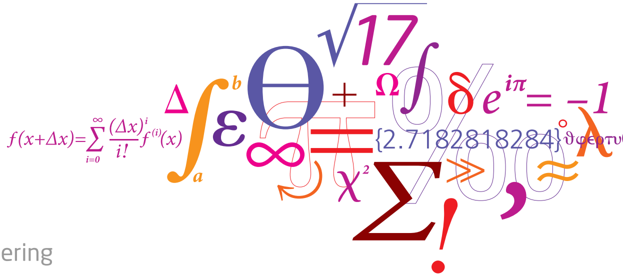
PGM foundations - Part 1

Representation, Conditional Independence and Inference

Filipe Rodrigues

Francisco Pereira


$f(x+\Delta x)=\sum_{i=0}^{\infty}\frac{(\Delta x)^i}{i!}f^{(i)}(x)$
$\sqrt{17}$
{2.7182818284}
$e^{i\pi}=-1$

DTU Management Engineering
Department of Management Engineering

Outline

- Representation
- The concept of inference
- Conditional Probability Tables (CPTs)
- D-separation

(Based on Michael Jordan, David Blei)

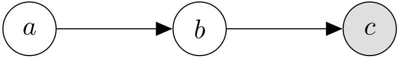
Learning objectives

At the end of this lecture, you should be able to:

- Understand the mathematical notation and representation conventions of directed graphical models
- Factorize the joint distribution that the PGM represents, taking advantage of conditional independence
- Perform exact inference in basic discrete PGMs represented with their Conditional Probability Tables (CPTs)
- Understand the concept of D-separation and prove conditional independences through the D-separation algorithm

PGM representation

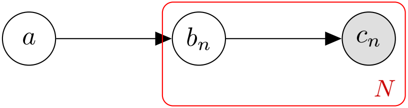
- An example graphical model



- **Nodes** represent random variables
  - shaded nodes correspond to observed variables
  - unshaded nodes denote unobserved variables
    (also known as hidden or latent variables)
- **Edges** express probabilistic relationships between the variables

PGM representation

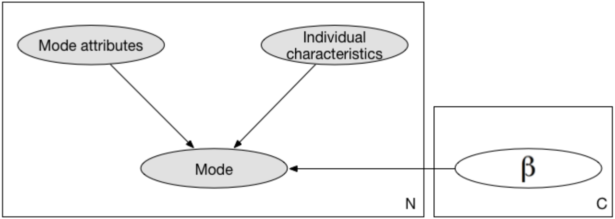
- An example graphical model



- **Nodes** represent random variables
  - shaded nodes correspond to observed variables
  - unshaded nodes denote unobserved variables (also known as hidden or latent variables)
- **Edges** express probabilistic relationships between the variables
- **Plates** indicate repetition

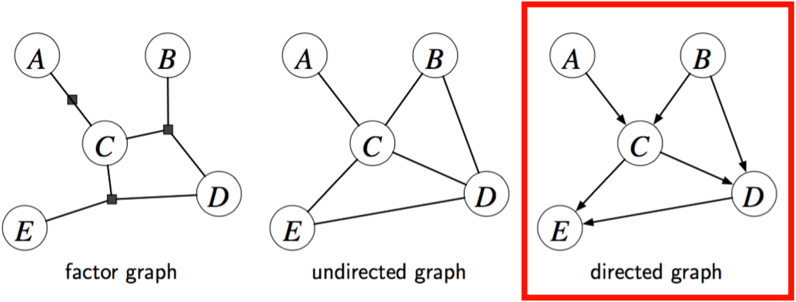
PGM representation

- A practical example



PGM representation

- There are other kinds of graphical models...



factor graph

undirected graph

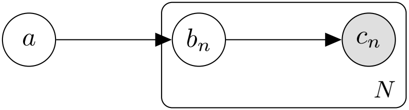
directed graph

  - Each has different properties and expressiveness
- We will mainly consider **directed** graphical models in this and coming lectures!

