$$\begin{aligned}
 \mu_{f3 \rightarrow A}(a_T) &= \\
 &= f3(b_T, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_T) \\
 &+ f3(b_F, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_F) \\
 &+ f3(b_T, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_T) \\
 &+ f3(b_F, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_F) \\
 &= 0,01 * 0 * 0,1 + 0,85 * 0 * 0,9 \\
 &+ 0,9 * 0,95 * 0,1 + 0,1 * 0,95 * 0,9 \\
 &= 0,171
 \end{aligned}$$

Example

$$\begin{aligned}\mu_{f3 \rightarrow A}(a) &= \sum_{\neg\{A\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus A} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{b, e} f3(b, e, a) \mu_{E \rightarrow f3}(e) \mu_{B \rightarrow f3}(b)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow A}(a_F) &= \\ &= f3(b_T, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_T) \\ &\quad + f3(b_F, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_F) \\ &\quad + f3(b_T, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_T) \\ &\quad + f3(b_F, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_F)\end{aligned}$$

Example

$$\mu_{B \rightarrow f3} = [0,1, 0,9]$$

$$\mu_{E \rightarrow f3} = [0, 0,95]$$

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

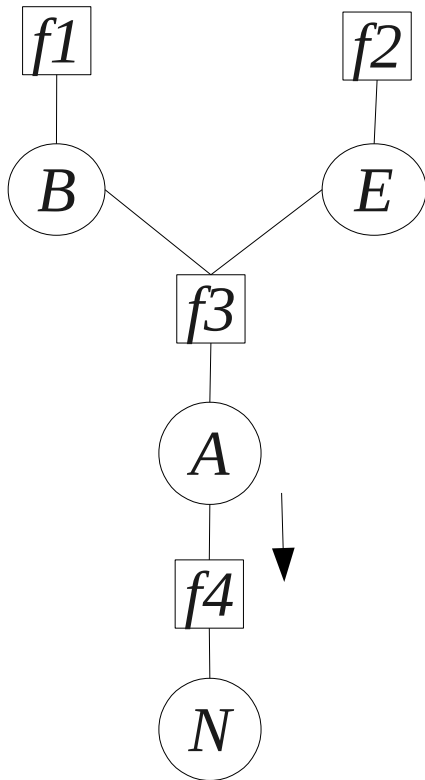
$$\begin{aligned}
 \mu_{f3 \rightarrow A}(a_F) &= \\
 &= f3(b_T, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_T) \\
 &+ f3(b_F, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{B \rightarrow f3}(b_F) \\
 &+ f3(b_T, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_T) \\
 &+ f3(b_F, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{B \rightarrow f3}(b_F) \\
 &= 0,99 * 0 * 0,1 + 0,15 * 0 * 0,9 \\
 &+ 0,1 * 0,95 * 0,1 + 0,9 * 0,95 * 0,9 \\
 &= 0,779
 \end{aligned}$$

Example

- So overall

$$\mu_{f3 \rightarrow A} = [0,171, 0,779]$$

Example



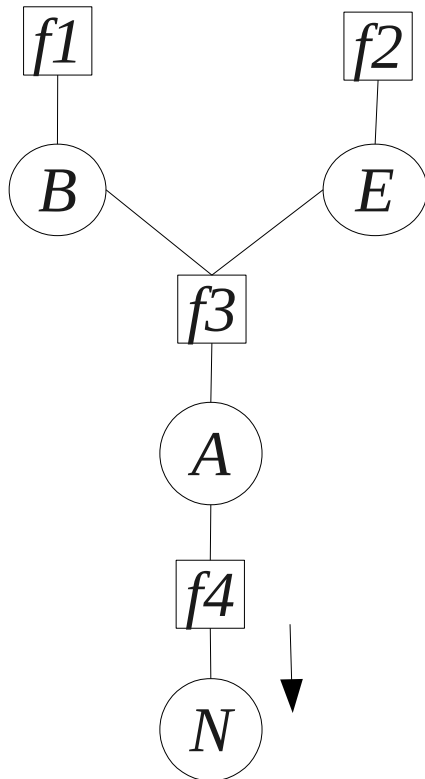
$$\mu_{A \rightarrow f4}(a) = \prod_{h \in nb(A) \setminus f4} \mu_{h \rightarrow A}(a) = \mu_{f3 \rightarrow A}(a)$$

$$\mu_{A \rightarrow f4} = [0,171, 0,779]$$

Example

$$\mu_{f4 \rightarrow N}(n) = \sum_{\neg\{N\}} (f4(a, n) \prod_{Y \in nb(f4) \setminus N} \mu_{Y \rightarrow f4}(y))$$

$$\begin{aligned} \mu_{f4 \rightarrow N}(n_T) &= \sum_a f4(a, n_T) \mu_{A \rightarrow f4}(a) \\ &= f4(a_T, n_T) \mu_{A \rightarrow f4}(a_T) + f4(a_F, n_T) \mu_{A \rightarrow f4}(a_F) \end{aligned}$$



$$\begin{aligned} \mu_{f4 \rightarrow N}(n_F) &= \sum_a f4(a, n_F) \mu_{A \rightarrow f4}(a) \\ &= f4(a_T, n_F) \mu_{A \rightarrow f4}(a_T) + f4(a_F, n_F) \mu_{A \rightarrow f4}(a_F) \end{aligned}$$

Example

$$\mu_{A \rightarrow f4} = [0,171, 0,779]$$

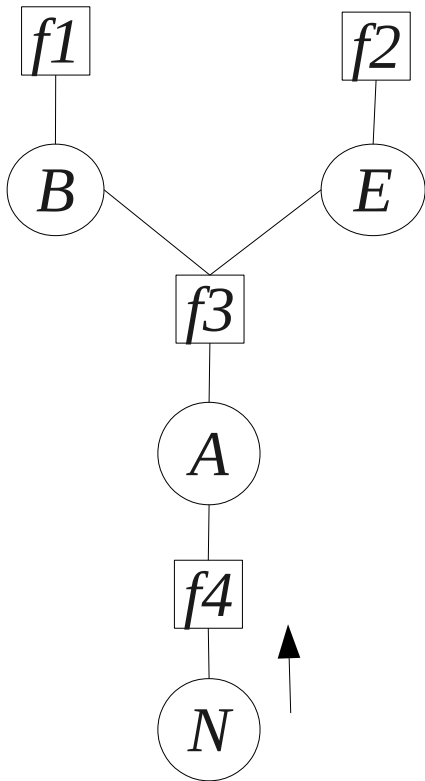
N	A	$f4$
no	no	0
no	yes	0
yes	no	0.1
yes	yes	0.95

$$\begin{aligned} \mu_{f4 \rightarrow N}(n_T) &= \\ &= f4(a_T, n_T) \mu_{A \rightarrow f4}(a_T) + f4(a_F, n_T) \mu_{A \rightarrow f4}(a_F) \\ &= 0,95 * 0,171 + 0,1 * 0,779 = 0,24035 \end{aligned}$$

$$\begin{aligned} \mu_{f4 \rightarrow N}(n_F) &= \\ &= f4(a_T, n_F) \mu_{A \rightarrow f4}(a_T) + f4(a_F, n_F) \mu_{A \rightarrow f4}(a_F) \\ &= 0 * 0,171 + 0 * 0,779 = 0 \end{aligned}$$

$$\mu_{f4 \rightarrow N} = [0,24035, 0]$$

Example



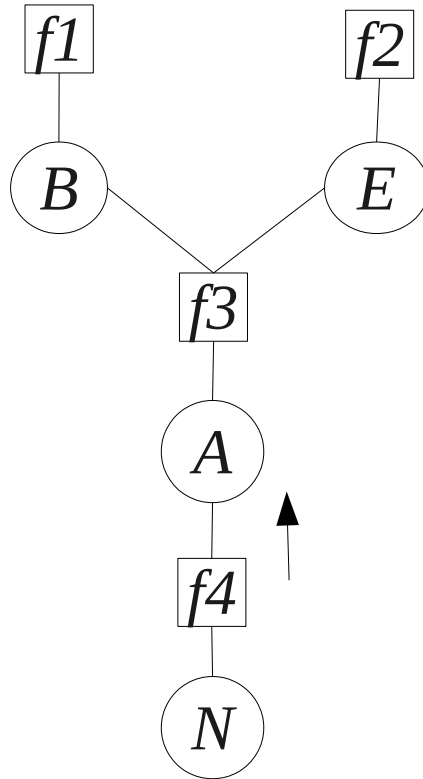
$$\mu_{N \rightarrow f4}(n) = \prod_{h \in nb(N) \setminus f4} \mu_{h \rightarrow N}(n)$$

$$\mu_{N \rightarrow f4}(n) = 1$$

$$\mu_{N \rightarrow f4} = [1, 1]$$

Example

$$\mu_{f4 \rightarrow A}(a) = \sum_{\neg\{A\}} (f4(a, n) \prod_{Y \in nb(f4) \setminus A} \mu_{Y \rightarrow f4}(y))$$



$$\begin{aligned} \mu_{f4 \rightarrow a_T}(a_T) &= f4(a_T, n_T) \mu_{N \rightarrow f4}(n_T) \\ &+ f4(a_T, n_F) \mu_{N \rightarrow f4}(n_F) \end{aligned}$$

$$\begin{aligned} \mu_{f4 \rightarrow A}(a_F) &= f4(a_F, n_T) \mu_{N \rightarrow f4}(n_T) \\ &+ f4(a_F, n_F) \mu_{N \rightarrow f4}(n_F) \end{aligned}$$

