Bayes’ Nets: Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

- Bayes’ nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we’ll be vague about how these interactions are specified
Bayes’ Nets

- A Bayes’ net is an efficient encoding of a probabilistic model of a domain

- Questions we can ask:
  - Inference: given a fixed BN, what is $P(X \mid e)$?
  - Representation: given a BN graph, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?

This slide deck courtesy of Dan Klein
Bayes’ Net Semantics

- Let’s formalize the semantics of a Bayes’ net
- A set of nodes, one per variable $X$
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over $X$, one for each combination of parents’ values
    \[ P(X|a_1 \ldots a_n) \]

  - CPT: conditional probability table
  - Description of a noisy “causal” process

A Bayes net = Topology (graph) + Local Conditional Probabilities
Probabilities in BNs

- Bayes’ nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:
    \[ P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

- Example:
  \[ P(+\text{cavity}, +\text{catch}, -\text{toothache}) \]

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies
Example: Traffic

\[ P(R) \]

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>3/4</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(T|R) \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>+t</th>
<th>3/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td>1/4</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(\neg r, \neg t) = \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td>1/2</td>
</tr>
<tr>
<td>\neg t</td>
<td>1/2</td>
<td></td>
</tr>
</tbody>
</table>
Example: Alarm Network

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>¬b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>¬e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| A   | J   | P(J|A) |
|-----|-----|------|
| +a  | +j  | 0.9  |
| +a  | ¬j  | 0.1  |
| ¬a  | +j  | 0.05 |
| ¬a  | ¬j  | 0.95 |

| A   | M   | P(M|A) |
|-----|-----|------|
| +a  | +m  | 0.7  |
| +a  | ¬m  | 0.3  |
| ¬a  | +m  | 0.01 |
| ¬a  | ¬m  | 0.99 |

| B   | E   | A   | P(A|B,E) |
|-----|-----|-----|---------|
| +b  | +e  | +a  | 0.95    |
| +b  | +e  | ¬a  | 0.05    |
| +b  | ¬e  | +a  | 0.94    |
| ¬b  | +e  | +a  | 0.29    |
| ¬b  | +e  | ¬a  | 0.71    |
| ¬b  | ¬e  | +a  | 0.001   |
| ¬b  | ¬e  | ¬a  | 0.999   |
Example: Independence

- For this graph, you can fiddle with $\theta$ (the CPTs) all you want, but you won’t be able to represent any distribution in which the flips are dependent!

<table>
<thead>
<tr>
<th>$P(X_1)$</th>
<th>$P(X_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>h 0.5</td>
<td>h 0.5</td>
</tr>
<tr>
<td>t 0.5</td>
<td>t 0.5</td>
</tr>
</tbody>
</table>
Topology Limits Distributions

- Given some graph topology $G$, only certain joint distributions can be encoded.
- The graph structure guarantees certain (conditional) independences.
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs.
- Full conditioning can encode any distribution.
When Bayes’ nets reflect the true causal patterns:
- Often simpler (nodes have fewer parents)
- Often easier to think about
- Often easier to elicit from experts

BNs need not actually be causal
- Sometimes no causal net exists over the domain (especially if variables are missing)
- E.g. consider the variables Traffic and Drips
- End up with arrows that reflect correlation, not causation

What do the arrows really mean?
- Topology may happen to encode causal structure
- Topology really encodes conditional independence
Example: Traffic

- Causal direction

\[
P(R)
\]

\[
\begin{array}{c|c}
  r & 1/4 \\
  \neg r & 3/4 \\
\end{array}
\]

\[
P(T|R)
\]

\[
\begin{array}{c|c|c}
  r & t & 3/4 \\
  r & \neg t & 1/4 \\
  \neg r & t & 1/2 \\
  \neg r & \neg t & 1/2 \\
\end{array}
\]

\[
P(T, R)
\]

\[
\begin{array}{c|c|c}
  r & t & 3/16 \\
  r & \neg t & 1/16 \\
  \neg r & t & 6/16 \\
  \neg r & \neg t & 6/16 \\
\end{array}
\]

\[
\text{Example: Traffic}
\]

- Causal direction

\[
P(R)
\]

\[
\begin{array}{c|c}
  r & 1/4 \\
  \neg r & 3/4 \\
\end{array}
\]

\[
P(T|R)
\]

\[
\begin{array}{c|c|c}
  r & t & 3/4 \\
  r & \neg t & 1/4 \\
  \neg r & t & 1/2 \\
  \neg r & \neg t & 1/2 \\
\end{array}
\]

\[
P(T, R)
\]

\[
\begin{array}{c|c|c}
  r & t & 3/16 \\
  r & \neg t & 1/16 \\
  \neg r & t & 6/16 \\
  \neg r & \neg t & 6/16 \\
\end{array}
\]
Example: Reverse Traffic