Recall our notation

- Unlike in the "standard" statistics notation, where:
  - X is a random variable and x is atom/event
  - We write e.g. $p(X = x)$
- In the machine learning literature, this notation is typically simplified
  - Lowercase letters, such as x, represent random variables
  - We simply write $p(x)$. Everything else should be clear from the context!
- This allows us to have
  - Bold letters denote vectors (e.g. **x**, where the i^{th} element is referred as x_i)
  - Matrices are represented by bold uppercase letters such as **X**
  - Roman letters, such as N, denote constants
- This is the notation that we will adopt from now on!

PGM representation



- PGMs represent a set of **conditional independence** relationships

$$c_n \perp\!\!\!\perp a \,|\, b_n \quad (c_n \text{ is conditionally independent of } a \text{ given } b_n)$$

  - if we observed b_n, then observing a tell us nothing about c_n

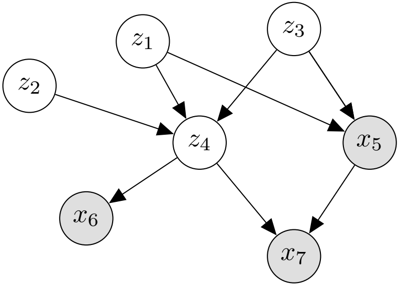
- A PGM specifies a **joint distribution** over variables and how it factorizes:

$$p(a, \mathbf{b}, \mathbf{c}) = p(a) \prod_{n=1}^{N} p(b_n|a)\, p(c_n|b_n)$$

where $\mathbf{b} = \{b_n\}_{n=1}^N$ and $\mathbf{c} = \{c_n\}_{n=1}^N$

From PGMs to joint distributions

- Another example

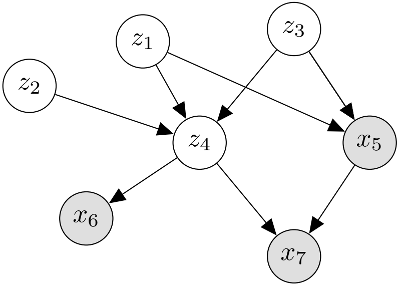


- Corresponding factorization of the joint distribution:

$$p(z_1, z_2, z_3, z_4, x_5, x_6, x_7) = ?$$

From PGMs to joint distributions

- Another example



- Corresponding factorization of the joint distribution:

$$\begin{aligned} p(z_1, z_2, z_3, z_4, x_5, x_6, x_7) &= p(z_1)\, p(z_2)\, p(z_3)\, p(z_4|z_1, z_2, z_3) \\ &\quad \times p(x_5|z_1, z_3)\, p(x_6|z_4)\, p(x_7|z_4, x_5) \end{aligned}$$

Why are factorizations so important?

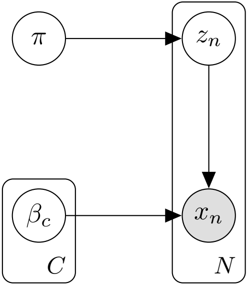
- Consider the joint distribution $p(z_1, z_2, z_3, z_4, x_5, x_6, x_7)$
- Assume each variable is binary
- How many parameters do we need to represent $p(z_1, z_2, z_3, z_4, x_5, x_6, x_7)$? You can think of it as a huge table...
  - You need $2^7 - 1 = 127$ parameters (entries in that table)!
- How about for the factorized version?

$$p(z_1)\, p(z_2)\, p(z_3)\, p(z_4|z_1, z_2, z_3)\, p(x_5|z_1, z_3)\, p(x_6|z_4)\, p(x_7|z_4, x_5)$$

  - Just $1 + 1 + 1 + 2^3 + 2^2 + 2 + 2^2 = 21$ parameters!
  - Much more efficient, right?
    We are exploiting the **conditional independencies** between the variables

Playtime

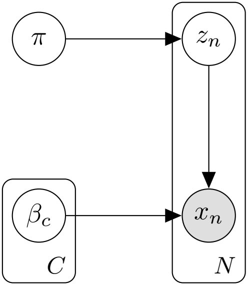
- What is the factorization of the joint distribution corresponding to this PGM?



$$p(\pi, \boldsymbol{\beta}, \mathbf{z}, \mathbf{x}) = ?$$

Playtime

- What is the factorization of the joint distribution corresponding to this PGM?

