ETH zürich

Directed acyclic graph (DAG) models

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Exchange students who will be away in August need to request a “Distance Examination“
  - Contact me if you need further details
- Series 1 will be uploaded later today

Last week

- Controlled experiments vs. observational studies
- Simpson's paradox

Salk Vaccine Field Trial

- Design 2
  - Grade 2: vaccine if parents consent (treatment)
  - Grade 2: no vaccine if no parental consent (control)
  - Grades 1 + 3: no vaccine (control)

Population

Treatment

Control

No consent

Consent

→ at risk (higher income)

↳ hard to draw conclusions

Salk Vaccine Field Trial

- Lesson: Treatment and control groups should be as similar as possible

Population → Consent, No consent

Consent → Treatment, Control

↳ control is chosen from the same population as the treatment group

Salk Vaccine Field Trial

- Design 3
  - Need a control and a treatment group from the same population
  - Only consider children of consenting parents
  - Randomize: 50% chance of being put in the control or the treatment group

  - Double-blinding:
    - Give placebo to control group and don’t tell anyone whether they are in control or treatment group
    - Ensure that effect is due to vaccine and not due to the “idea of getting treatment”
    - Doctors (who decide whether child contracted polio during the experiment) were not told whether a child got real vaccine or placebo

  - Randomized controlled double-blind experiment

Internal vs external validity

- Internal validity: Validity of conclusions drawn within context of particular study

Population

Consent

No consent

Treatment

Control

- External validity: Generalizability of empirical findings to new environments, settings or populations

Internal vs external validity

- Internal validity: Validity of conclusions drawn within context of particular study
  - Conclusions from experiment are applicable to experimental units used in the experiment
  - If units are a representative sample from some population of units, conclusions also valid
  - Best design: double-blind randomized controlled trial (RCT)
  - External validity limited by internal validity: If a causal conclusion drawn within a study is invalid, then generalizations of that inference to other contexts will also be invalid

Internal vs external validity

- External validity: Generalizability of empirical findings to new environments, settings or populations
  - Threats:
    - Only particular subpopulation (e.g. college students, volunteers, …)
    - Situation (e.g. lighting, noise, treatment administration, investigator, timing, …)
    - …
- Transportability: “license” to transfer causal effects learned in experimental studies to a new population
  - Sometimes possible under causal assumptions (Pearl and Bareinboim (2014))
    - Recalibration, transport formulas
  - Need to characterize commonalities and differences between populations

Salk Vaccine Field Trial

Design 2:

	Size	Rate*
Grade 2 (consent)	225’000	25
Grades 1 & 3	725’000	54
Grade 2 (no consent)	125’000	44

*(= per 100’000)

Design 3: RCT

	Size	Rate*
Treatment (consent)	200’000	28
Control (consent)	200’000	71
No consent	350’000	$\boxed{46}$

*(= per 100’000)

- Design 2 biased against the vaccine
- Design 3 shows effectiveness of vaccine

→ vaccine is effective but effect is underestimated

Simpson's paradox: Kidney stones

	Treatment A	Treatment B
Patients with small stones	**93%** (81/87)	87% (234/270)
Patients with large stones	**73%** (192/263)	69% (55/80)
Overall	78% (273/350)	**83%** (289/350)

Recovery rates

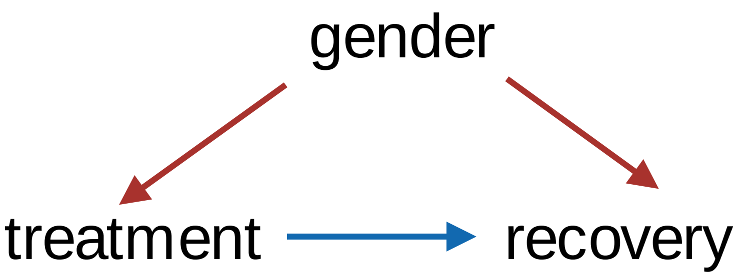
- Treatment A: Open surgery
- Treatment B: Percutaneous nephrolithotomy (less invasive treatment)

$$\begin{tikzcd} & S \arrow[dl] \arrow[dr] & \\ T \arrow[rr] & & R \end{tikzcd}$$

