ECE 175B: Probabilistic Reasoning and Graphical Models
Lecture 5: Basic Properties of BNs

Florian Meyer & Ken Kreutz-Delgado

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Example 1

- Consider a simple example where we discuss the joint distribution $P(x_1, x_2)$
- Three different graphical models can be established

$P(x_1, x_2)$

UG

$P(x_1, x_2) = P(x_2|x_1)P(x_1)$

DAG1

$P(x_1, x_2) = P(x_1|x_2)P(x_2)$

DAG2

Example 1

- Case I: Here, x_1 represents the reading of a thermometer and x_2 represents the sidewalk temperature
  - Thermometer reading and sidewalk temperature are correlated, but not causally related
  - All three models can be used for prediction of one given the other, based on the conditional likelihood $P(x_1|x_2)$ or $P(x_2|x_1)$
  - None can be used for control purpose

$P(x_1, x_2)$

UG

$P(x_1, x_2) = P(x_2|x_1)P(x_1)$

DAG1

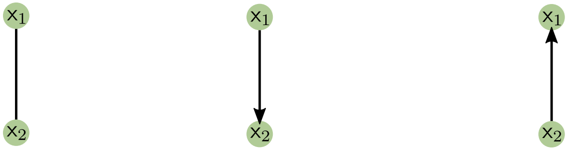
$P(x_1, x_2) = P(x_1|x_2)P(x_2)$

DAG2

Example 1

- Case II: Here, x_1 represents a thermometer reading and x_2 represents an air conditioner setting (it is assumed that there a are no latent variables)
    - The air conditioner setting causes the reading of the thermometer
    - All three models can be used for prediction
    - Only DAG2 can be used for control based on $P(\mathsf{x}_1|\mathsf{x}_2)$
    - However, if there are “lurking latent variables”, even DAG2 might be inadequate due to “confounding”

$P(\mathsf{x}_1, \mathsf{x}_2)$ $\quad$ $P(\mathsf{x}_1, \mathsf{x}_2) = P(\mathsf{x}_2|\mathsf{x}_1)P(\mathsf{x}_1)$ $\quad$ $P(\mathsf{x}_1, \mathsf{x}_2) = P(\mathsf{x}_1|\mathsf{x}_2)P(\mathsf{x}_2)$



UG $\quad$ **DAG1** $\quad$ **DAG2**

Example 1

- All three graphical models might be used for prediction
- DAG1 and DAG2 might be useful for control (observation vs intervention), i.e.,
  - neither might be appropriate
  - otherwise either one or the other, but not both, can be used
- In general, control is much harder since it's related to causation; one has to match the distribution with the experimental conditions, otherwise apparent paradoxes may arise

$P(x_1, x_2)$

UG

$P(x_1, x_2) = P(x_2|x_1)P(x_1)$

DAG1

$P(x_1, x_2) = P(x_1|x_2)P(x_2)$

DAG2

Example 2

- Consider the game were we toss a coin three times independently, i.e., $x_i \in \{0, 1\}, i = 1, 2, 3$ represents the outcome of three tosses, where 0 is "tail" and 1 is "head"
- We win the game if there is at least one "head" in three trials, i.e., $x_4 \in \{0, 1\}$ represents the result of the game where $x_4 = 1$ if we win the game and 0 otherwise
- Let us define $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and try two different factorizing orders of $P(\mathcal{X})$

$$\text{(a)} \quad P(\mathcal{X}) = P(x_1|x_2, x_3, x_4)P(x_2|x_3, x_4)P(x_3|x_4)P(x_4)$$

$$\begin{aligned} \text{(b)} \quad P(\mathcal{X}) &= P(x_4|x_1, x_2, x_3)P(x_3|x_2, x_1)P(x_2|x_1)P(x_1) \\ &= P(x_4|x_1, x_2, x_3)P(x_3)P(x_2)P(x_1) \end{aligned}$$