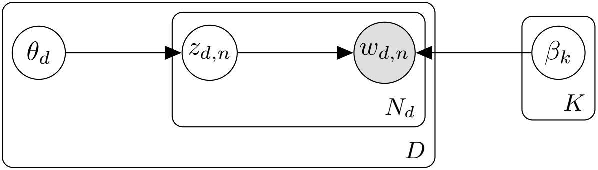


$$p(\pi, \boldsymbol{\beta}, \mathbf{z}, \mathbf{x}) = p(\pi) \left(\prod_{c=1}^{C} p(\beta_c) \right) \prod_{n=1}^{N} p(z_n|\pi)\, p(x_n|z_n, \boldsymbol{\beta})$$

Careful with the parenthesis!

Playtime

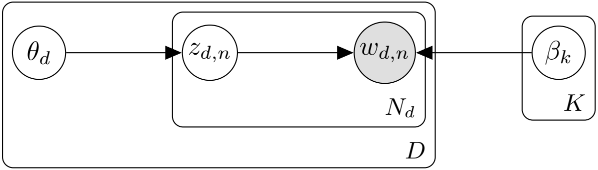
- What is the factorization of the joint distribution corresponding to this PGM?



$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = ?$$

Playtime

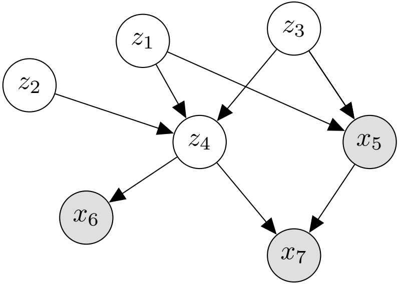
- What is the factorization of the joint distribution corresponding to this PGM?



$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \left(\prod_{k=1}^{K} p(\beta_k) \right) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{d,n} | \theta_d) \, p(w_{d,n} | z_{d,n}, \boldsymbol{\beta})$$

Inference

- **Model + Data → Insights**
- Answer various types of questions about the data by computing the posterior distribution of the latent variables given the observed ones



- Example: $p(z_2|x_5, x_6, x_7) = ?$

Two fundamental rules that you should always remember...

- Product rule of probability (or chain rule)

$$p(x, z) = p(x|z)\, p(z)$$

- Sum rule of probability (or marginalization rule)

$$p(x) = \sum_z p(x, z)$$

...or, if z is continuous

$$p(x) = \int p(x, z)\, dz$$

- This is also called *marginalizing over* z. But more on that later... :-)

Inference

- **Exact** inference
  - Set of latent variables $\mathbf{z} = \{z_m\}_{m=1}^{M}$
  - Observed variables $\mathbf{x} = \{x_n\}_{n=1}^{N}$
  - Using Bayes' theorem, the **posterior distribution** of $\mathbf{z}$ can be computed as

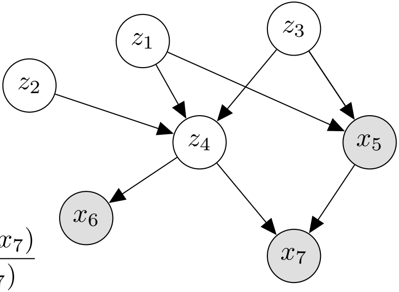
$$\overbrace{p(\mathbf{z}|\mathbf{x})}^{\text{posterior}} = \frac{\overbrace{p(\mathbf{x},\mathbf{z})}^{\text{joint}}}{p(\mathbf{x})} = \frac{\overbrace{p(\mathbf{x}|\mathbf{z})}^{\text{likelihood}}\ \overbrace{p(\mathbf{z})}^{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{evidence}}}$$

  - The **model evidence** , or marginal likelihood, can be computed by making use of the sum rule of probability to give

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})$$

Inference

- Returning to the previous example...