Example

$$\mu_{N \rightarrow f4} = [1, 1]$$

N	A	$f4$
no	no	0
no	yes	0
yes	no	0.1
yes	yes	0.95

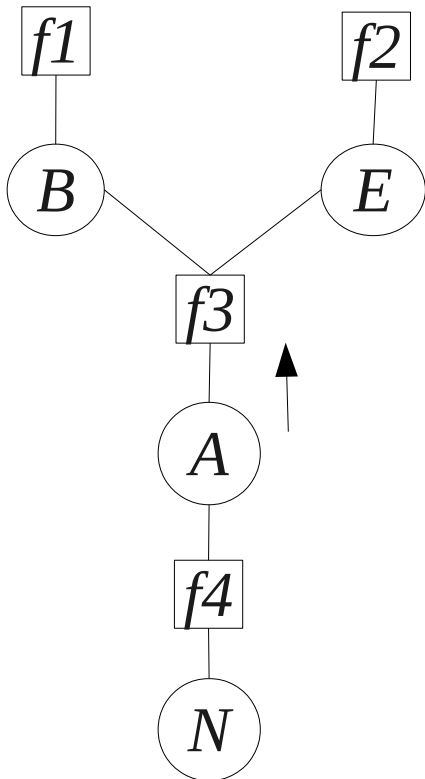
$$\begin{aligned} \mu_{f4 \rightarrow a_T}(a_T) &= f4(a_T, n_T) \mu_{N \rightarrow f4}(n_T) + f4(a_T, n_F) \mu_{N \rightarrow f4}(n_F) \\ &= 0.95 * 1 + 0 * 1 = 0.95 \end{aligned}$$

$$\begin{aligned} \mu_{f4 \rightarrow A}(a_F) &= f4(a_F, n_T) \mu_{N \rightarrow f4}(n_T) + f4(a_F, n_F) \mu_{N \rightarrow f4}(n_F) \\ &= 0.1 * 1 + 0 * 1 = 0.1 \end{aligned}$$

$$\mu_{f4 \rightarrow A} = [0.95, 0.1]$$

Example

$$\mu_{f4 \rightarrow A} = [0,95,0,1]$$

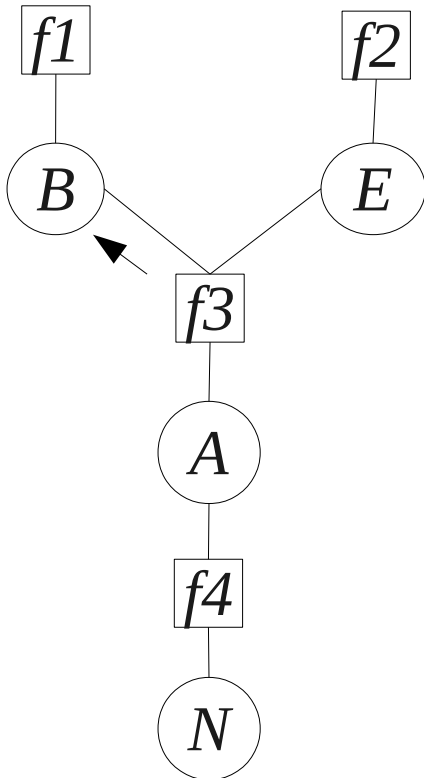


$$\mu_{A \rightarrow f3}(a) = \prod_{h \in nb(A) \setminus f3} \mu_{h \rightarrow A}(a) = \mu_{f4 \rightarrow A}(a)$$

$$\mu_{A \rightarrow f3} = [0,95,0,1]$$

Example

$$\begin{aligned}\mu_{f3 \rightarrow B}(b) &= \sum_{\neg\{B\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus B} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{a, e} f3(b, e, a) \mu_{E \rightarrow f3}(e) \mu_{A \rightarrow f3}(a)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow b_T}(b_T) &= f3(b_T, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_T, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_F) \\ &\quad + f3(b_T, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_T, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_F)\end{aligned}$$

Example

$$\mu_{A \rightarrow f3} = [0,95,0,1]$$

$$\mu_{E \rightarrow f3} = [0,0,95]$$

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

$$\mu_{f3 \rightarrow b_T}(b_T)$$

$$= f3(b_T, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_T)$$

$$+ f3(b_T, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_F)$$

$$+ f3(b_T, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_T)$$

$$+ f3(b_T, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_F)$$

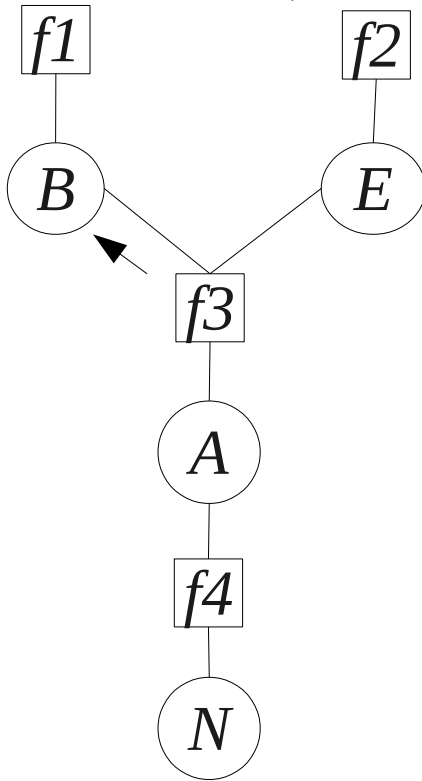
$$= 0,01 * 0 * 0,95 + 0,99 * 0 * 0,1$$

$$+ 0,9 * 0,95 * 0,95 + 0,1 * 0,95 * 0,1$$

$$= 0,82175$$

Example

$$\begin{aligned}\mu_{f3 \rightarrow B}(b) &= \sum_{\neg\{B\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus B} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{a, e} f3(b, e, a) \mu_{E \rightarrow f3}(e) \mu_{A \rightarrow f3}(a)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow B}(b_F) &= f3(b_F, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_F, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_F) \\ &\quad + f3(b_F, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_F, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_F)\end{aligned}$$

Example

$$\mu_{A \rightarrow f3} = [0,95,0,1]$$

$$\mu_{E \rightarrow f3} = [0,0,95]$$

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

$$\begin{aligned}
 &\mu_{f3 \rightarrow B}(b_F) \\
 &= f3(b_F, e_T, a_T) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_T) \\
 &+ f3(b_F, e_T, a_F) \mu_{E \rightarrow f3}(e_T) \mu_{A \rightarrow f3}(a_F) \\
 &+ f3(b_F, e_F, a_T) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_T) \\
 &+ f3(b_F, e_F, a_F) \mu_{E \rightarrow f3}(e_F) \mu_{A \rightarrow f3}(a_F) \\
 &= 0,85 * 0 * 0,95 + 0,15 * 0 * 0,1 \\
 &+ 0,1 * 0,95 * 0,95 + 0,9 * 0,95 * 0,1 \\
 &= 0,17575
 \end{aligned}$$

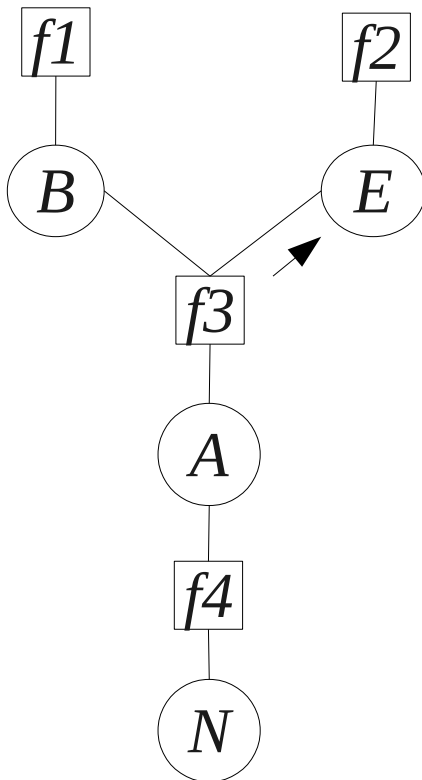
Example

- So overall

$$\mu_{f3 \rightarrow B} = [0,82175, 0,17575]$$

Example

$$\begin{aligned}\mu_{f3 \rightarrow E}(e) &= \sum_{\neg\{E\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus E} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{a, b} f3(b, e, a) \mu_{B \rightarrow f3}(b) \mu_{A \rightarrow f3}(a)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow E}(e_T) &= f3(b_T, e_T, a_T) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_T, e_T, a_F) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_F) \\ &\quad + f3(b_F, e_T, a_T) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_F, e_T, a_F) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_F)\end{aligned}$$