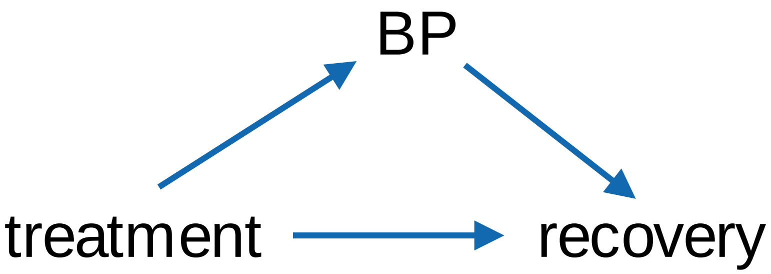
Charig et al. (1986)

Simpson’s paradox and causal diagrams

- Same numbers, different conclusions….
  - Must use additional information: “story behind the data”, causal assumptions
- Consider total causal effect of treatment on recovery
  - Possible scenarios:



gender is a confounder;
control for gender

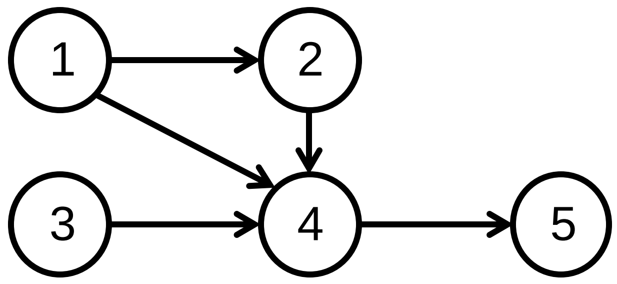


BP is an intermediate variable;
don’t control for BP

Today

- Internal vs. external validity
- Graph terminology
- Directed acyclic graph (DAG) models
- Markov properties
- d-separation
- Probabilistic reasoning using DAG models

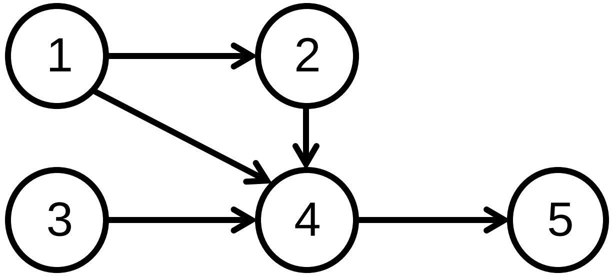
Graph terminology



- A graph $G = (\boldsymbol{V}, \boldsymbol{E})$ consists of vertices (nodes) $\boldsymbol{V}$ and edges $\boldsymbol{E}$
- There is at most one edge between every pair of vertices
- Two vertices are adjacent if there is an edge between them
- If all edges are directed $(i \rightarrow j)$, the graph is called directed
- A path between i and j is a sequence of distinct vertices $(i, \dots, j)$ such that successive vertices are adjacent
- A directed path from i to j is a path between i and j where all edges are pointing towards j, i.e., $i \rightarrow \cdots \rightarrow j$

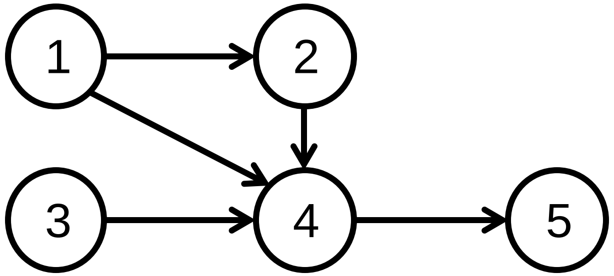
Graph terminology

- A cycle is a path $(i, j, \ldots, k)$ plus an edge between k and i
- A directed cycle is a directed path $(i, j, \ldots, k)$ from i to k, plus an edge $k \rightarrow i$
- A directed acyclic graph (DAG) is a directed graph without directed cycles



Example

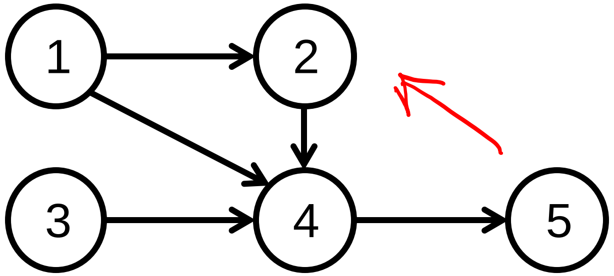
- Which of the following are (directed) paths?
  - (4,2,1,3) Not a path! (1 and 3 not adjacent)
  - (3,4,2,1,4,5) Not a path! (vertices are not distinct)
  - (3,4,2,1) A path but not directed
  - (3,4,5) A directed path