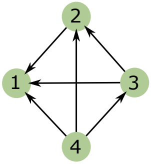
Example 2

- Factorization (a) results in the largest memory requirements since we need to specify and store 15 values; the BN w.r.t factorization (a) is a dense, complete graph

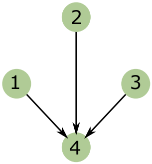
$$\underbrace{P(\mathcal{X})}_{1+2+4+8=15} = \underbrace{P\left(\widehat{x_1}^1|\widehat{x_2}^2, \widehat{x_3}^2, \widehat{x_4}^2\right)}_{2^3=8} \underbrace{P\left(\widehat{x_3}^1|\widehat{x_2}^2, \widehat{x_1}^2\right)}_{2^2=4} \underbrace{P\left(\widehat{x_2}^1|\widehat{x_1}^2\right)}_{2} \underbrace{P\left(\widehat{x_1}^1\right)}_{1}$$

- Factorization (b) results in the lowest memory requirements since we need to only store 11 values; the BN w.r.t (b) is a tree in skeleton

$$\underbrace{P(\mathcal{X})}_{1+1+1+8=11} = \underbrace{P(\widehat{x_4}^1|\widehat{x_1}^2, \widehat{x_2}^2, \widehat{x_3}^2)}_{2^3=8} \underbrace{P(\widehat{x_3}^1)}_{1} \underbrace{P(\widehat{x_2}^1)}_{1} \underbrace{P(\widehat{x_1}^1)}_{1}$$



(a)

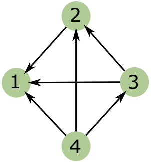


(b)

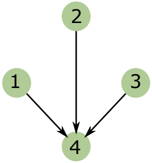
Example 2

- Different factorizations may generate different BNs with different edge density, which then influence the complexity to specify all possible values of probability

- A good factorization with optimal complexity should conform to the causal intuition



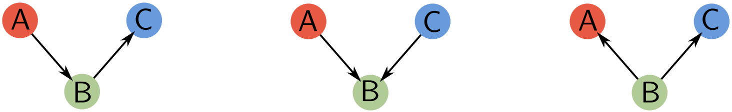
(a)



(b)

Inference on BN

- What does a BN express?
  Causation, independence, conditional independence, etc.



$A \perp\!\!\!\perp C$? or $A \perp\!\!\!\perp C \mid B$?

- What can BN be used for?
  Inference (Jeffrey's rule, Bayes' rule)

- This will be discussed in detail later

Domain Knowledge

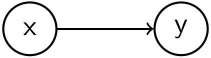
- So far we largely assumed that all distributions are fully specified for the inference tasks; this is very hard to do in practice
- In machine learning and related fields, distributions are learned from data; here learning becomes the problem of integrating data with domain knowledge to form a model for the problem of interest
- What is actually done is that a "domain expert"
  - first specify the factorization order, refine the parent set for every x_j, $j = 1, \ldots, N$ and builds a BN using "domain knowledge" and "common sense"
  - then learns the specific values for each $P(\mathrm{x}_j|\mathbf{pa}(\mathrm{x}_j))$, $j = 1, \ldots, N$ (e.g., via maximum likelihood estimators), which is a "divide and conquer" learning strategy

Chapter 9, "Bayesian Reasoning and Machine Learning" by D. Barber

BNs are Straightforward to Modify and Extend/Grow

Example 1: Given the proposition "This animal can fly"

Let $\mathrm{y} = y \in \{\underbrace{1}_{\text{true}}, \underbrace{0}_{\text{false}}\}$, $\mathrm{x} = x \in \mathcal{X} = \{\text{bird, dog, cat}\}$, and consider the simple model



We initially specify $P(\mathrm{y} = 1|\mathrm{x} = \text{bird}) = 1$ but then we remember penguins!