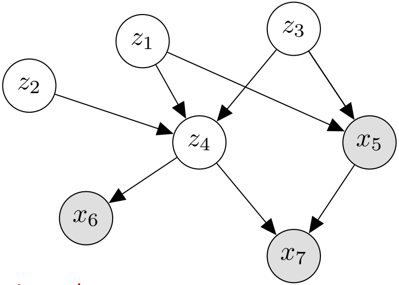


- Assuming discrete variables:

$$p(z_2|x_5,x_6,x_7) = \frac{p(z_2,x_5,x_6,x_7)}{p(x_5,x_6,x_7)}$$
$$\propto \sum_{z_1}\sum_{z_3}\sum_{z_4} p(z_1,z_2,z_3,z_4,x_5,x_6,x_7)$$

- In this case, it can be computed exactly (using the sum rule of probability)!
- Notice that, in this case, we don't need to compute $p(x_5,x_6,x_7)$!
  We can just renormalize the numerator in the end
  (hence the "proportional to" sign)

Inference



- What if x_6 is missing?
  - No problem! Just marginalize over its values

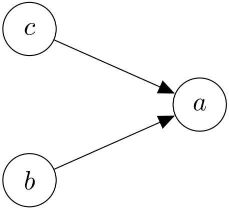
$$p(z_2|x_5,x_7) = \frac{p(z_2,x_5,x_7)}{p(x_5,x_7)}$$
$$\propto \sum_{z_1}\sum_{z_3}\sum_{z_4}\sum_{x_6} p(z_1,z_2,z_3,z_4,x_5,x_6,x_7)$$

  - Corresponds to considering all possible values of x_6 and averaging over them (weighted by their respective probabilities)
- PGMs provide a consistent way of handling **missing data**

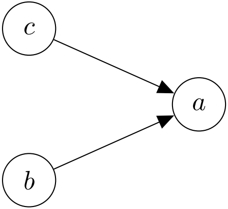
Conditional Probability Tables (CPTs)

- For now, our PGMs have only discrete random variables
- Each node has associated a **Conditional Probability Table**
  - It maps all possible values of its incoming set of arcs...
  - ...to all possible values of the node itself

- For example



Conditional Probability Tables (CPTs)



- This relationship could be defined by:

	$a=0$	$a=1$
$b=0, c=0$	0.7	0.3
$b=0, c=1$	0.3	0.7
$b=1, c=0$	0.5	0.5
$b=1, c=1$	0.1	0.9

Table: $p(a|b,c)$

- It factorizes as:

$$p(a,b,c) = p(a|b,c)\,p(b)\,p(c)$$

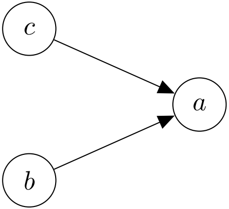
$b=0$	$b=1$
0.4	0.6

Table: $p(b)$

$c=0$	$c=1$
0.7	0.3

Table: $p(c)$

Conditional Probability Tables (CPTs)



- Joint distribution factorizes as:

$$p(a,b,c) = p(a|b,c)\,p(b)\,p(c)$$

- Imagine we observe $b=1$. Let's calculate $p(a|b=1)$

$$
\begin{aligned}
p(a|b=1) &= \frac{\sum_c p(a,b=1,c)}{p(b=1)} = \frac{\sum_c p(a|b=1,c)\,\cancel{p(b=1)}\,p(c)}{\cancel{p(b=1)}} \\
&= \sum_c p(a|b=1,c)\,p(c) \\
&= p(a|b=1,c=0)\,p(c=0) + p(a|b=1,c=1)\,p(c=1) \\
&= p(a|b=1,c=0) \times 0.7 + p(a|b=1,c=1) \times 0.3
\end{aligned}
$$

Conditional Probability Tables (CPTs)

$$p(a|b=1) = p(a|b=1, c=0) \times 0.7 + p(a|b=1, c=1) \times 0.3$$

- Considering $p(a|b,c)$:

	$a=0$	$a=1$
$b=0, c=0$	0.7	0.3
$b=0, c=1$	0.3	0.7
$b=1, c=0$	0.5	0.5
$b=1, c=1$	0.1	0.9

- We have:

$$p(a=1|b=1) = 0.5 \times 0.7 + 0.9 \times 0.3 = 0.62$$
$$p(a=0|b=1) = 0.5 \times 0.7 + 0.1 \times 0.3 = 0.38$$

- Thus $p(a|b=1)$ will be:

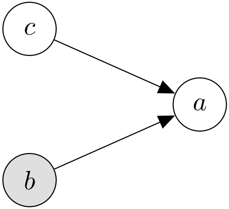
$a=0$	$a=1$
0.38	0.62

Playtime!