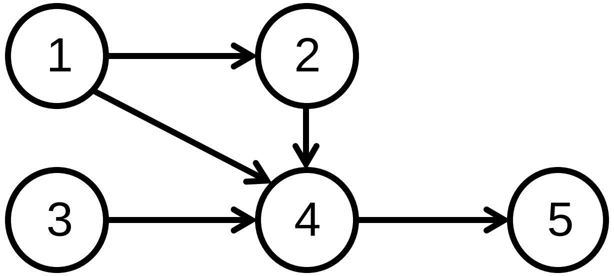
Example

- $(1,2,4,1)$ is a cycle, but not a directed cycle

- This graph is a DAG
- Adding the edge $5 \rightarrow 2$ yields a directed cycle $(2,4,5,2)$
  - Then no longer a DAG



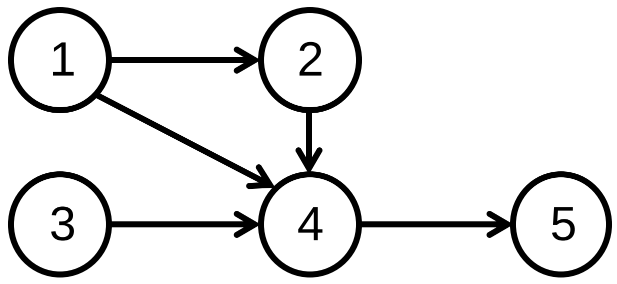
Graph terminology



- If $i \to j$, then i is a parent of j, and j is a child of i
- If there is a directed path from i to j, then i is an ancestor of j and j is a descendant of i
- Each vertex is also an ancestor and descendant of itself
- The sets of parents, children, descendants and ancestors of i in G are denoted by $\mathrm{pa}(i, G)$, $\mathrm{ch}(i, G)$, $\mathrm{desc}(i, G)$, $\mathrm{an}(i, G)$
- We omit G if the graph is clear from the context

Graph terminology

- We write sets of vertices in bold face
- The previous definitions are applied disjunctively to sets
  - Example: $\mathrm{pa}(\boldsymbol{S}) = \bigcup_{k \in \boldsymbol{S}} \mathrm{pa}(k)$

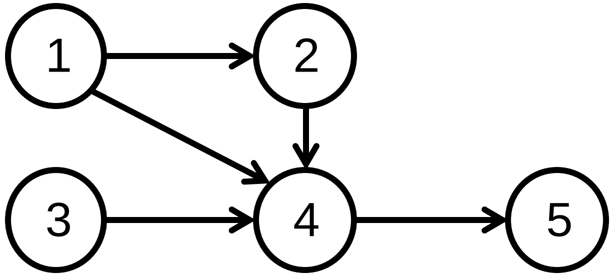


- The non-descendants of $\boldsymbol{S}$ are the complement of $\mathrm{desc}(\boldsymbol{S})$:

$$\mathrm{nondesc}(\boldsymbol{S}) \coloneqq \boldsymbol{V} \setminus \mathrm{desc}(\boldsymbol{S})$$

Example

- 1 is a parent of 4, and 4 is a child of 1
- 1 is an ancestor of 5, and 5 is a descendant of 1
- Determine children, parents, descendants and ancestors of 4:
  - $\mathrm{ch}(4) =$ $\{5\}$
  - $\mathrm{pa}(4) =$ $\{1, 2, 3\}$
  - $\mathrm{desc}(4) =$ $\{4, 5\}$
  - $\mathrm{an}(4) =$ $\{1, 2, 3, 4\}$



