Bayesian Networks

- Syntax
- Semantics
- Parametrized Distributions
Bayesian Networks - Structure

- Are a simple, graphical notation for conditional independence assertions
  - hence for compact specifications of full joint distributions

- A BN is a directed graph with the following components:
  - **Nodes**: one node for each variable
  - **Edges**: a directed edge from node $N_i$ to node $N_j$ indicates that variable $X_i$ has a direct influence upon variable $X_j$
In addition to the structure, we need a **conditional probability distribution** for the random variable of each node given the random variables of its parents.

- i.e. we need $P(X_i \mid \text{Parents}(X_i))$

Nodes/variables that are not connected are (conditionally) independent:

- Weather is independent of Cavity
- Toothache is independent of Catch given Cavity
Running Example: Alarm

- **Situation:**
  - I'm at work
  - John calls to say that the in my house alarm went off
    - but Mary (my neighbor) did not call
  - The alarm will usually be set off by burglars
    - but sometimes it may also go off because of minor earthquakes

- **Variables:**
  - *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*

- **Network topology reflects causal knowledge:**
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call
Alarm Example

Bayesian Networks

© J. Fürnkranz

TU Darmstadt, SS 2009

Einführung in die Künstliche Intelligenz
Local Semantics of a BN

- Each node is conditionally independent of its nondescendants given its parents

\[
P(X \mid U_1, \ldots, U_m, Z_{1j}, \ldots, Z_{nj}) = P(X \mid U_1, \ldots, U_m)
\]
Markov Blanket

- **Markov Blanket:**
  - parents + children + children's parents

Each node is conditionally independent of all other nodes given its markov blanket

\[
P(X \mid U_1, \ldots, U_m, Y_1, \ldots, Y_n, Z_{1j}, \ldots, Z_{nj}) = P(X \mid all\ variables)
\]
Global Semantics of a BN

- The conditional probability distributions define the joint probability distribution of the variables of the network

\[ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(X_i)) \]

- Example:
  - What is the probability that the alarm goes off and both John and Mary call, but there is neither a burglary nor an earthquake?

\[ P(j \land m \land a \land \neg b \land \neg e) = \]

\[ = P(j \mid a) \cdot P(m \mid a) \cdot P(a \mid \neg b, \neg e) \cdot P(\neg b) \cdot P(\neg e) \]

\[ = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \approx 0.00063 \]
Theorem

Local Semantics ⇔ Global Semantics

Proof:
- order the variables so that parents always appear before children
- apply chain rule
- use conditional independence
Constructing Bayesian Networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables \( X_1, \ldots, X_n \)
2. For \( i = 1 \) to \( n \)
   - add \( X_i \) to the network
   - select parents from \( X_1, \ldots, X_{i-1} \) such that
     \[
     P(X_i | Parents(X_i)) = P(X_i | X_1, \ldots, X_{i-1})
     \]

This choice of parents guarantees the global semantics:

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}
\]

\[
= \prod_{i=1}^{n} P(X_i | Parents(X_i)) \quad \text{(by construction)}
\]
Example

- Suppose we first select the ordering \textit{Mary Calls, John Calls, Alarm, Burglary, Earthquake},

\[ P(J \mid M) = P(J)? \]

If Mary calls, it is more likely that John calls as well.
Example

- Suppose we first select the ordering 
  *MaryCalls, JohnCalls, Alarm, Burglary, Earthquake*.

\[ P(A | J, M) = P(A) ? \]

If Mary and John call, the probability that the alarm has gone off is larger than if they don't call.

Node A needs parents J or M
Example

- Suppose we first select the ordering
  \textit{MaryCalls, JohnCalls, Alarm, Burglary, Earthquake},

\[ P(A | J, M) = P(A | J) ? \]

If John and Mary call, the probability that the alarm has gone off is higher than if only John calls.
Example

Suppose we first select the ordering *MaryCalls, JohnCalls, Alarm, Burglary, Earthquake*.

\[ P(A | J, M) = P(A | M) ? \]

If John and Mary call, the probability that the alarm has gone off is higher than if only Mary calls.
Example

- Suppose we first select the ordering
  *MaryCalls, JohnCalls, Alarm, Burglary, Earthquake*,

\[ P(B \mid A, J, M) \neq P(B) \]

Knowing whether Mary or John called and whether the alarm went off influences my knowledge about whether there has been a burglary.