- Solve $p(c|b = 1)$
- Estimated time: 20 min

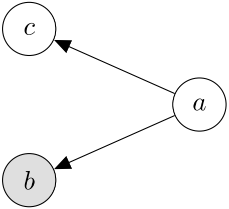
Conditional Probability Tables (CPTs)

- Indeed, b and c are independent.
  Just look at the factorization...



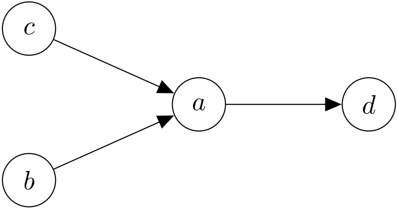
$$p(a, b, c) = p(a|b, c)\, p(b)\, p(c)$$

- What if we had this instead?



$$p(a, b, c) = p(b|a)\, p(c|a)\, p(a)$$

Conditional Probability Tables (CPTs)

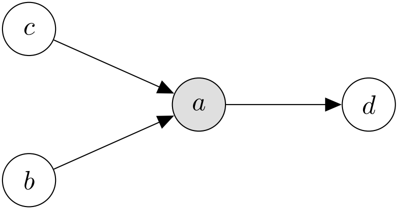


- Another relationship, another CPT:

	$d = 0$	$d = 1$
$a = 0$	0.6	0.4
$a = 1$	0.2	0.8

Table: $p(d|a)$

Conditional Probability Tables (CPTs)



- Another relationship, another CPT:

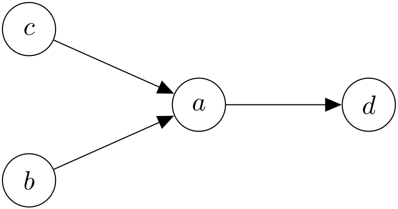
	$d = 0$	$d = 1$
$a = 0$	0.6	0.4
$a = 1$	0.2	0.8

Table: $p(d|a)$

- If a is observed, then we can get the distribution of d directly
- We can conclude that $d \perp\!\!\!\perp b, c \,|\, a$
- And also $p(a, b, c, d) = p(a|b, c)\, p(d|a)\, p(b)\, p(c)$

Conditional Probability Tables (CPTs)

- Full PGM:



- Another relationship, another CPT:

	$d = 0$	$d = 1$
$a = 0$	0.6	0.4
$a = 10$	0.2	0.8

Table: $p(d|a)$

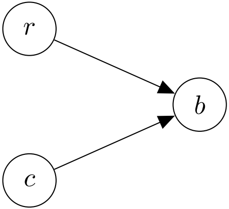
- Of course, if we do **not** observe a, then d will depend on the values of b and c

Some useful rules

	We want	We have	We do
1.	$p(a,b)$	$p(a,b,c)$	$p(a,b) = \sum_c p(a,b,c)$
2.	$p(a\|b,c)$	$p(a,b,c)$	$p(a\|b,c) = \frac{p(a,b,c)}{\sum_a p(a,b,c)}$
3.	$p(a\|b)$	$p(a,b,c)$	$p(a\|b) = \frac{\sum_c p(a,b,c)}{\sum_c \sum_a p(a,b,c)}$
4.	$p(a\|b)$	$p(b\|a), p(a)$	$p(a\|b) = \frac{p(b\|a)\, p(a)}{\sum_a p(b\|a)\, p(a)}$
5.	$p(a\|b)$	$p(a\|b,c), p(c)$	$p(a\|b) = \sum_c p(a\|b,c)\, p(c)$

- Note that these are just applications of the sum and product rules of probability!