Example

$$\mu_{B \rightarrow f3} = [0,1,0,9]$$

$$\mu_{A \rightarrow f3} = [0,95,0,1]$$

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

$$\mu_{f3 \rightarrow E}(e_T)$$

$$= f3(b_T, e_T, a_T) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_T)$$

$$+ f3(b_T, e_T, a_F) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_F)$$

$$+ f3(b_F, e_T, a_T) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_T)$$

$$+ f3(b_F, e_T, a_F) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_F)$$

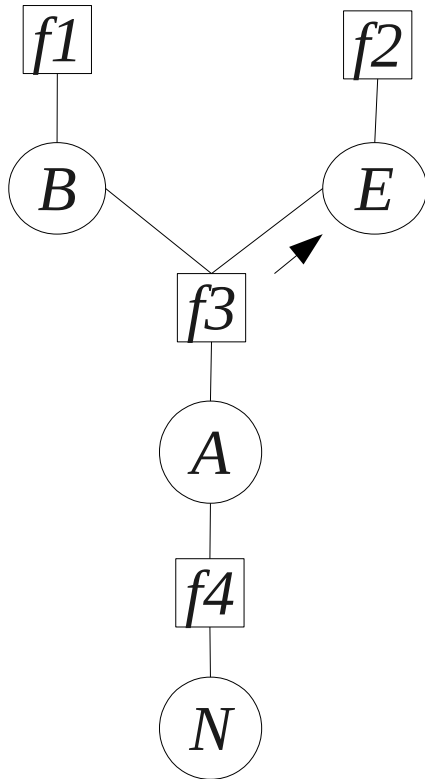
$$= 0,01 * 0,1 * 0,95 + 0,99 * 0,1 * 0,1$$

$$+ 0,85 * 0,9 * 0,95 + 0,15 * 0,9 * 0,1$$

$$= 0,7511$$

Example

$$\begin{aligned}\mu_{f3 \rightarrow E}(e) &= \sum_{\neg\{E\}} (f3(b, e, a) \prod_{Y \in nb(f3) \setminus E} \mu_{Y \rightarrow f3}(y)) \\ &= \sum_{a, b} f3(b, e, a) \mu_{B \rightarrow f3}(b) \mu_{A \rightarrow f3}(a)\end{aligned}$$



$$\begin{aligned}\mu_{f3 \rightarrow E}(e_F) &= f3(b_T, e_F, a_T) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_T, e_F, a_F) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_F) \\ &\quad + f3(b_F, e_F, a_T) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_T) \\ &\quad + f3(b_F, e_F, a_F) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_F)\end{aligned}$$

Example

$$\mu_{B \rightarrow f3} = [0,1, 0,9]$$

$$\mu_{A \rightarrow f3} = [0,95,0,1]$$

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

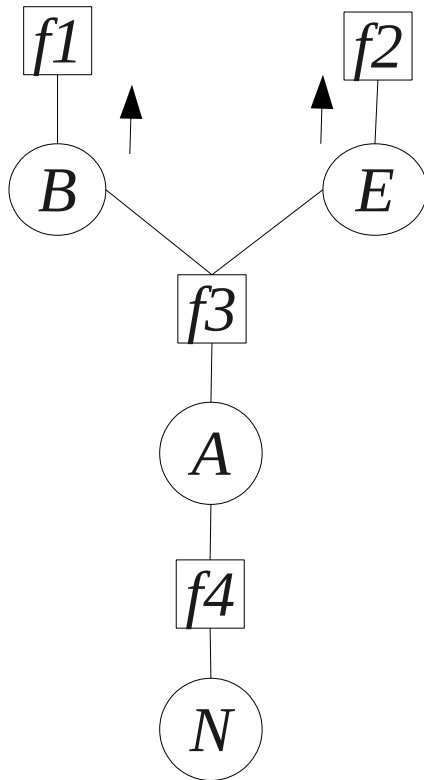
$$\begin{aligned}
 &\mu_{f3 \rightarrow E}(e_F) \\
 &= f3(b_T, e_F, a_T) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_T) \\
 &+ f3(b_T, e_F, a_F) \mu_{B \rightarrow f3}(b_T) \mu_{A \rightarrow f3}(a_F) \\
 &+ f3(b_F, e_F, a_T) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_T) \\
 &+ f3(b_F, e_F, a_F) \mu_{B \rightarrow f3}(b_F) \mu_{A \rightarrow f3}(a_F) \\
 &= 0,9 * 0,1 * 0,95 + 0,1 * 0,1 * 0,1 \\
 &+ 0,1 * 0,9 * 0,95 + 0,9 * 0,9 * 0,1 \\
 &= 0,253
 \end{aligned}$$

Example

- So overall

$$\mu_{f3 \rightarrow E} = [0,7511,0,253]$$

Example



$$\mu_{B \rightarrow f1}(b) = \prod_{h \in nb(B) \setminus f1} \mu_{h \rightarrow B}(b)$$

$$\mu_{B \rightarrow f1}(b) = \mu_{f3 \rightarrow B}(b)$$

$$\mu_{B \rightarrow f1} = [0,82175,0,17575]$$

$$\mu_{E \rightarrow f2}(e) = \prod_{h \in nb(E) \setminus f2} \mu_{h \rightarrow E}(e)$$

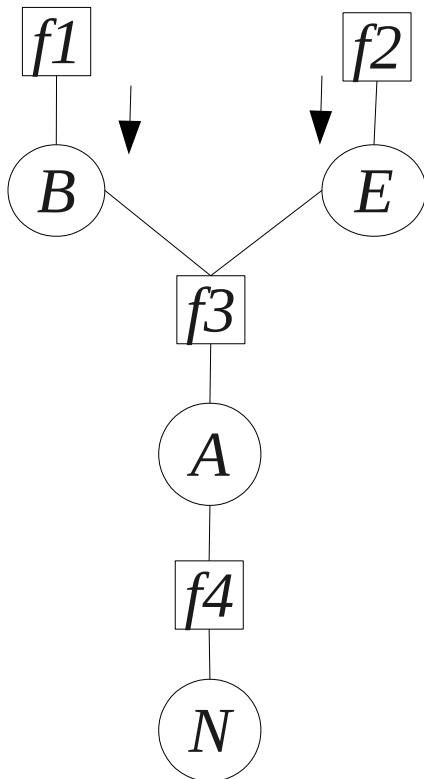
$$\mu_{E \rightarrow f2}(e) = \mu_{f3 \rightarrow E}(e)$$

$$\mu_{E \rightarrow f2} = [0,7511,0,253]$$

Example

$$\mu_{f1 \rightarrow B}(b) = \sum_{\neg\{B\}} (f1(b) \prod_{Y \in nb(f1) \setminus f1} \mu_{Y \rightarrow f1}(y)) = f1(b)$$

$$\mu_{f2 \rightarrow E}(e) = \sum_{\neg\{E\}} (f2(e) \prod_{Y \in nb(f2) \setminus f2} \mu_{Y \rightarrow f2}(y)) = f2(e)$$



.....

- Convergence

Example: Marginal

$$b_B(b) = \prod_{f \in nb(B)} \mu_{f \rightarrow B}(b)$$
$$\mu_{f1 \rightarrow B} = [0, 1, 0, 9]$$
$$\mu_{f3 \rightarrow B} = [0, 82175, 0, 17575]$$
$$b_B(b) = \mu_{f1 \rightarrow B}(b) \mu_{f3 \rightarrow B}(b)$$

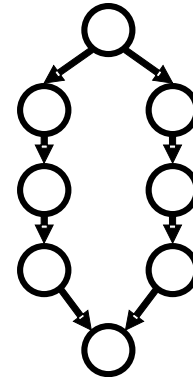
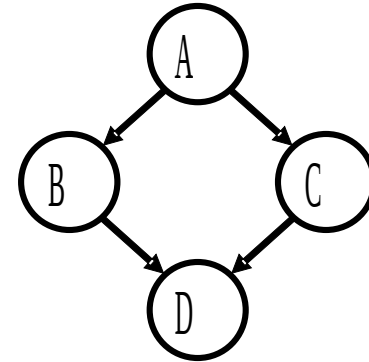
$$b_B(b_T) = \mu_{f1 \rightarrow B}(b_T) \mu_{f3 \rightarrow B}(b_T) = 0,1 * 0,82175 = 0,082175$$

$$b_B(b_F) = \mu_{f1 \rightarrow B}(b_F) \mu_{f3 \rightarrow B}(b_F) = 0,9 * 0,17575 = 0,158175$$