Node B needs parents A, J or M.
Example

- Suppose we first select the ordering \textit{MaryCalls, JohnCalls, Alarm, Burglary, Earthquake},

\[ P(B \mid A, J, M) = P(B \mid A)? \]

If I know that the alarm has gone off, knowing that John or Mary have called does not add to my knowledge of whether there has been a burglary or not.

Thus, no edges from M and J, only from B.
Example

- Suppose we first select the ordering
  \( Mary\text{Calls}, John\text{Calls}, Alarm, Burglary, Earthquake, \)

\[
P(E \mid B, A, J, M) \neq P(E \mid A)?
\]

Knowing whether there has been an Earthquake does not suffice to determine the probability of an earthquake, we have to know whether there has been a burglary as well.
**Example**

- **Suppose we first select the ordering**

  *MaryCalls, JohnCalls, Alarm, Burglary, Earthquake*,

\[ P(E \mid B, A, J, M) = P(E \mid A)? \quad \times \]

\[ P(E \mid B, A, J, M) = P(E \mid A, B)? \quad \checkmark \]

If we know whether there has been an alarm and whether there has been burglary, no other factors will determine our knowledge about whether there has been an earthquake.
Example - Discussion

- Deciding conditional independence is hard in non-causal direction
  - for assessing whether \( X \) is conditionally independent of \( Z \) ask the question: *If I add variable \( Z \) in the condition, does it change the probabilities for \( X \)?*
  - causal models and conditional independence seem hardwired for humans!
- Assessing conditional probabilities is also hard in non-causal direction
- Network is less compact
  - more edges and more parameters to estimate
- Worst possible ordering *MaryCalls, JohnCalls Earthquake, Burglary, Alarm* → fully connected network

![Bayesian Network Diagram](image-url)
Reasoning with Bayesian Networks

- Belief Functions (margin probabilities)
  - given the probability distribution, we can compute a margin probability at each node, which represents the belief into the truth of the proposition
  - → the margin probability is also called the belief function

- New evidence can be incorporated into the network by changing the appropriate belief functions
  - this may not only happen in unconditional nodes!

- changes in the margin probabilities are then propagated through the network
  - propagation happens in forward (along the causal links) and backward direction (against them)
    - e.g., determining a symptom of a disease does not cause the disease, but changes the probability with which we believe that the patient has the disease
Example: Medical Diagnosis

- Structure of the network

- Visit to Asia
- Smoking
- Tuberculosis
- Lung Cancer
- Bronchitis
- XRay Result
- Dyspnea
- Medical Difficulties
- Diagnostic Tests
- Patient Information

Slide by Buede, Tatman, Bresnik, 1998
Example: Medical Diagnosis

- Adding Probability Distributions

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
</tr>
<tr>
<td>Patient Information</td>
<td></td>
</tr>
<tr>
<td>Medical Difficulties</td>
<td></td>
</tr>
<tr>
<td>Diagnostic Tests</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Stress</td>
<td>Present</td>
</tr>
<tr>
<td>Diagnostic Tests</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Brachitis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td>Present</td>
</tr>
<tr>
<td>Present</td>
<td>0.90</td>
</tr>
<tr>
<td>Absent</td>
<td>0.10</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.70</td>
</tr>
<tr>
<td>False</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.80</td>
</tr>
<tr>
<td>False</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.10</td>
</tr>
<tr>
<td>False</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.70</td>
</tr>
<tr>
<td>False</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.80</td>
</tr>
<tr>
<td>False</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuberculosis or Cancer</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tub or Can</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>Present</td>
</tr>
<tr>
<td>False</td>
<td>Present</td>
</tr>
<tr>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>True</td>
<td>0.10</td>
</tr>
<tr>
<td>False</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Example: Medical Diagnosis

- **Belief functions**

  - **Visit To Asia**
    - Visit: 1.00
    - No Visit: 99.0

  - **Smoking**
    - Smoker: 50.0
    - NonSmoker: 50.0

  - **Tuberculosis**
    - Present: 1.04
    - Absent: 99.0

  - **Lung Cancer**
    - Present: 5.50
    - Absent: 94.5

  - **Bronchitis**
    - Present: 45.0
    - Absent: 55.0

  - **XRay Result**
    - Abnormal: 11.0
    - Normal: 89.0

  - **Dyspnea**
    - Present: 43.6
    - Absent: 56.4
Example: Medical Diagnosis

- Interviewing the patient results in change of probability for variable for “visit to Asia”
Example: Medical Diagnosis

- Patient is also a smoker...