- Reverse causality?

\[
\begin{array}{|c|c|}
\hline
T & P(T) \\
\hline
\text{t} & 9/16 \\
\text{\neg t} & 7/16 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
R & P(R|T) \\
\hline
\text{t} & \text{r} & 1/3 \\
\text{\neg r} & 2/3 \\
\hline
\text{\neg t} & \text{r} & 1/7 \\
\text{\neg r} & 6/7 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
P(T, R) & r & t & 3/16 \\
\text{r} & \text{\neg t} & 1/16 \\
\text{\neg r} & t & 6/16 \\
\text{\neg r} & \text{\neg t} & 6/16 \\
\hline
\end{array}
\]
Changing Bayes’ Net Structure

- The same joint distribution can be encoded in many different Bayes’ nets
  - Causal structure tends to be the simplest

- Analysis question: given some edges, what other edges do you need to add?
  - One answer: fully connect the graph
  - Better answer: don’t make any false conditional independence assumptions
Example: Alternate Alarm

If we reverse the edges, we make different conditional independence assumptions.

To capture the same joint distribution, we have to add more edges to the graph.
Bayes’ Nets

- So far: how a Bayes’ net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Key idea: conditional independence
  - Today: assembled BNs using an intuitive notion of conditional independence as causality
  - Next: formalize these ideas
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)
Example: Naïve Bayes

- Imagine we have one cause $y$ and several effects $x$:

$$P(y, x_1, x_2 \ldots x_n) = P(y)P(x_1|y)P(x_2|y) \ldots P(x_n|y)$$

- This is a naïve Bayes model
- We’ll use these for classification later
Example: Alarm Network

\[
P(b, e, \neg a, j, m) =
\]
The Chain Rule

- Can always factor any joint distribution as an incremental product of conditional distributions

\[ P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots \]

\[ P(X_1, X_2, \ldots X_n) = \prod_{i} P(X_i|X_1 \ldots X_{i-1}) \]

- Why is the chain rule true?

- This actually claims nothing…

- What are the sizes of the tables we supply?
Example: Alarm Network

\[ \prod_{i} P(X_i | \text{Parents}(X_i)) = P(B) \cdot P(E) \cdot P(A | B, E) \cdot P(J | A) \cdot P(M | A) \]
Bayes’ Net Semantics

- Let’s formalize the semantics of a Bayes’ net

- A set of nodes, one per variable X

- A directed, acyclic graph

- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents’ values
    \[ P(X|a_1 \ldots a_n) \]

- CPT: conditional probability table
- Description of a noisy “causal” process

A Bayes net = Topology (graph) + Local Conditional Probabilities
Example: Alarm Network

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>−b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>−e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| A | J | P(J|A) |
|---|---|------|
| +a | +j | 0.9 |
| +a | −j | 0.1 |
| −a | +j | 0.05 |
| −a | −j | 0.95 |

| A | M | P(M|A) |
|---|---|------|
| +a | +m | 0.7 |
| +a | −m | 0.3 |
| −a | +m | 0.01 |
| −a | −m | 0.99 |

| B | E | A | P(A|B,E) |
|---|---|---|---------|
| +b | +e | +a | 0.95 |
| +b | +e | −a | 0.05 |
| +b | −e | +a | 0.94 |
| +b | −e | −a | 0.06 |
| −b | +e | +a | 0.29 |
| −b | +e | −a | 0.71 |
| −b | −e | +a | 0.001 |
| −b | −e | −a | 0.999 |
Size of a Bayes’ Net

- How big is a joint distribution over N Boolean variables?
  \[2^N\]

- How big is an N-node net if nodes have up to k parents?
  \[O(N \times 2^{k+1})\]

- Both give you the power to calculate \(P(X_1, X_2, \ldots, X_n)\)
- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also turns out to be faster to answer queries (coming)
Building the (Entire) Joint

- We can take a Bayes’ net and build any entry from the full joint distribution it encodes.

\[
P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))
\]

- Typically, there’s no reason to build ALL of it.
- We build what we need on the fly.

- To emphasize: every BN over a domain implicitly defines a joint distribution over that domain, specified by local probabilities and graph structure.
Bayes’ Nets So Far

- We now know:
  - What is a Bayes’ net?
  - What joint distribution does a Bayes’ net encode?