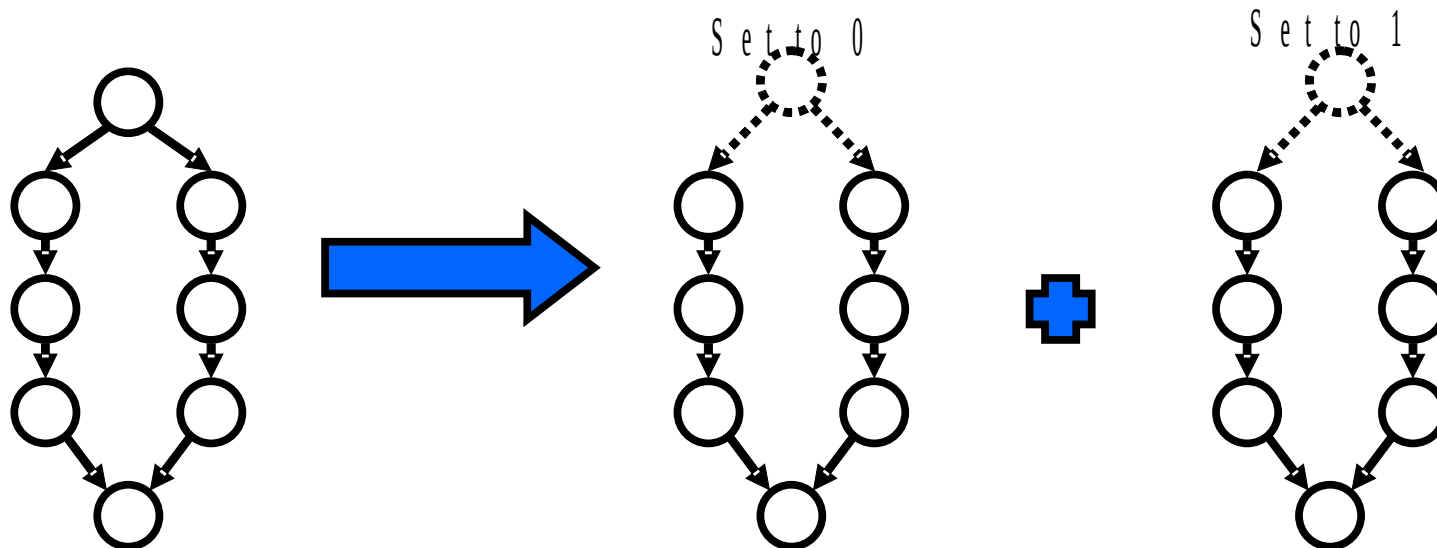
$$P(b_T) = \frac{b_B(b_T)}{b_B(b_T) + b_B(b_F)} = \frac{0,082175}{0,082175 + 0,158175} = 0,3418972$$

General Networks

- Networks that have a cycle in their undirected version
- Three possibilities
 - Conditioning
 - Clustering
 - Approximations

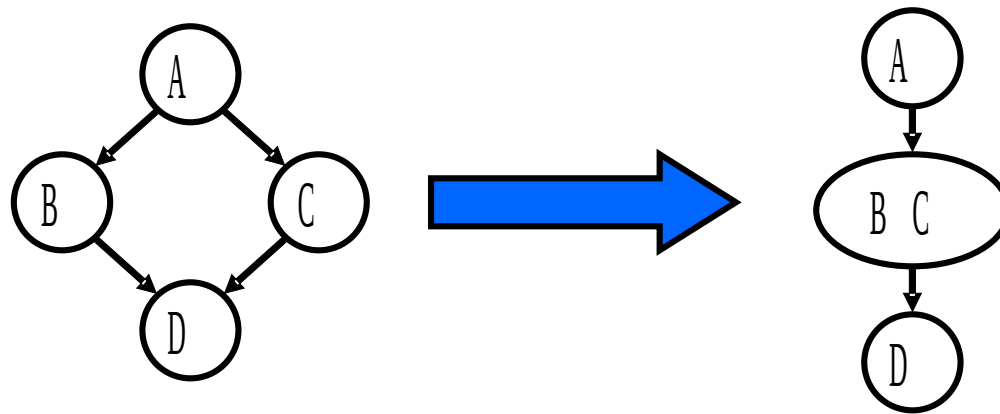


Conditioning



Clustering

- Group together variables so that the resulting network is a polytree and use belief propagation



- Problem: how to find a good clustering?

Join Trees

- Technique for clustering variables
- Steps:
 - Obtain an undirected version of the network
 - Perform a graph operation on it (triangulation)
 - Each clique is a compound variable
 - Add direction to the edges

Junction Tree

- The resulting inference algorithm [Lauritzen, Spiegelhalter 1988] is called
 - Junction tree algorithm (jt), or
 - Clique propagation

Variable Elimination

- Each conditional probability table is seen as a factor $f(x_1, \dots, x_i)$, i.e. functions from variables values to real numbers
- Operations on factors:
 - Product: given $f_1(x_1, \dots, x_i, y_1, \dots, y_j)$ and $f_2(y_1, \dots, y_j, z_1, \dots, z_k)$, the product $f_1 \times f_2$ is defined as
$$(f_1 \times f_2)(x_1, \dots, x_i, y_1, \dots, y_j, z_1, \dots, z_k) = f_1(x_1, \dots, x_i, y_1, \dots, y_j) \times f_2(y_1, \dots, y_j, z_1, \dots, z_k)$$
 - Sum out a variable: sum out x_1 from $f(x_1, \dots, x_i)$
$$(\sum_{x_1} f)(x_1, \dots, x_i) = f(X_1 = x_1^1, \dots, x_i) + \dots + f(X_1 = x_1^m, \dots, x_i)$$

Function sum-out

- Function sum-out(List of factors F, variable z)
 - 1) Remove from F all the factors, say $f_1, \dots, f_k$, that contain z
 - 2) Add the new factor

$$\sum_z \prod_{i=1}^k f_i$$

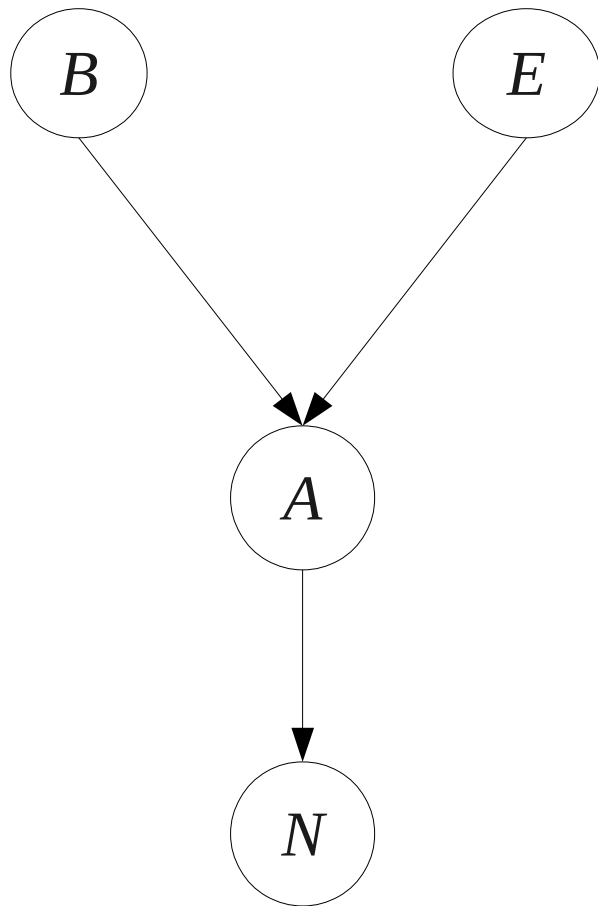
to F

- 3) return F

Variable Elimination

- Function $VE(F, q, e, O)$
 - F set of factors, q query, e evidence, O order of variables in $U \setminus Q \setminus E$
- 1) Set the observed variables E in all factors to their observed value e
 - 2) While O is not empty
 - a) Remove the first variable v from O
 - b) $F = \text{sum-out}(F, v)$
 - 3) Set $h =$ the product of all the factors in F
 - 4) Return $h(q) / \sum_Q h(Q)$

Example



B	$f1$
no	0,9
yes	0,1

E	$f2$
no	0,95
yes	0,05

B	E	A	$f3$
no	no	no	0,9
no	no	yes	0,1
no	yes	no	0,15
no	yes	yes	0,85
yes	no	no	0,1
yes	no	yes	0,9
yes	yes	no	0,99
yes	yes	yes	0,01

N	A	$f4$
no	no	0,9
no	yes	0,05
yes	no	0,1
yes	yes	0,95

Example

- Query: $P(B=\text{yes}|N=\text{yes},E=\text{no})=P(B|N,\sim E)$
- Elimination order $O=\{A\}$
- $VE(\{f1,f2,f3,f4\},B,\{N,\sim E\},\{A\})$

1) Set the observed variables E in all factors to their observed value e

<i>B</i>	<i>f1</i>
no	0,9
yes	0,1

<i>A</i>	<i>f4</i>
no	0,1
yes	0,95

<i>f2</i>
0,95

<i>B</i>	<i>A</i>	<i>f3</i>
no	no	0,9
no	yes	0,1
yes	no	0,1
yes	yes	0,9

Example

- Query: $P(B=\text{yes}|N=\text{yes},E=\text{no})=P(B|N,\sim E)$
- Elimination order $O=\{A\}$
- $VE(\{f1,f2,f3,f4\},B,\{N,\sim E\},\{A\})$