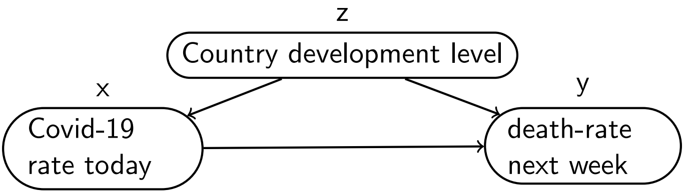
We can fix this in a variety of ways

- By setting $\overbrace{P(\mathrm{y} = 1|\mathrm{x} = \text{bird}) = 0.98}^{\text{re-specification}}$
- By growing $\mathcal{X} = \{\text{dog, cat, penguin, kiwi, emu}, \ldots\}$
- By being more specific: birds = "birds in the neighbourhood of a pet shop"

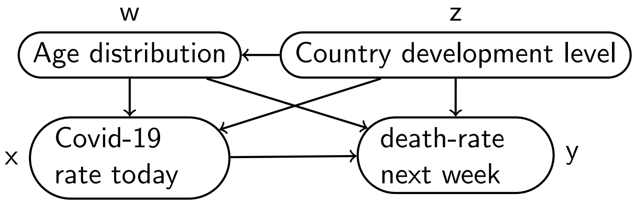
BNs are Straightforward to Modify and Extend/Grow

Example 2:

- The initial model is given by $P(y|x,z)P(x|z)P(z)$

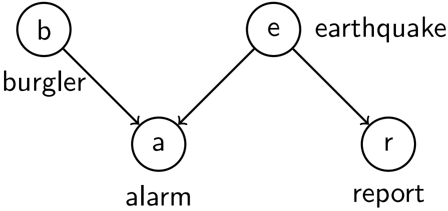


- The extended model is $P(y|x,w,z)P(x,w,z)P(w|z)P(z)$



“Explaining-Away” Property of Collider Nodes

Example 3.1 of BRML: Sally gets a text that her burglar alarm has gone off. She worries, but subsequently she sees a news report that there had been an earthquake. This scenario can be modelled as follows:



The random variables are $a, r, b, e \in \{\underbrace{1}_{\text{true}}, \underbrace{0}_{\text{false}}\}$

Note the structure of the subgraph $b \longrightarrow a \longleftarrow e$

We say that a is a collider node

“Explaining-Away” Property of Collider Nodes

- We can read the following factorization from the BN

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

- We assume the following specifications

$P(\mathrm{a} = 1|b, e)$	b	e
0.999	1	1
0.99	1	0
0.99	0	1
0.001	0	0

$P(\mathrm{r} = 1|e)$	e
1	1
0	0

$P(\mathrm{b} = 1)$
0.01

$P(\mathrm{e} = 1)$
10^{-6}

"Explaining-Away" Property of Collider Nodes

- Let us calculate the conditional probabilities $P(\mathrm{b}=1|\mathrm{a}=1)$ and $P(\mathrm{b}=1|\mathrm{a}=1,\mathrm{r}=1)$

$$\begin{aligned}
P(\mathrm{b}=1|\mathrm{a}=1) &= \frac{P(\mathrm{a}=1,\mathrm{b}=1)}{P(\mathrm{a}=1)} \\
&= \frac{\sum_{e,r} P(\mathrm{a}=1,\mathrm{b}=1,e,r)}{\sum_{b} P(\mathrm{a}=1,b)} \\
&= 99\%
\end{aligned}$$

$$\begin{aligned}
P(\mathrm{b}=1|\mathrm{a}=1,\mathrm{r}=1) &= \frac{P(\mathrm{a}=1,\mathrm{b}=1,\mathrm{r}=1)}{P(\mathrm{a}=1,\mathrm{r}=1)} \\
&= \frac{\sum_{e} P(\mathrm{a}=1,\mathrm{b}=1,\mathrm{r}=1,e)}{\sum_{b} P(\mathrm{a}=1,\mathrm{r}=1,b)} \\
&= 1\%
\end{aligned}$$

- The reporting of an earthquake explains-away the hypothesis of a burglary

Types of Evidence (§3.2, BRML)

- In the previous example, Sally obtained knowledge of the instantiated values of a and r
- Such a collection of observations is called evidence, denoted as $\mathcal{E} = \{\mathsf{a}, \mathsf{r}\}$, i.e., a and r are the evidence nodes
- Given evidence, we wish to compute probabilities on the non-evidence nodes, conditional on the evidence