![Bayesian Network Diagram](image-url)
Example: Medical Diagnosis

- but fortunately the X-ray is normal...
Example: Medical Diagnosis

- but then again patient has difficulty in breathing.
More Complex Example: Car Diagnosis

- **Initial evidence:** Car does not start
- **Test variables**
- **Hidden variables:** ensure spare structure, reduce parameters

Variables for possible failures:

![Diagram of a Bayesian network for car diagnosis](diagram.png)
More Complex: Car Insurance
Example: Alarm Network

- Monitoring system for patients in intensive care
Example: Pigs Network

- Determines pedigree of breeding pigs
  - used to diagnose PSE disease
  - half of the network structure shown here
Compactness of a BN

A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.

Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1 - p$).

If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.

I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$).
Compact Conditional Distributions

CPT grows exponentially with number of parents
CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
\[ X = f(\text{Parents}(X)) \] for some function \( f \)

E.g., Boolean functions
\[ \text{NorthAmerican} \iff \text{Canadian} \lor \text{US} \lor \text{Mexican} \]

E.g., numerical relationships among continuous variables
\[ \frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation} \]
Compact Conditional Distributions
Independent Causes

Noisy-OR distributions model multiple noninteracting causes
1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability $q_i$ for each cause alone

$$P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg \text{Fever})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02 = 0.2 × 0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 × 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 × 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 × 0.2 × 0.1</td>
</tr>
</tbody>
</table>

Number of parameters linear in number of parents
Hybrid Networks

Discrete \((\text{Subsidy? and Buys?})\); continuous \((\text{Harvest and Cost})\)

Option 1: discretization—possibly large errors, large CPTs
Option 2: finitely parameterized canonical families

1) Continuous variable, discrete+continuous parents (e.g., \text{Cost})
2) Discrete variable, continuous parents (e.g., \text{Buys?})
Continuous Conditional Distributions

We need one conditional density function for each child variable given its continuous parents, for each possible assignment to discrete parents.

Most common is the linear Gaussian model, e.g.:

\[
P(Cost = c \mid Harvest = h, Subsidy? = \text{true})
\]

\[
= N(a_th + b_t, \sigma_t)(c)
\]

\[
= \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{c - (a_th + b_t)}{\sigma_t} \right)^2 \right)
\]

Mean Cost varies linearly with Harvest, variance is fixed.

Linear variation is unreasonable over the full range, but works OK if the likely range of Harvest is narrow.
Continuous Conditional Distributions

\[ P(c \mid h, \text{subventionen}) \]

\[ P(c \mid h, \neg \text{subventionen}) \]

\[ P(c \mid h) = P(c \mid h, \text{subventionen}) + P(c \mid h, \neg \text{subventionen}) \]

All-continuous network with LG distributions

\[ \Rightarrow \text{full joint distribution is a multivariate Gaussian} \]

Discrete+continuous LG network is a **conditional Gaussian** network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values
Discrete Variables with Continuous Parents

Probability of $\text{Buys?}$ given $\text{Cost}$ should be a “soft” threshold:

Probit distribution uses integral of Gaussian:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy$$

$$P(\text{Buys?} = \text{true} \mid \text{Cost} = c) = \Phi\left(\frac{-c + \mu}{\sigma}\right)$$
Why Probit?

1. It’s sort of the right shape

2. Can view as hard threshold whose location is subject to noise
Discrete Variables with Continuous Parents

Sigmoid (or logit) distribution also used in neural networks:

\[
P(Buys? = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + e^{\frac{-\mu - c}{\sigma}}}
\]

Sigmoid has similar shape to probit but much longer tails:
Real-World Applications of BN

- **Industrial**
  - Processor Fault Diagnosis - by Intel
  - Auxiliary Turbine Diagnosis - GEMS by GE
  - Diagnosis of space shuttle propulsion systems - VISTA by NASA/Rockwell
  - Situation assessment for nuclear power plant – NRC

- **Military**
  - Automatic Target Recognition - MITRE
  - Autonomous control of unmanned underwater vehicle - Lockheed Martin
  - Assessment of Intent
Real-World Applications of BN

- Medical Diagnosis
  - Internal Medicine
  - Pathology diagnosis - Intellipath by Chapman & Hall
  - Breast Cancer Manager with Intellipath

- Commercial
  - Financial Market Analysis
  - Information Retrieval
  - Software troubleshooting and advice - Windows 95 & Office 97
  - Pregnancy and Child Care – Microsoft
  - Software debugging - American Airlines’ SABRE online reservation system