1) Set the observed variables E in all factors to their observed value e

<i>B</i>	<i>f1</i>
no	0,9
yes	0,1

<i>A</i>	<i>f4</i>
no	0,1
yes	0,95

<i>f2</i>
0,95

<i>B</i>	<i>A</i>	<i>f3</i>
no	no	0,9
no	yes	0,1
yes	no	0,1
yes	yes	0,9

Example

a) Remove the first variable v from O : $v=A$

b) $F=\text{sum-out}(F,A)$

1) Remove from F all the factors that contain A : f_4, f_3

A	f_4
no	0,1
yes	0,95

B	A	f_3
no	no	0,9
no	yes	0,1
yes	no	0,1
yes	yes	0,9

2) Add the new factor

$$\sum_a f_3 \times f_4$$

Example

A	f_4
no	0,1
yes	0,95

B	A	f_3
no	no	0,9
no	yes	0,1
yes	no	0,1
yes	yes	0,9

$$f_3 \times f_4$$

B	A	$f_3 \times f_4$
no	no	0,090
no	yes	0,095
yes	no	0,010
yes	yes	0,855

$$\sum_a f_3 \times f_4$$

B	$\sum_a f_3 \times f_4$
no	0,185
yes	0,865

Example

3) Set h = the product of all the factors in F

B	$f1$
no	0,9
yes	0,1

$f2$
0,95

B	$\Sigma_a f3 \times f4$
no	0,185
yes	0,865

B	h
no	0,158175
yes	0,082175

4) Return $h(B)/\Sigma_b h(b) = 0,082175/0,24035 = 0,341897$

Approximate Methods

- Stochastic simulation:
 - Generate N samples from BN
 - Count: N_e : samples that satisfy $\mathbf{e}$, N_{qe} samples that satisfy $\mathbf{q}, \mathbf{e}$
 - $P(\mathbf{q}|\mathbf{e}) = N_{qe} / N_e$
- Loopy belief propagation:
 - bp in networks with cycles
 - Experiments have shown that it converges also in network with cycles, often to good quality solutions

Stochastic Simulation

- Let $X_1, \dots, X_n$ be a topological sort of the variables
- For $i=1$ to n
 - Find parents, if any, of X_i . Call them $X_{p(i,1)}, X_{p(i,2)}, \dots, X_{p(i,p(i))}$
 - Recall the values that those parents were randomly given: $x_{p(i,1)}, x_{p(i,2)}, \dots, x_{p(i,p(i))}$
 - Look up in the cpt for:
$$P(X_i = x_i \mid X_{p(i,1)} = x_{p(i,1)}, X_{p(i,2)} = x_{p(i,2)}, \dots, X_{p(i,p(i))} = x_{p(i,p(i))})$$
 - Randomly choose x_i according to this probability

Problems in Building BN

- Assessing conditional independence is not always easy for humans
- Usually done on the basis of causal information
- Assigning a number to each cpt entry is also difficult for humans

Problems in Building BN

- Often we do not have an expert but we are given a set of observations $D = \{\mathbf{u}^1, \dots, \mathbf{u}^N\}$
- $\mathbf{u}^j$ is an assignment to all the variables $\mathbf{U} = \{X_1, \dots, X_n\}$
- How to infer the parameters and/or the structure from D ?

Learning

- We want to find a BN over $\mathbf{U}$ such that the probability of the data $P(D)$ is maximized
- $P(D)$ is also called the **likelihood** of the data
- We assume that all the samples are **independent and identically distributed** (iid) so

$$P(D) = \prod_{i=1}^N P(\mathbf{u}^i)$$

- Often the natural log of $P(D)$ (**log likelihood**) is considered

$$\log P(D) = \sum_{i=1}^N \log P(\mathbf{u}^i)$$

Learning BN

- Tasks
 - Computing the parameters given a fixed structure or
 - finding the structure and the parameters
- Properties of data:
 - complete data: in each data vectors $\mathbf{u}^j$, the values of all the variables are observed
 - incomplete data

Parameter Learning from Complete Data

- Parameters to be learned

$$\theta_{x_i|\pi_i} = P(x_i|\pi_i)$$

- for all $x_i, \pi_i, i=1,\dots,n$
- The values of the parameters that maximize the likelihood can be computed in closed form

Maximum Likelihood Parameters

- Given by relative frequency
- If N_y be the number of vectors of D where $\mathbf{Y}=\mathbf{y}$.

$$\theta_{x_i|\pi_i} = \frac{N_{x_i, \pi_i}}{N_{\pi_i}}$$

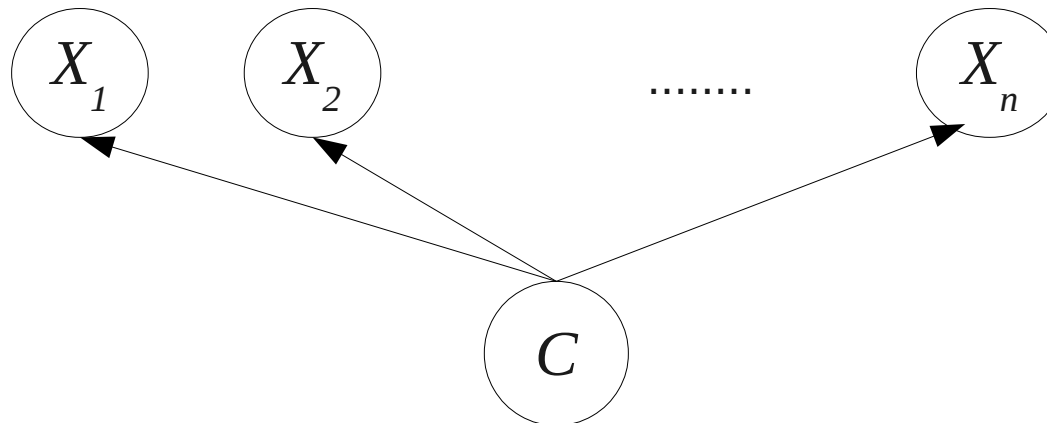
- Counting: for each i , for each value π_i we must collect

$$C_{\pi_i} = \langle N_{x_i^1, \pi_i}, \dots, N_{x_i^{v_i}, \pi_i} \rangle$$

- where v_i is the number of values of X_i

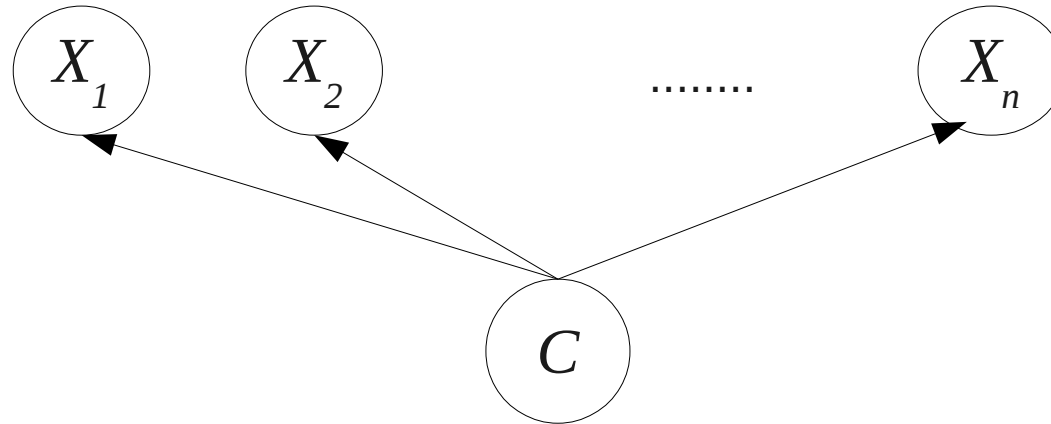
Naive Bayes Special Case

- We want to perform classification
- One variable C represents the class
- The variables $\mathbf{X}$ represent the attributes
- Model:



- X_i independent from X_j given C

Naive Bayes Special Case



- Conditional probability tables (case of Boolean variables):

	C=true	C=false
X_i =true	$P(X_i=\text{true} C=\text{true})$	$P(X_i=\text{true} C=\text{false})$
X_i =false	$P(X_i=\text{false} C=\text{true})$	$P(X_i=\text{false} C=\text{false})$

Example

No	Outlook	Temp	Humid	Windy	Class
D1	sunny	mild	normal	T	P
D2	sunny	hot	high	T	N
D3	sunny	hot	high	F	N
D4	sunny	mild	high	F	N
D5	sunny	cool	normal	F	P
D6	overcast	mild	high	T	P
D7	overcast	hot	high	F	P
D8	overcast	cool	normal	T	P
D9	overcast	hot	normal	F	P
D10	rain	mild	high	T	N
D11	rain	cool	normal	T	N
D12	rain	mild	normal	F	P
D13	rain	cool	normal	F	P
D14	rain	mild	high	F	P