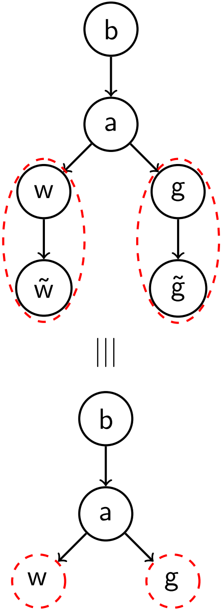
  Example: $P(\mathsf{b} = 1|\mathcal{E}) = P(\mathsf{b} = 1|\mathsf{a} = 1, \mathsf{r} = 1)$
- In our example, the evidence is a "hard evidence" which means that it is certain and reliable
- Let's dicuss the meaning of hard evidence more in detail

Types of Evidence (§3.2, BRML)

- Suppose Dr. Good never lies; he is a reliable attestor. And suppose that you are certain that he answered “Yes” to the question “did you hear the alarm?”; then you have certain and reliable evidence
- But suppose that you are uncertain that Dr. Good said “Yes” because, say, you dropped the cell-phone; then you have uncertain but reliable evidence
- Suppose Dr. Evil will lie when it’s in his best interest but otherwise will tell the truth; he is an unreliable attestor: If you are certain he said “yes”, you have certain but unreliable evidence; if you are uncertain, you have uncertain and unreliable evidence

We will consider the case of hard evidence and uncertain but reliable evidence only

Uncertain Evidence (§3.2.1, BRML)



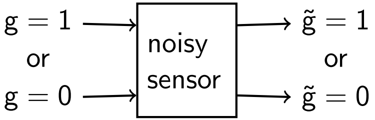
- $\mathrm{b} = 1 =$ burgled
- $\mathrm{a} = 1 =$ alarm
- $\mathrm{g} = 1 =$ alarm definitely detected by Mrs. Gibson
- $\tilde{\mathrm{g}} = 1 =$ alarm maybe detected by Mrs. Gibson
- $\mathrm{w} = 1 =$ alarm definitely detected by Dr. Watson
- $\tilde{\mathrm{w}} = 1 =$ alarm maybe detected by Dr. Watson

All our observed data is collected as evidence $\mathcal{E}$ or $\tilde{\mathcal{E}}$

- Ideal hard data $\mathcal{E} = \{\mathrm{w}, \mathrm{g}\}$
- Possibly uncertain data $\tilde{\mathcal{E}} = \{\tilde{\mathrm{w}}, \tilde{\mathrm{g}}\}$

Uncertain Evidence (§3.2.1, BRML)

g̃ and w̃ are like noisy sensors



- We can model g by $P(g|\tilde{g})$ and w by $P(w|\tilde{w})$
- From our BN, we can compute the ideal $P(b|\mathcal{E}) = P(b|w, g)$

Jeffrey's Rule:

$$
\begin{aligned}
P(b|\tilde{\mathcal{E}}) &= \sum_{\mathcal{E}\text{-values}} P(b, \mathcal{E}|\tilde{\mathcal{E}}) \\
&= \sum_{\mathcal{E}\text{-values}} P(b|\mathcal{E}, \tilde{\mathcal{E}})P(\mathcal{E}|\tilde{\mathcal{E}}) \\
&= \sum_{\mathcal{E}\text{-values}} P(b|\mathcal{E})P(\mathcal{E}|\tilde{\mathcal{E}}) \qquad (\because \text{b} \perp\!\!\!\perp \tilde{\mathcal{E}}|\mathcal{E}) \\
&= \mathbb{E}_{\mathcal{E}|\tilde{\mathcal{E}}}\left[P(b|\mathcal{E})\right] \qquad \text{where } P(\mathcal{E}|\tilde{\mathcal{E}}) = P(g|\tilde{g})P(w|\tilde{w})
\end{aligned}
$$