Travel mode choice - a possible story



- Every day, John needs to decide whether to go to work by bike $(b = 1)$, or just take his car $(b = 0)$?
  - It depends on whether he has schedule constraints $(c = 1)$: e.g. a meeting far away may imply the need for a car
  - It depends on whether it rains $(r = 1)$, or not $(r = 0)$

$r = 1$	$r = 0$
0.7	0.3

Table: $p(r)$

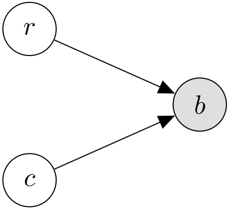
$c = 1$	$c = 0$
0.3	0.7

Table: p(c)

	$b = 1$	$b = 0$
$c = 1, r = 1$	0.1	0.9
$c = 1, r = 0$	0.2	0.8
$c = 0, r = 1$	0.3	0.7
$c = 0, r = 0$	0.8	0.2

Table: $p(b|c, r)$

Mode choice - a possible story



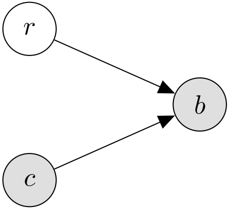
- We observe that he took his car $(b = 0)$
- What is the probability that it is raining?
  - $p(r = 1|b = 0) = ?$
- Notice that we have $p(b, r, c) = p(b|r, c)\, p(r)\, p(c)$

$$p(r = 1|b = 0) \overset{1,2}{=} \frac{\sum_{c\in\{0,1\}} p(b = 0, r = 1, c)}{p(b = 0)} = \frac{\sum_{c\in\{0,1\}} p(b = 0|r = 1, c)\, p(r = 1)\, p(c)}{p(b = 0)}$$

$$\overset{3}{=} \frac{\sum_{c\in\{0,1\}} p(b = 0|r = 1, c)\, p(r = 1)\, p(c)}{\sum_{r\in\{0,1\}} \sum_{c\in\{0,1\}} p(b = 0, r, c)} = \frac{\sum_c p(b = 0|r = 1, c)\, p(r = 1)\, p(c)}{\sum_r \sum_c p(b = 0|r, c)\, p(c)\, p(r)}$$

$$= \frac{(0.9 \times 0.3 + 0.7 \times 0.7) \times 0.7}{(0.9 \times 0.3 + 0.7 \times 0.7) \times 0.7 + (0.8 \times 0.3 + 0.2 \times 0.7) \times 0.3} = \frac{0.532}{0.646} = 0.824$$

Explaining away



- What if we **also** observe that the schedule is constrained, $c = 1$
- Should the probability that it is raining change?...
  - $p(r = 1|b = 0, c = 1) = ?$

$$p(r = 1|b = 0, c = 1) = \frac{p(b = 0|r = 1, c = 1)\, p(r = 1)\, \cancel{p(c = 1)}}{\sum_{r\in\{0,1\}} p(b = 0|r, c = 1)\, p(r)\, \cancel{p(c = 1)}}$$

$$= \frac{p(b = 0|r = 1, c = 1)\, p(r = 1)}{\sum_r p(b = 0|r, c = 1)\, p(r)} = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.8 \times 0.3)} = \frac{0.63}{0.87} = 0.72$$

- What happened is that knowing the choice of car (b=0) was **explained away** by the fact that the schedule is constrained/tight
- As if you believe less that John does not pick the bike due to the rain

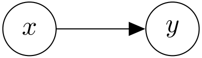
Independence properties

- Just by analysing the representation, we simplify the calculations!
  - Observed data vs Latent variables (color of node)
  - Arrow directions
  - Conditional independence rules (D-separation)
- The Bayesian network assumption says:

"Each variable is conditionally independent of its non-descendants, given its parents"

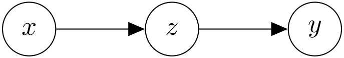
D-separation

- When does x influence y?
- Direct connection:

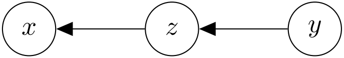


- Indirect connection:

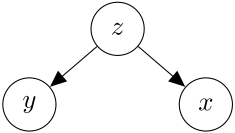
Causal effect:



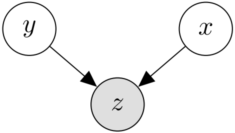
Evidential effect:



Common (latent) cause:



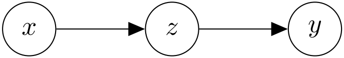
Common (observed) effect:



D-separation

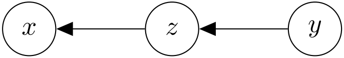
- When influence can flow from x to y via z, we say that the trail x, y, z is *active* (otherwise, it is *blocked*)