	C=P	C=N
Humid=normal	6/9=0.66666	1/5=0.2
Humid=high	3/9=0.33333	4/5=0.8

Queries

- Computing the probability of a class given values for the attributes: $P(c|x_1, \dots, x_n)$

$$P(c|x_1, \dots, x_n) = \frac{P(c, x_1, \dots, x_n)}{P(x_1 \dots x_n)} = \frac{P(x_1, \dots, x_n|c) P(c)}{P(x_1 \dots x_n)}$$

- Since the attributes are independent given the class

$$P(c|x_1, \dots, x_n) = \frac{P(x_1|c) \dots P(x_n|c) P(c)}{P(x_1 \dots x_n)}$$

Example

- We want to classify <Outlook=sunny,Temp=cool,Humid=high,Windy=T>
- We have to compute
 $P(\text{Class}=P | \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T})$
- We compute only the parameters we need

$$P(\text{Class}=P)=9/14=0.64$$

$$P(\text{Class}=N)=5/14=0.36$$

$$P(\text{Outlook}=\text{sunny} | \text{Class}=P)=2/9=0.222$$

$$P(\text{Outlook}=\text{sunny} | \text{Class}=N)=3/5=0.6$$

$$P(\text{Temp}=\text{cool} | \text{Class}=P)=3/9=0.333$$

$$P(\text{Temp}=\text{cool} | \text{Class}=N)=1/5=0.2$$

$$P(\text{Humid}=\text{high} | \text{Class}=P)=3/9=0.333$$

$$P(\text{Humid}=\text{high} | \text{Class}=N)=4/5=0.8$$

$$P(\text{Windy}=\text{T} | \text{Class}=P)=3/9=0.33$$

$$P(\text{Windy}=\text{T} | \text{Class}=N)=3/5=0.6$$

Example

$$P(\text{Class}=\text{P}, \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) = 0.0053$$

$$P(\text{Class}=\text{N}, \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) = 0.0206$$

- We can compute $P(\text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T})$ by marginalization:

$$P(\text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) =$$

$$P(\text{Class}=\text{P}, \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) +$$

$$P(\text{Class}=\text{N}, \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) =$$

$$0.0053 + 0.0206 = 0.0259$$

- So

$$P(\text{Class}=\text{P} | \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) = 0.0053 / 0.0259 = 0.205$$

$$P(\text{Class}=\text{N} | \text{Outlook}=\text{sunny}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Windy}=\text{T}) = 0.0206 / 0.0259 = 0.795$$

Structure Learning from Complete Data

- Perform a local search in the space of possible structures
- HGC algorithm [Heckerman, Geiger, Chickering 95]:
 - Start with a structure BestG' (possibly empty)
 - Repeat
 - $\text{BestG} = \text{BestG}'$
 - Let $\text{Ref} = \{G \mid G \text{ is obtained from } \text{BestG}' \text{ by adding, deleting or reversing an arc}\}$
 - Let $\text{BestG}' = \arg\max_G \{\text{score}(G) \mid G \in \text{Ref}\}$
 - while $\text{score}(\text{BestG}') - \text{score}(\text{BestG}) > 0$

Structure Score

$$\text{score}(G) = \log P(D|G)$$

$$\begin{aligned} P(D|G) &= \int \rho(D, \Theta|G) d\Theta \\ &= \int P(D|\Theta, G) \rho(\Theta|G) d\Theta \end{aligned}$$

- where

$$\rho(\Theta|G) = \prod_{i, \pi_i} \rho(\theta_{\pi_i})$$

$$\theta_{\pi_i} = \langle \theta_{x_i^1|\pi_i}, \dots, \theta_{x_i^{v_i}|\pi_i} \rangle$$

- and $\rho(\theta_{\pi_i})$ is the prior density of the vector θ_{π_i}

Prior Density of the Parameters

- A common choice for the form of the prior density is the **Dirichlet probability density**
- In this case $\rho(\theta_{\pi_i})$ is described by v_i parameters

$$C'_{\pi_i} = \langle N'_{x_i^1, \pi_i}, \dots, N'_{x_i^{v_i}, \pi_i} \rangle$$

- Prior counts: it is as if we had previously observed some data on which the counts are N'_{x_i, π_i}

Structure Score

- If the priors for the parameters are Dirichlet, then the score is called BD (for Bayesian Dirichlet) and

$$BD(G) = \sum_i BD_i(G)$$

- where $BD_i(G)$ depends only on C_i and C'_i , the counts for the family of X_i

$$C_i = \langle C_{\pi_i^1}, \dots, C_{\pi_i^{r_i}} \rangle$$
$$C'_i = \langle C'_{\pi_i^1}, \dots, C'_{\pi_i^{r_i}} \rangle$$

Structure Score

$$BD_i(G) = \sum_{\pi_i} \log \frac{\Gamma(N_{\pi_i})}{\Gamma(N_{\pi_i} + N'_{\pi_i})} + \sum_{x_i} \log \frac{\Gamma(N_{x_i, \pi_i} + N'_{x_i, \pi_i})}{\Gamma(N_{x_i, \pi_i})}$$

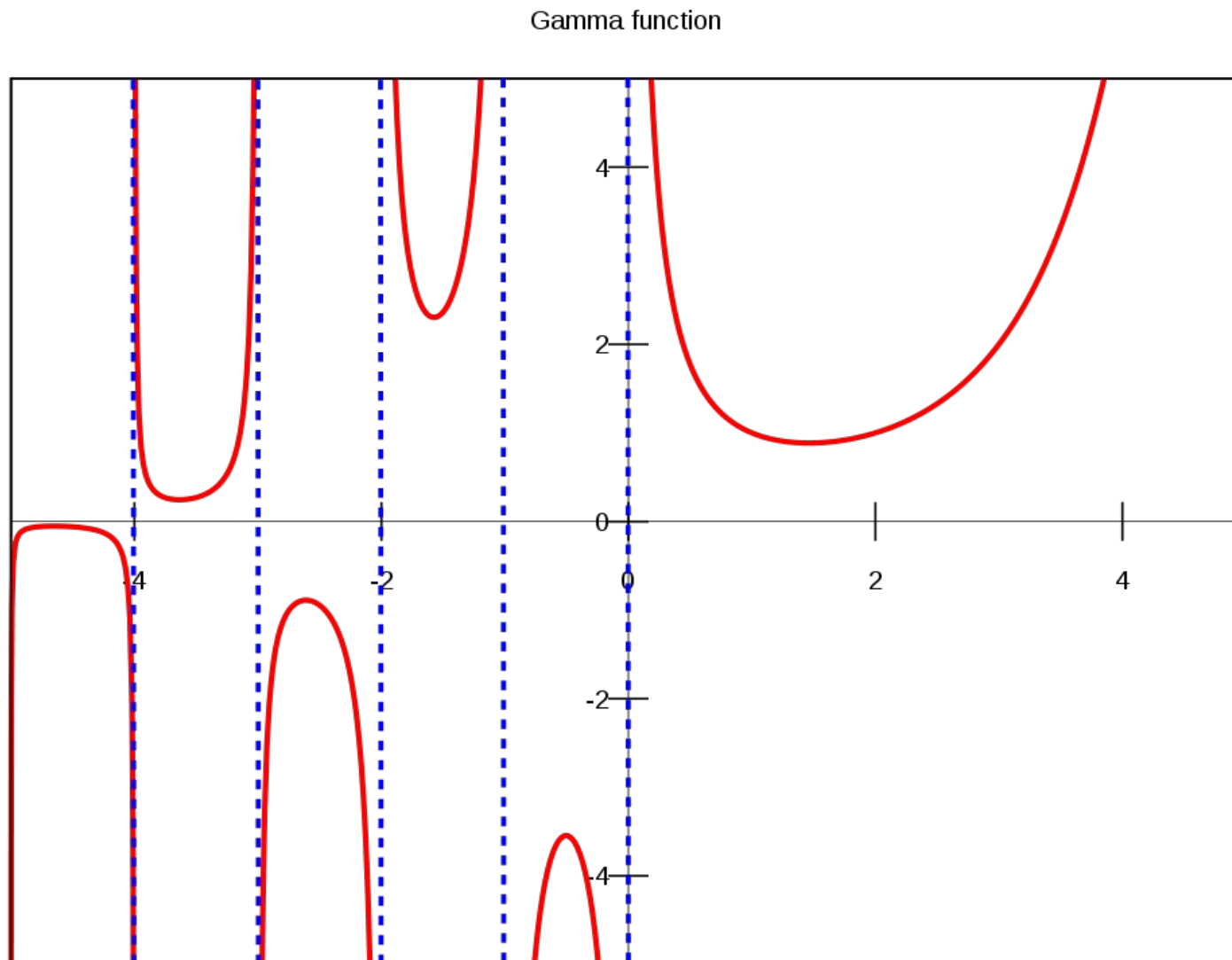
- Where Γ is the Gamma function, an extension of the factorial function with its argument shifted down by 1, to real and complex numbers. That is, if n is a positive integer:

$$\Gamma(n) = (n-1)!$$

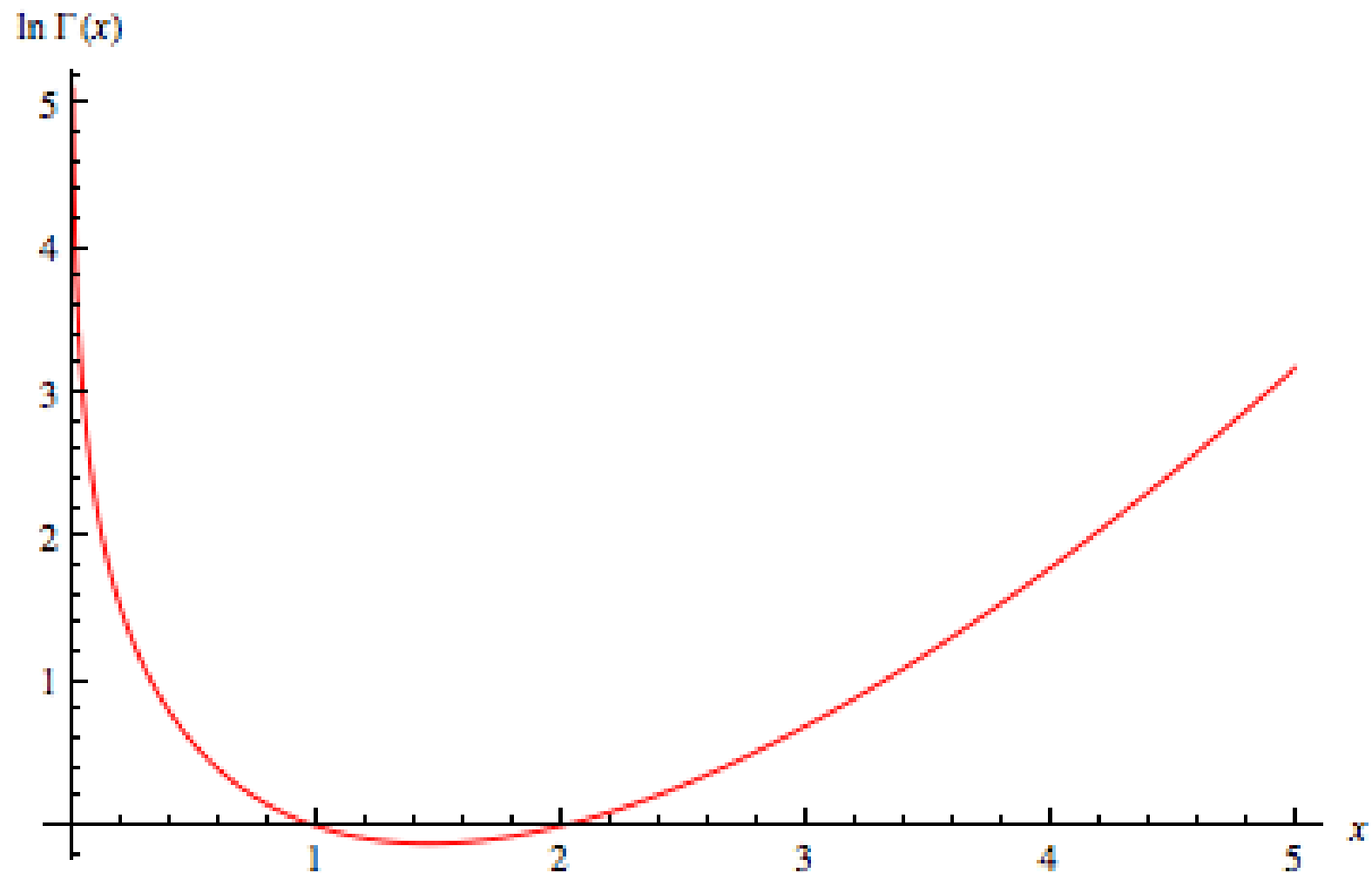
- otherwise

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

Gamma Function



Log Gamma Function



Structure Score

- $BD(G)$ is **decomposable**:
 - It can be computed independently for each family
- Each edge operation involves
 - 1 family (addition, deletion) or
 - 2 families (reversal)
- $BD(G)$ can be quickly computed from $BD(\text{BestG}')$ by changing only the score of the affected families

Parameter Learning from Incomplete Data

- The maximum likelihood parameters cannot be computed in closed form
- An iterative algorithm is necessary: the EM algorithm
- Finds a (possibly) local maximum of the likelihood

EM Algorithm

- Initialize the parameters at random Θ
- Repeat
 - Expectation step:
 - compute the probability $P(y|\mathbf{e})$ of each value y of the missing attributes using (G, Θ) and inference where $\mathbf{e}$ are the values of the observed attributes
 - Compute Θ by maximum likelihood
 - Relative frequency for each family
 - If a variable Y is unobserved in an example $\mathbf{e}$ that matches x_i, π_i , instead of adding 1 to N_{x_i, π_i} we add $P(y|\mathbf{e})$

Structure Learning from Incomplete Data

- There is no decomposable score
- HGC would not be efficient
- Structural EM:
 - Start with a structure BestG' (possibly empty)
 - Repeat
 - $\text{BestG} = \text{BestG}'$
 - Compute the parameters of BestG with EM
 - Optimize a lower bound of the likelihood of the observed data
 - Let BestG' the optimum
 - Until no improvement

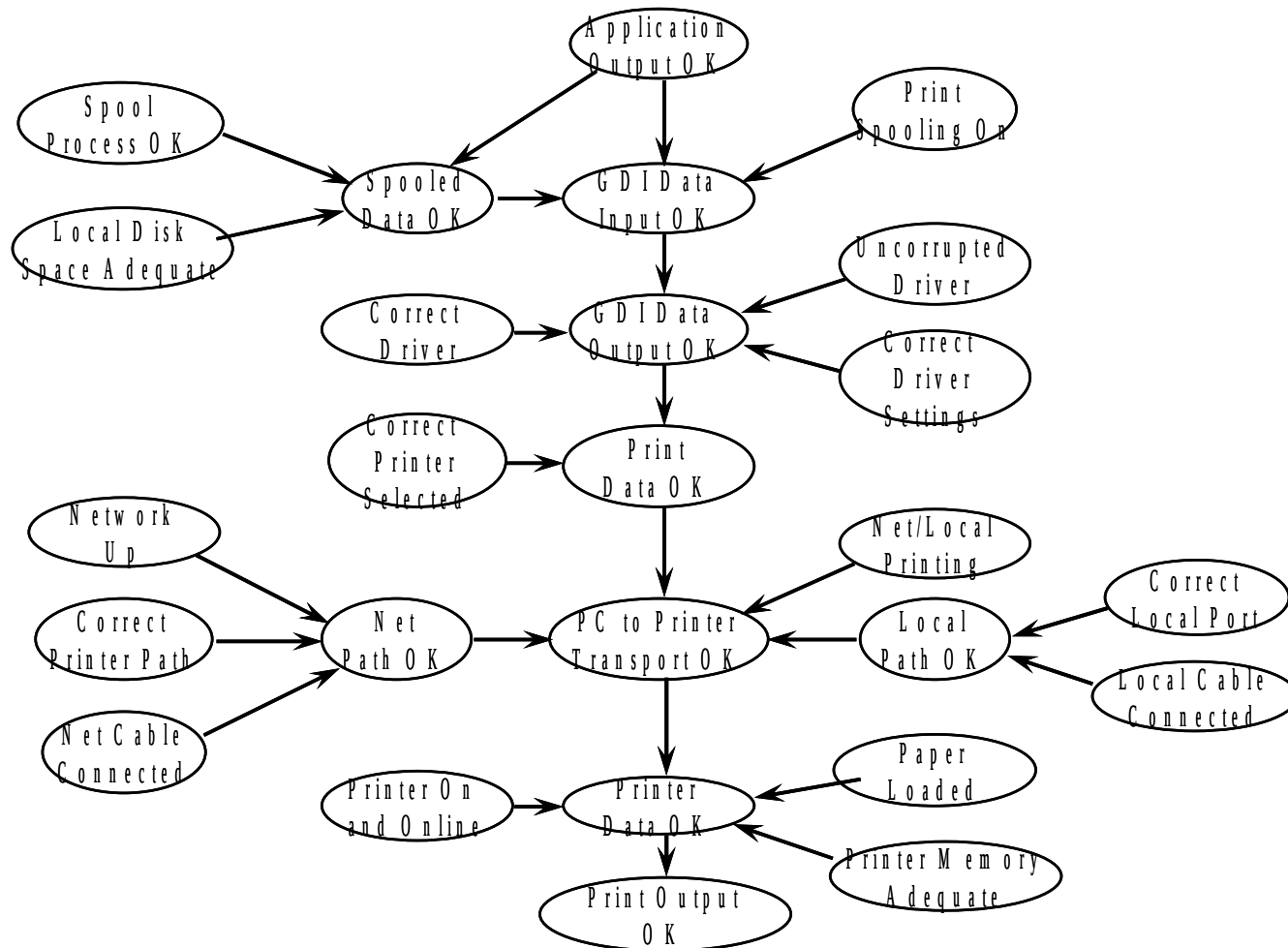
Applications of BN

- Monitoring of emergency care patients
- Model of barley crops yield
- Diagnosis of carpal tunnel syndrome
- Insulin dose adjustment (DBN) in diabetes
- Predicting hails in northern Colorado
- Evaluating insurance applications