Example

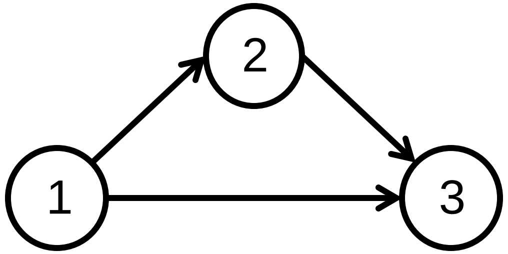
- pa({2,4}) = $\{1\} \cup \{1,2,3\} = \{1,2,3\}$
- desc({2,3}) = $\{2,4,5\} \cup \{3,4,5\} = \{2,3,4,5\}$
- nondesc({2,3}) = $\{1,2,3,4,5\} \setminus \{2,3,4,5\} = \{1\}$

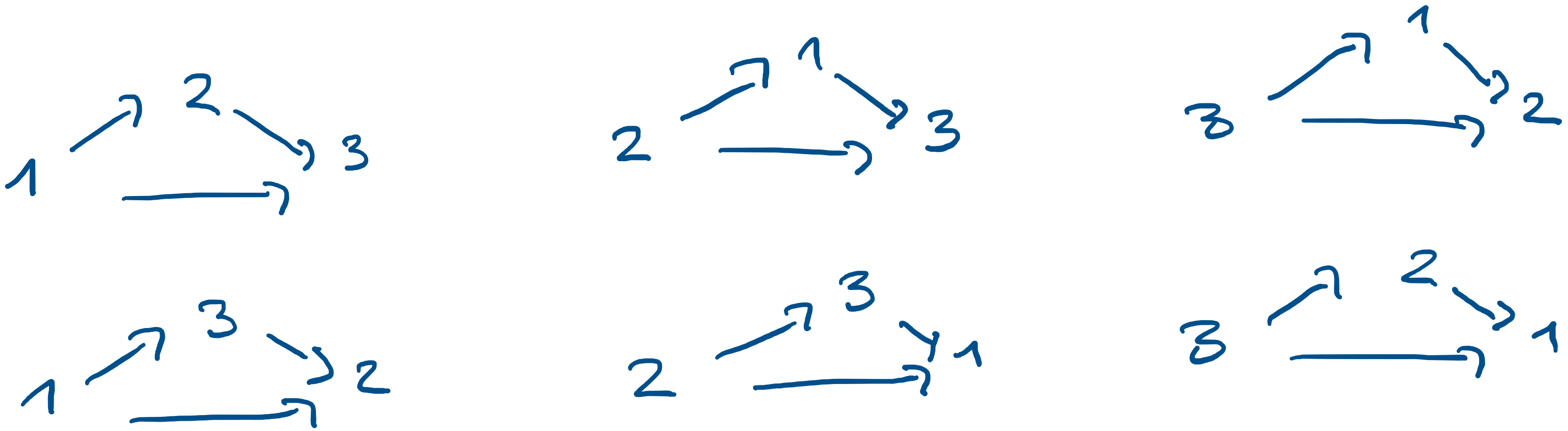
$$
\begin{tikzcd}
1 \arrow[r] \arrow[dr] & 2 \arrow[d] & \\
3 \arrow[r] & 4 \arrow[r] & 5
\end{tikzcd}
$$

Graph terminology

- We call G fully connected if all pairs of nodes are adjacent.
- How many possibilities?

"full DAG"





$p!$ possibilities $\quad p = |V|$

Clicker question – Number of DAGs

$\binom{p}{2} = \frac{p(p-1)}{2}$ distinct pairs of nodes $\quad p = 3 \quad \binom{p}{2} = 3$

Ignoring acyclicity:

3 options for every pair A, B

A B

$A \to B$

$A \leftarrow B$

$$3^{p(p-1)/2} = 3^3 = 27$$

Subtract 2 graphs with cycles $1 \to 2 \to 3 \to 1$ $\quad 1 \to 3 \to 2 \to 1$

$\Rightarrow$ 25 DAGs with 3 nodes

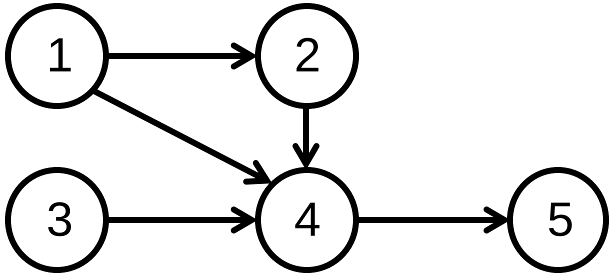
Number of DAGs with p nodes

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

Table 1.2: The number of DAGs depending on the number p of nodes, taken from `http://oeis.org/A003024` (Feb 2015).