- Now: properties of that joint distribution (independence)
  - Key idea: conditional independence
  - Last class: assembled BNs using an intuitive notion of conditional independence as causality
  - Today: formalize these ideas
  - Main goal: answer queries about conditional independence and influence

- Next: how to compute posteriors quickly (inference)
Bayes Nets: Assumptions

- Assumptions we are required to make to define the Bayes net when given the graph:

\[ P(x_i | x_1 \cdots x_{i-1}) = P(x_i | \text{parents}(X_i)) \]

- Probability distributions that satisfy the above ("chain-rule→Bayes net") conditional independence assumptions
  - Often guaranteed to have many more conditional independences
  - Additional conditional independences can be read off the graph

- Important for modeling: understand assumptions made when choosing a Bayes net graph
Example

- Conditional independence assumptions directly from simplifications in chain rule:

- Additional implied conditional independence assumptions?
Conditional Independence

- Reminder: independence
  - $X$ and $Y$ are independent if
    \[ \forall x, y \quad P(x, y) = P(x)P(y) \quad \rightarrow \quad X \independent Y \]
  - $X$ and $Y$ are conditionally independent given $Z$
    \[ \forall x, y, z \quad P(x, y|z) = P(x|z)P(y|z) \quad \rightarrow \quad X \independent Y|Z \]
  - (Conditional) independence is a property of a distribution
D-separation: Outline

- Study independence properties for triples

- Any complex example can be analyzed using these three canonical cases
Independence in a BN

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter example
  - Example:

```
X -- Y -- Z
```

- Question: are X and Z necessarily independent?
  - Answer: no. Example: low pressure causes rain, which causes traffic.
  - X can influence Z, Z can influence X (via Y)
  - Addendum: they *could* be independent: how?
Causal Chains

- This configuration is a “causal chain”

\[ P(x, y, z) = P(x)P(y|x)P(z|y) \]

- Is X independent of Z given Y?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z|y) \]

Yes!

- Evidence along the chain “blocks” the influence

X: Low pressure
Y: Rain
Z: Traffic
Another basic configuration: two effects of the same cause

- Are $X$ and $Z$ independent?

- Are $X$ and $Z$ independent given $Y$?

\[
P(z | x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x | y)P(z | y)}{P(y)P(x | y)} = P(z | y)
\]

Yes!

Observing the cause blocks influence between effects.
Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: the ballgame and the rain cause traffic, but they are not correlated
    - Still need to prove they must be (try it!)
  - Are X and Z independent given Y?
    - No: seeing traffic puts the rain and the ballgame in competition as explanation?
  - This is backwards from the other cases
    - Observing an effect activates influence between possible causes.

X: Raining
Z: Ballgame
Y: Traffic
The General Case

- Any complex example can be analyzed using these three canonical cases

- General question: in a given BN, are two variables independent (given evidence)?

- Solution: analyze the graph
Reachability

- **Recipe:** shade evidence nodes

- **Attempt 1:** if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent

- **Almost works, but not quite**
  - Where does it break?
  - **Answer:** the v-structure at T doesn’t count as a link in a path unless “active”
Reachability (D-Separation)

- **Question:** Are X and Y conditionally independent given evidence vars \( \{Z\} \)?
  - Yes, if X and Y “separated” by Z
  - Look for active paths from X to Y
  - No active paths = independence!

- **A path is active if each triple is active:**
  - Causal chain \( A \rightarrow B \rightarrow C \) where B is unobserved (either direction)
  - Common cause \( A \leftarrow B \rightarrow C \) where B is unobserved
  - Common effect (aka v-structure) \( A \rightarrow B \leftarrow C \) where B or one of its descendents is observed

- **All it takes to block a path is a single inactive segment**
D-Separation

- Given query $X_i \perp \!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
- Shade all evidence nodes
- For all (undirected!) paths between and
  - Check whether path is active
    - If active return $X_i \perp \!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
- (If reaching this point all paths have been checked and shown inactive)
  - Return $X_i \perp \!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
Example

\[ R \perp B \]
\[ R \perp B | T \]
\[ R \perp B | T' \]
Example

\begin{align*}
L \perp T' | T & \quad \text{Yes} \\
L \perp B & \quad \text{Yes} \\
L \perp B | T & \\
L \perp B | T' & \\
L \perp B | T, R & \quad \text{Yes} \\
\end{align*}
Example

- **Variables:**
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I’m sad

- **Questions:**
  
  \[ T \perp D \]
  
  \[ T \perp D \mid R \quad \text{Yes} \]
  
  \[ T \perp D \mid R, S \]
Given a Bayes net structure, can run d-separation to build a complete list of conditional independences that are necessarily true of the form

\[ X_i \perp\!\!\!\!\!\!\!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_m}\} \]

This list determines the set of probability distributions that can be represented.
Causality?