Applications of BN

- Deciding on the amount of fungicides to be used against attack of mildew in wheat
- Assisting experts of electromyography
- Pedigree of breeding pigs
- Modeling the biological processes of a water purification plant
- Printer troubleshooting (Microsoft Windows)

Printer Troubleshooting (Windows 95)



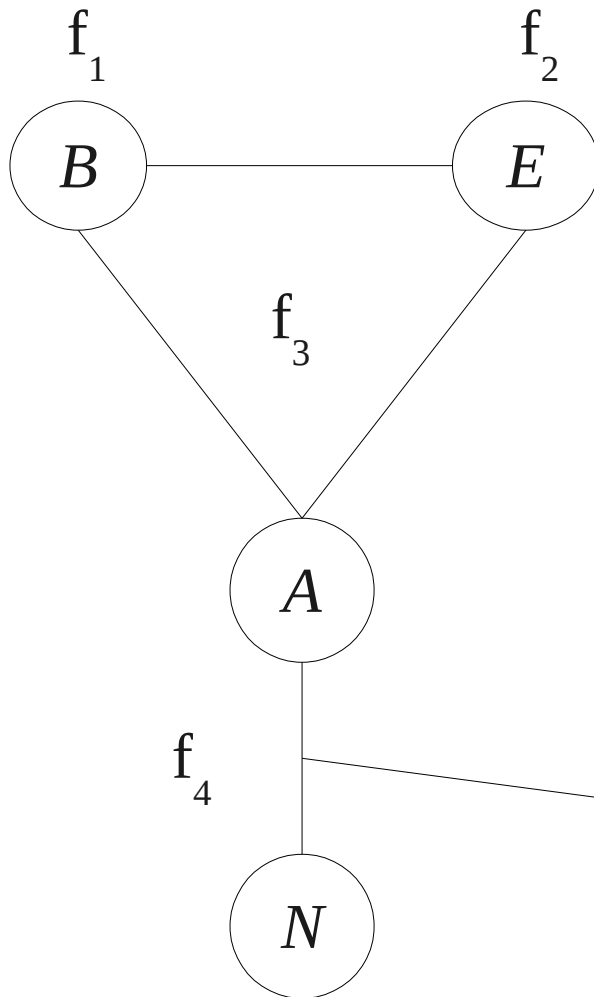
Applications

- Office Assistant in MS Office (“smiley face”)
 - Bayesian network based free-text help facility
 - help based on past experience (keyboard/mouse use) and the task user is currently doing

Markov Networks (MN)

- Alternative to BN
- Also called Random Fields, undirected graphical models
- Undirected graph
- Conditional independence represented by graph separation
- Probability distribution as the product of a set of **potentials** or **factors** (functions of a subset of variables) divided by a normalization constant
- Potentials over cliques

Example



- Four potentials f_1 , f_2 , f_3 and f_4

N	A	f_4
no	no	0,9
no	yes	0,05
yes	no	0,1
yes	yes	0,05

Markov Networks

- Probability

$$P(\mathbf{u}) = \frac{\prod_c f_c(\mathbf{x}_c)}{Z}$$

$$Z = \sum_u \prod_c f_c(\mathbf{x}_c)$$

- Z is called **partition function**, ensures that the probabilities sum to 1

Loglinear Models

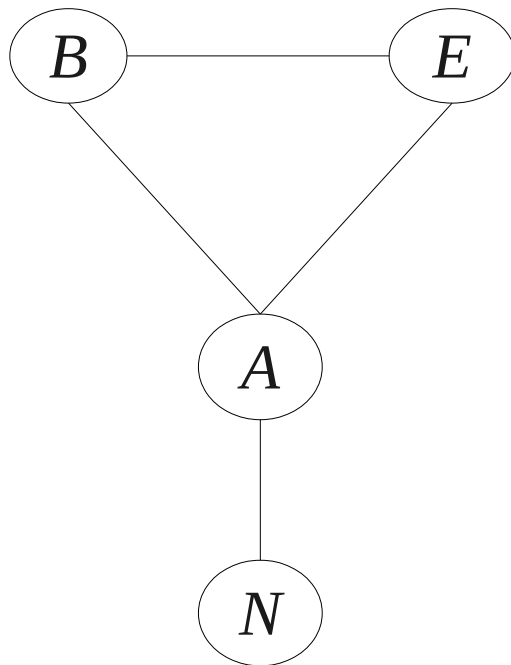
- If all the potentials are >0 , they can be represented as exponential functions, i.e., f_4 can be represented as $f_4 = \exp(w_4 F_4)$
- where F_4 is any real function of f_4 arguments and w_4 is a real weight. Then

$$P(\mathbf{u}) = \frac{\exp \sum_c w_c F_c(\mathbf{x}_c)}{Z}$$

$$Z = \sum_u \exp \sum_c w_c F_c(\mathbf{x}_c)$$

How to tell independence

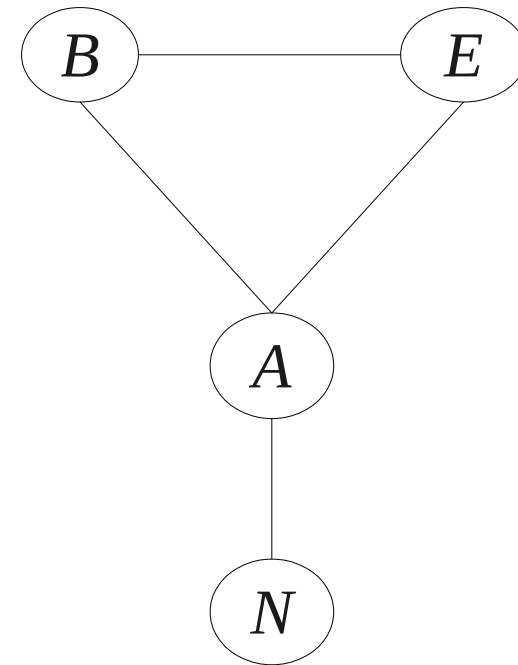
- Definition: $\mathbf{X}$ and $\mathbf{Y}$ are **independent** given a set of variables $\mathbf{Z}$ ($I\langle\mathbf{X},\mathbf{Z},\mathbf{Y}\rangle$) iff every path from $\mathbf{X}$ to $\mathbf{Y}$ passes through a variable of $\mathbf{Z}$



- $I\langle B, \{\}, N \rangle$ No
- $I\langle B, A, N \rangle$ Yes
- $I\langle B, E, N \rangle$ No
- $I\langle \{B, E\}, A, N \rangle$ Yes

Markov Blanket

- The Markov blanket of a variable V is the set of variables connected to V with an arc
- E.g.: the Markov blanket of E is $\{B, A\}$



Markov Network

- Inference:
 - Algorithms similar to those for BN (bp, cp, ve, ss...)
 - Same complexity
- MN can represent some independencies that BN can not represent and vice versa
- Advantage: we do not have to avoid cycles
- Disadvantage: MN parameters are more difficult to interpret

BN Software

- List of BN software
<http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html>
- BNT: inference and learning, Matlab, open source
- MSBNx: inference, by Microsoft, free closed source
- OpenBayes: inference and learning, Python, open source
- BNJ: inference and learning, Java, open source
- Weka: learning, Java, open source

BN Software

- libDAI - A free and open source C++ library for Discrete Approximate Inference in graphical models
<http://people.kyb.tuebingen.mpg.de/jorism/libDAI/>
- FastInf: memory and time efficient approximate inference in large relational undirected graphical models
<http://compbio.cs.huji.ac.il/FastInf/fastInf>

Resources

- Daphne Koller, Nir Friedman, Probabilistic graphical models: principles and techniques, MIT Press: 2009, ISBN 978-0-262-01319-2
- Probabilistic Reasoning in Intelligent Systems by Judea Pearl. Morgan Kaufmann: 1998.
- Probabilistic Reasoning in Expert Systems by Richard Neapolitan. Wiley: 1990.
- List of BN Models and Datasets
<http://www.cs.huji.ac.il/labs/compbio/Repository/>

References

- [Pearl 85] Pearl, J., "Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning," UCLA CS Technical Report 850021, Proceedings, Cognitive Science Society, UC Irvine, 329-334, August 15-17, 1985.
- [Pearl 00] Pearl, J., Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000
- [Suermondt et al. 90] Henri Jacques Suermondt, Gregory F. Cooper, David Heckerman, "A combination of cutset conditioning with clique-tree propagation in the Pathfinder system", UAI '90.

References

- [Pearl 88] Judea Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann: 1998.
- [Lauritzen, Spiegelhalter 1988]
- [Heckerman, Geiger, Chickering 95] D. Heckerman, D. Geiger, D. M. Chickering: “Learning Bayesian Networks: The Combination of Knowledge and Statistical Data”, Machine Learning, 20(3), 1995