DAGs and random variables

- Each vertex represents a random variable: vertex i represents random variable X_i
- If $\boldsymbol{A} \subseteq \boldsymbol{V}$, then $X_{\boldsymbol{A}} \coloneqq \{X_i : i \in \boldsymbol{A}\}$



- Edges denote relationships between pairs of variables (we will make this more precise)

Factorization of the joint density

- We can connect a distribution to a DAG in the following way:
- We always have:

$$f(x_1, \dots, x_p) = f(x_1)f(x_2|x_1) \dots f(x_p|x_1, \dots x_{p-1})$$

"chain rule of probability"

- A set of variables $X_{\mathrm{pa}(j)}$ is said to be Markovian parents of X_j if it is a minimal subset of $\{X_1, \dots X_{j-1}\}$ such that $f(x_j|x_1, \dots, x_{j-1}) = f(x_j|x_{\mathrm{pa}(j)})$
  - Note: Markovian parents depend on the chosen ordering of the variables.
- Then $f(x_1, \dots, x_p) = \prod_{j=1}^{p} f(x_j|x_{\mathrm{pa}(j)})$ "factorization property"
- We can draw a DAG accordingly; the distribution is said to factorize according to this DAG.

Example

- Consider (X_1, X_2, X_3) and suppose that $X_1 \perp\!\!\!\perp X_3 | X_2$ is the only (conditional) independence:

$$f(x_1|x_2,x_3) = f(x_1|x_2) \text{ and } f(x_3|x_1,x_2) = f(x_3|x_2)$$

  - Then $f(x_1,x_2,x_3) = f(x_1)f(x_2|x_1)f(x_3|x_1,x_2) = f(x_1)f(x_2|x_1)f(x_3|x_2)$
  - DAG: $1 \to 2 \to 3$
- Or
  - $f(x_3,x_2,x_1) = f(x_3)f(x_2|x_3)f(x_1|x_2,x_3) = f(x_3)f(x_2|x_3)f(x_1|x_2)$
  - DAG: $3 \to 2 \to 1$
- Or
  - $f(x_1,x_3,x_2) = f(x_1)f(x_3|x_1)f(x_2|x_1,x_3)$
  - DAG: $$\begin{tikzcd} 1 \arrow[r] \arrow[rr, bend right] & 3 \arrow[r] & 2 \end{tikzcd}$$

Factorization of the joint density

- A distribution can factorize according to several DAGs
- Every distribution factorizes according to a full DAG
  - Note: there are *p!* possibilities

- Sometimes a distribution factorizes according to a sparse DAG, i.e., a DAG with few edges.
  - E.g. first-order Markov chain:
    - $f(x_1, \dots, x_p) = f(x_1) f(x_2|x_1) \dots f(x_p|x_1, \dots, x_{p-1}) = f(x_1) f(x_2|x_1) \dots f(x_p|x_{p-1})$
    - DAG: $1 \rightarrow 2 \rightarrow \cdots \rightarrow p$

DAG models

- A DAG model or Bayesian network is a combination (G, f), where G is a DAG and f is a distribution that factorizes according to G

- DAG models can be used for various purposes:
  - Estimating the joint density from low order conditional densities
  - Reading off conditional independencies from the DAG
  - Probabilistic reasoning (expert systems)
  - Causal inference

Estimating the joint density

- Estimating the joint density of many variables is generally difficult.
  - Example: The joint distribution of p binary variables requires $2^p - 1$ parameters.

1 binary variable

0	1
$1-p$	p

1 parameter

2 binary variables

0		
1		

3 parameters

3 binary variables

$x_1 \to x_2 \to x_3$

$p(x_1)\,p(x_2|x_1)\,p(x_3|x_1,x_2)$

1 par 2 par 4 par 7 parameters

⋮

Estimating the joint density

- Estimating the joint density of many variables is generally difficult.
  - Example: The joint distribution of p binary variables requires $2^p - 1$ parameters.

- But if you know that the distribution factorizes according to a DAG, then you only need to estimate $f(x_i | x_{\text{pa}(i)})$ for $i = 1, \dots, p$.

- If the parent sets are small, this means we only need to estimate low order conditional densities.