Causal trail:



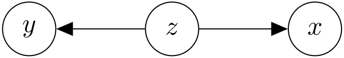
Active iff z is not observed

Evidential trail:



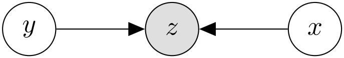
Active iff z is not observed

Common cause:



Active iff z is not observed

Common effect:



Active iff z **or one of its descendants are observed**

D-separation: a simple(r) algorithm

For any expression "is $\mathbf{x}$ independent of $\mathbf{y}$ given $\mathbf{z}$" (formally, $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$)[1]

1. Draw the *ancestral graph*
    - It is the part of the original graph that has only the variable sets $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, and all their ancestors among them
2. *Moralize* the graph by *marrying* the parents
    - For each pair of variables with a common child, draw an undirected edge (line) between them (if a variable has more than two parents, draw lines between every pair of parents)
3. *Disorient* the graph by replacing all edges for undirected ones
4. Delete the variables $\mathbf{z}$ (and any other observed variables not explicitly included in $\mathbf{z}$), and their edges

[1] Note that $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ can themselves be sets of variables!

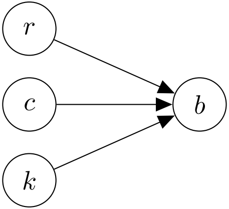
D-separation: a simple(r) algorithm (cont.)

Analysis of the result

- If **x** and **y** are **disconnected**, then they are conditionally independent given **z**!
  - Being disconnected means that there is no possible path between **x** and **y** in the resulting graph
- Otherwise, they are not proven to be independent

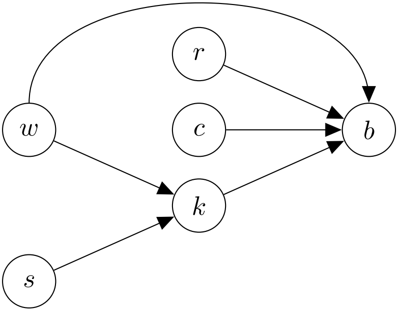
Mode choice - a possible story

- Every day, John needs to decide whether to go to work by bike $(b = 1)$, or to just take his car $(b = 0)$?
    - It depends on whether he has schedule constraints $(c = 1)$: e.g. a meeting far away may imply the need for a car
    - It depends on whether it rains $(r = 1)$, or not $(r = 0)$
    - It also depends on whether he needs to pickup and drop off his kids, k

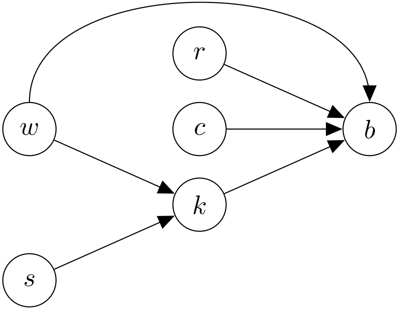


Mode choice - a possible story

- We can dig further in this problem...
  - If there is no school on that day ($s = 0$), he probably won't need to bring his kids at all
  - His wife w, may bring the kids
  - His wife may need to take the car (in which case, he has to take the kids by bike)



Playtime!



- Using the D-separation algorithm, try to prove that:

$$s \perp\!\!\!\perp b \,|\, k$$

$$s \perp\!\!\!\perp b \,|\, \{k, w\}$$

$$\{r, c\} \perp\!\!\!\perp k$$

$$\{r, c\} \perp\!\!\!\perp s \,|\, \{k, b\}$$

- Estimated time: 20 min

Readings

- Main reading: Chapter 8: “Graphical Models”, pages 359-362 and pages 372-379 of Chris Bishop’s book, “Pattern Recognition and Machine Learning” (PRML) Further readings:
- “D-separation: How to determine which variables are independent in a Bayes net”. Jessica Noss. EECS MIT. http://web.mit.edu/jmn/www/6.034/d-separation.pdf
- Chapter 10: “Directed graphical models”, pages 307-311 and pages 324-327 of Kevin Murphy’s book “Machine Learning: A Probabilistic Perspective”
- Koller, D., and Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.