- When Bayes’ nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts

- BNs need not actually be causal
  - Sometimes no causal net exists over the domain
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation

- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology only guaranteed to encode conditional independence
Example: Traffic

- Basic traffic net
- Let’s multiply out the joint

\[ P(R) \]

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>( \neg r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1/4</td>
<td></td>
</tr>
<tr>
<td>( \neg r )</td>
<td>3/4</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(T|R) \]

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>( \neg r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>( \neg t )</td>
<td>1/4</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(T, R) \]

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>( \neg r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>3/16</td>
<td></td>
</tr>
<tr>
<td>( \neg t )</td>
<td>1/16</td>
<td></td>
</tr>
<tr>
<td>( \neg r ) t</td>
<td>6/16</td>
<td></td>
</tr>
<tr>
<td>( \neg r ) ( \neg t )</td>
<td>6/16</td>
<td></td>
</tr>
</tbody>
</table>
Example: Reverse Traffic

- Reverse causality?

\[
\begin{array}{c|c}
\text{t} & 9/16 \\
\hline
\text{¬t} & 7/16 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{t} & \text{r} & 1/3 \\
\hline
\text{¬r} & 2/3 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{r} & \text{t} & 3/16 \\
\hline
\text{r} & \text{¬t} & 1/16 \\
\hline
\text{¬r} & \text{t} & 6/16 \\
\hline
\text{¬r} & \text{¬t} & 6/16 \\
\end{array}
\]
Example: Coins

- Extra arcs don’t prevent representing independence, just allow non-independence

\[
P(X_1) = \begin{pmatrix} h & 0.5 \\ t & 0.5 \end{pmatrix}, \quad P(X_2) = \begin{pmatrix} h & 0.5 \\ t & 0.5 \end{pmatrix}, \quad P(X_1) = \begin{pmatrix} h & 0.5 \\ t & 0.5 \end{pmatrix}, \quad P(X_2 | X_1) = \begin{pmatrix} h | h & 0.5 \\ t | h & 0.5 \\ h | t & 0.5 \\ t | t & 0.5 \end{pmatrix}
\]

- Adding unneeded arcs isn’t wrong, it’s just inefficient
Summary

- Bayes nets compactly encode joint distributions

- Guaranteed independencies of distributions can be deduced from BN graph structure

- D-separation gives precise conditional independence guarantees from graph alone

- A Bayes’ net’s joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution
Example: Alarm Network

\[ \prod_{i} P(X_i|\text{Parents}(X_i)) = P(B) \cdot P(E) \cdot P(A|B, E) \cdot P(J|A) \cdot P(M|A) \]
Reachability (the Bayes’ Ball)

- Correct algorithm:
  - Shade in evidence
  - Start at source node
  - Try to reach target by search

- States: pair of (node X, previous state S)

- Successor function:
  - X unobserved:
    - To any child
    - To any parent if coming from a child
  - X observed:
    - From parent to parent

- If you can’t reach a node, it’s conditionally independent of the start node given evidence
Example: Independence

- For this graph, you can fiddle with $\theta$ (the CPTs) all you want, but you won’t be able to represent any distribution in which the flips are dependent!

\[
\begin{align*}
\text{All distributions} &
\end{align*}
\]
Topology Limits Distributions

- Given some graph topology $G$, only certain joint distributions can be encoded.
- The graph structure guarantees certain (conditional) independences.
- (There might be more independence.)
- Adding arcs increases the set of distributions, but has several costs.
- Full conditioning can encode any distribution.
Expectimax Evaluation

- Evaluation functions quickly return an estimate for a node’s true value (which value, expectimax or minimax?)
- For minimax, evaluation function scale doesn’t matter
  - We just want better states to have higher evaluations (get the ordering right)
  - We call this **insensitivity to monotonic transformations**
- For expectimax, we need *magnitudes* to be meaningful

This slide deck courtesy of Dan Klein at UC Berkeley
Multi-Agent Utilities

- Similar to minimax:
  - Terminals have utility tuples
  - Node values are also utility tuples
  - Each player maximizes its own utility
  - Can give rise to cooperation and competition dynamically…
Maximum Expected Utility

- Why should we average utilities? Why not minimax?

- Principle of maximum expected utility:
  - A rational agent should choose the action which maximizes its expected utility, given its knowledge

- Questions:
  - Where do utilities come from?
  - How do we know such utilities even exist?
  - Why are we taking expectations of utilities (not, e.g. minimax)?
  - What if our behavior can’t be described by utilities?