Clicker question – Estimating the joint density

$p(x_1)\, p(x_2 | x_1)\, p(x_3 | x_2) \cdots p(x_p | x_{p-1})$

1 par 2 par 2 par - - - 2 par

$2p-1$ parameters

DAG models

- A DAG model or Bayesian network is a combination (G, f), where G is a DAG and f is a distribution that factorizes according to G

- DAG models can be used for various purposes:
  - Estimating the joint density from low order conditional densities
  - Reading off conditional independencies from the DAG
  - Probabilistic reasoning (expert systems)
  - Causal inference

Reading off conditional independencies: Markov property

- First-order Markov models: the future is independent of the past given the present

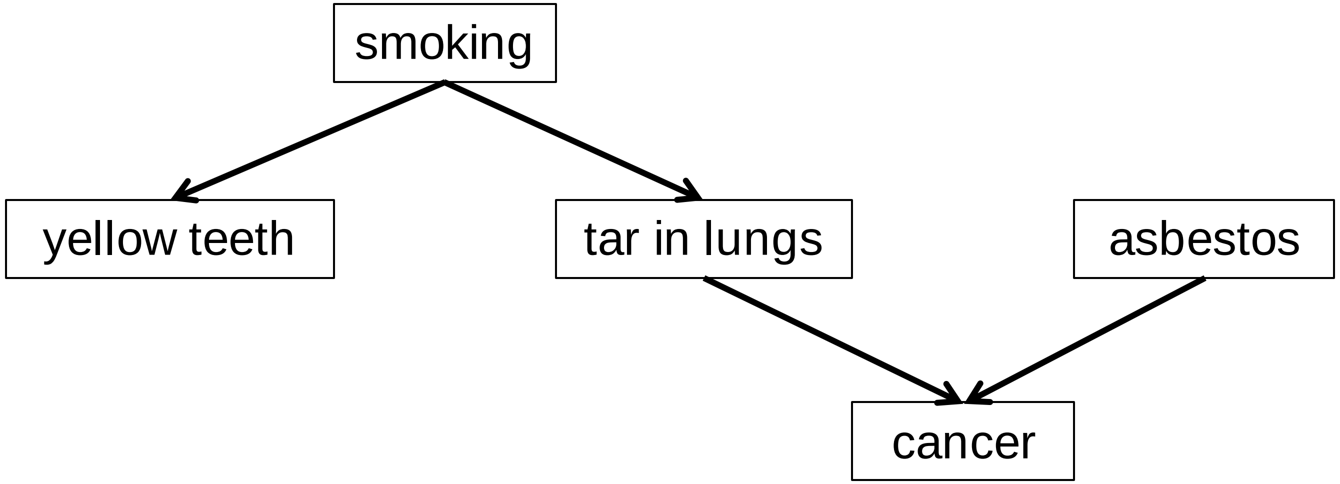
$$1 \to 2 \to \cdots \to (t-1) \to t \to (t+1)$$

$$X_{t+1} \perp\!\!\!\perp \{X_{t-1}, X_{t-2}, \ldots, X_1\} \mid X_t$$

- In DAG models, we have a similar (local) Markov property.
Let $\boldsymbol{S}$ be any collection of nodes. Then:

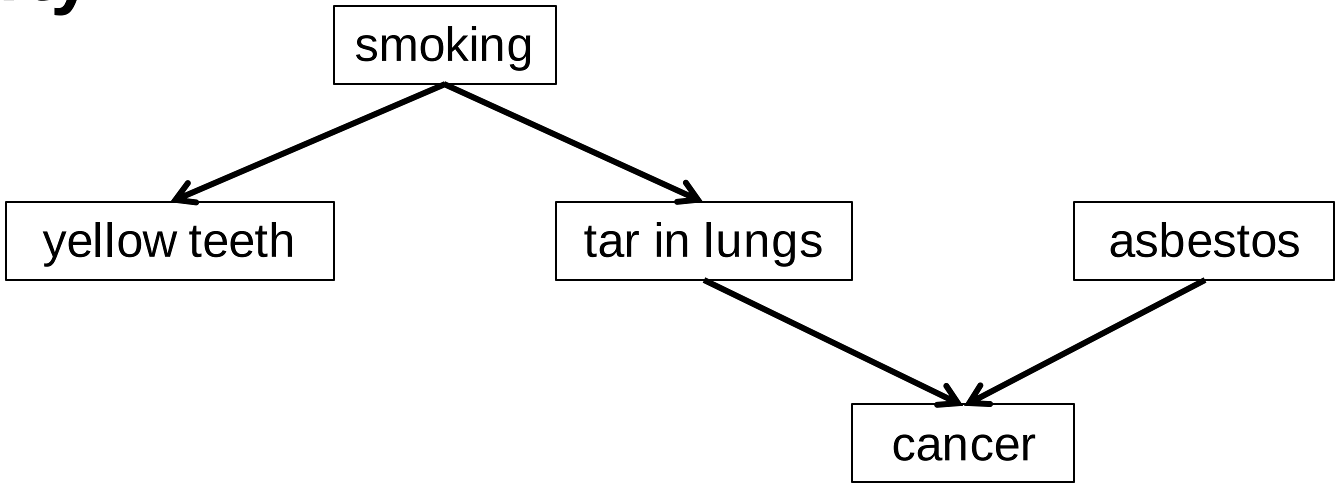
$$X_{\boldsymbol{S}} \perp\!\!\!\perp X_{\mathrm{nondesc}(\boldsymbol{S})\backslash \mathrm{pa}(\boldsymbol{S})} \mid X_{\mathrm{pa}(\boldsymbol{S})}$$

Example



- Take $\boldsymbol{S} = \{\text{yellow teeth}\}$ and apply the local Markov property
- Then:
  - $\text{pa}(\text{yellow teeth}) = \{\text{smoking}\}$
  - $\text{nondesc}(\text{yellow teeth}) = \{\text{smoking}, \text{tar}, \text{cancer}, \text{asbestos}\}$
- Hence, $\text{yellow teeth} \perp\!\!\!\perp \{\text{tar}, \text{cancer}, \text{asbestos}\} \mid \text{smoking}$ in any distribution that factorizes according to this DAG

Markov property



- Is $\text{tar} \perp\!\!\!\perp \text{asbestos} \mid \text{cancer}$?
- The local Markov property cannot be used to read off arbitrary conditional (in)dependencies. For this we have d-separation.

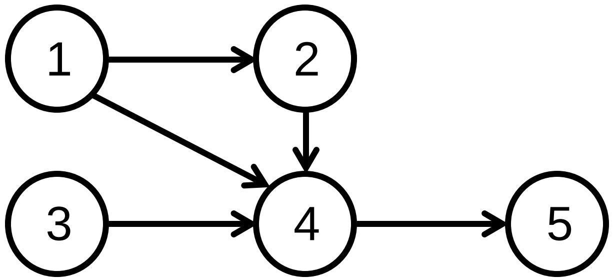
Graph terminology

- Need new terminology:
  - A non-endpoint node i is a collider on a path if the path contains $\rightarrow i \leftarrow$ (arrows collide at i).
  - Otherwise, it is a non-collider on the path.

- Is 4 a collider in the given graph? bad question

4 is a collider on the path (3, 4, 1)

4 is a non-collider on the path (3, 4, 5)



→ collider status is always relative to a path

d-separation

- A path between i to j is blocked by a set $\boldsymbol{S}$ (not containing i or j) if at least one of the following holds:
  - There is a non-collider on the path that is in $\boldsymbol{S}$; or
  - There is a collider on the path such that neither this collider nor any descendants are in $\boldsymbol{S}$.
- A path that is not blocked is active.

- If all paths between $i \in \boldsymbol{A}$ and $j \in \boldsymbol{B}$ are blocked by $\boldsymbol{S}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$. Otherwise they are d-connected given $\boldsymbol{S}$.

- Denote d-separation by $\perp$

Global Markov property

- Definition:
  A distribution P with density p satisfies the global Markov property with respect to a DAG G if:

$$\boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G \Rightarrow X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} \mid X_{\boldsymbol{S}} \text{ in } P$$

- Theorem (Pearl, 1988):
  A distribution P with density p satisfies the global Markov property with respect to G if and only if p factorizes according to G.

Example

Example

$X \rightarrow Z \rightarrow Y$ "Z blocks the path" $X \perp Y \mid Z$

"chain"

$X \leftarrow Z \leftarrow Y$

confounder

"fork" $X \leftarrow Z \rightarrow Y$ "Z blocks the path" $X \perp Y \mid Z$

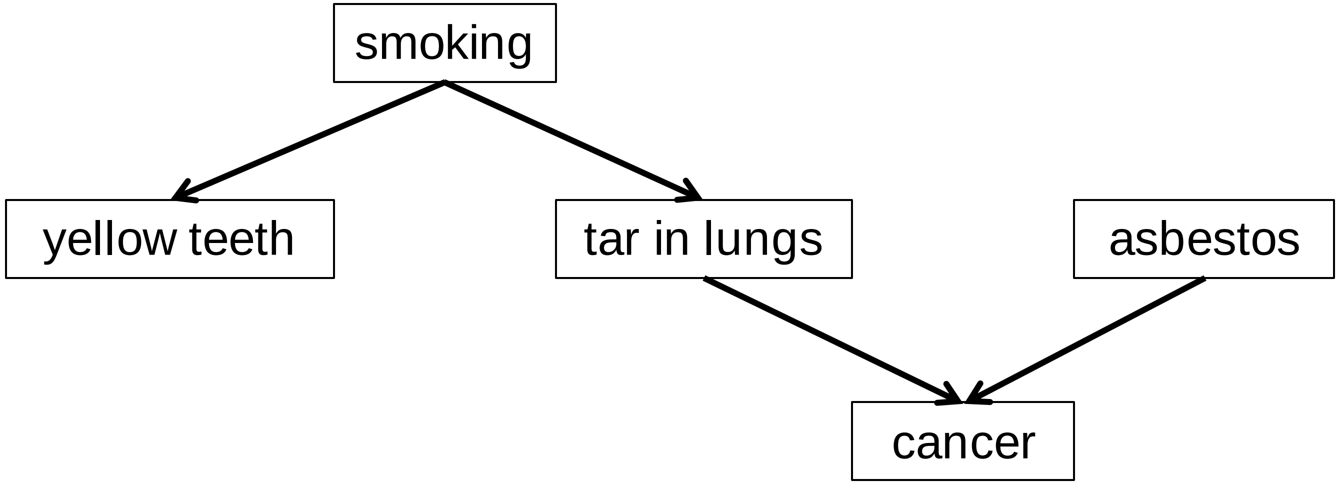
"v-structure"

"immorality"

$X \rightarrow Z \leftarrow Y$ "Z unblocks the path" $X \not\perp Y \mid Z$

collider

Example



- Which of the following hold?
  - yellow teeth $\perp$ cancer | smoking Yes
  - tar $\perp$ asbestos Yes
  - tar $\perp$ asbestos | cancer No
  - yellow teeth $\perp$ asbestos | cancer No

Clicker question – d-separation – Question 2

- Consider the following graph.

Is the path between i and j blocked by ...

$$i \rightarrow l \rightarrow k \leftarrow j$$

... $S = \{\}$? Yes.

... $S = \{l\}$? Yes.

... $S = \{k\}$? No.

... $S = \{k, l\}$? Yes. → only need to block the path somewhere!

DAG models

- A DAG model or Bayesian network is a combination (G, f), where G is a DAG and f is a distribution that factorizes according to G

- DAG models can be used for various purposes:
  - Estimating the joint density from low order conditional densities
  - Reading off conditional independencies from the DAG
  - Probabilistic reasoning (expert systems)
  - Causal inference

Probabilistic reasoning

- Conditional probabilities are rather counterintuitive for many people
- DAGs allow us to obtain conditional probabilities efficiently, using a “message passing” algorithm.
- We will apply this in R (without discussing the details behind the algorithms).

DAG models

- A DAG model or Bayesian network is a combination (G, f), where G is a DAG and f is a distribution that factorizes according to G

- DAG models can be used for various purposes:
  - Estimating the joint density from low order conditional densities
  - Reading off conditional independencies from the DAG
  - Probabilistic reasoning (expert systems)
  - Causal inference

Recap

- Concepts to know:
  - Internal vs. external validity
  - Graph terminology
  - Use cases of directed acyclic graph (DAG) models
  - Markov properties (local, global, factorization)
  - d-separation

References and acknowledgments

- Slides adapted from M. Maathuis
- Some examples from script by J. Peters & N. Meinshausen