CSC 580 Principles of Machine Learning

12 A closer look at PGMs; Hidden Markov Models

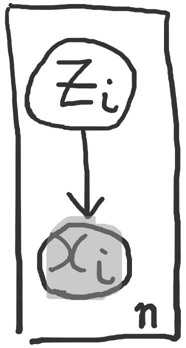
Chicheng Zhang

Department of Computer Science

THE UNIVERSITY OF ARIZONA

*slides credit: built upon CSC 580 Fall 2021 lecture slides by Kwang-Sung Jun

Background: A deeper look at conditional independence



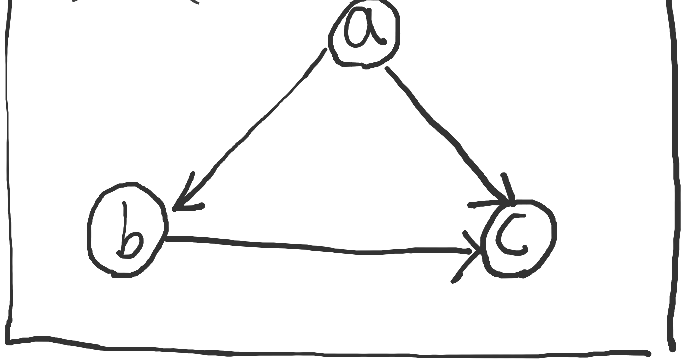
- Recall the graphical representation (plate notation) specifies the dependency
- More precisely, it specifies how a joint distribution can be factored in *a structured way*
- Remark: We focus on directed graphical models (Bayes nets)
  - another world: undirected models

- Intro example:
  - $P(A = a, B = b, C = c) = P(C = c \mid A = a, B = b) \cdot P(A = a, B = b)$

    $= P(C = c \mid A = a, B = b) \cdot P(B = b \mid A = a) \cdot P(A = a)$

  - Graphical representation:

  For each conditional distribution, add direct links from *the nodes being conditioned* to *the node whose distribution is of interest*



Warning: notation convention

- Notation easily gets overwhelming, no easy way out.
  - Fully-specified notation: explicit, but takes too long to process
  - Simplified notation: concise, but takes time to train yourself to be familiar
- Probabilistic models: For fully-specified notation, we always need to specify the random variable and the value that it takes separately.

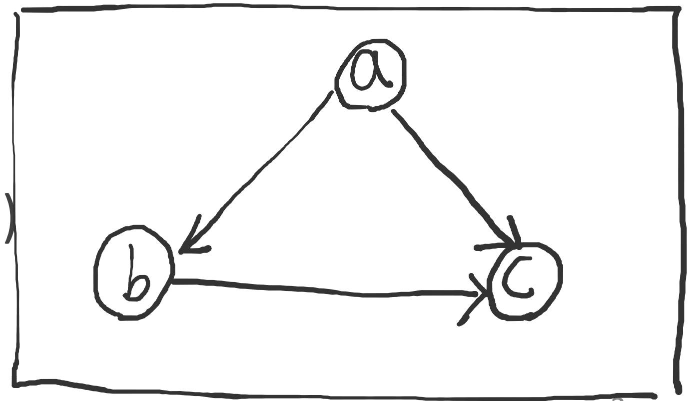
- E.g. $P(A = a, B = b, C = c) = P(\, C = c \mid A = a, B = b\,) \cdot P(A = a, B = b)$

$$= P(\, C = c \mid A = a, B = b\,) \cdot P(\, B = b \mid A = a\,) \cdot P(A = a)$$

- Simplified notation: $P(a, b, c) = P(\, c \mid a, b\,) \cdot P(a, b)$

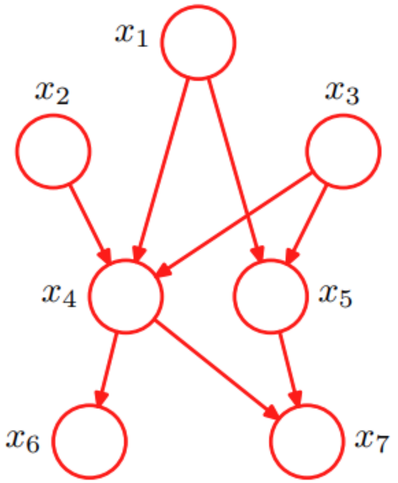
$$= P(\, c \mid a, b\,) \cdot P(\, b \mid a\,) \cdot P(a)$$

- i.e. reserve symbol a for values taken by random variable A (same for B, C)
- We will use simplified notation throughout this lecture



PGM: flexible modeling of data distributions

- Q: what kind of distribution does this graph represent?



- $P(x_1, x_2, \dots, x_7) = P(x_1)P(x_2)P(x_3)P(\, x_4 \mid x_1, x_2, x_3 \,) \cdot$
  $P(\, x_5 \mid x_1, x_3 \,)P(\, x_6 \mid x_4 \,)P(x_7 \mid x_4, x_5)$

- For a general directed acyclic graph (DAG) G with K nodes $x_1, \dots, x_K$,
  $P(x_1, x_2, \dots, x_K) = \prod_{k=1}^{K} P(x_k \mid \mathrm{pa}_k)$,

  Parent nodes of x_k in G

- Fact: this implicitly implies $P(\, x_k \mid \mathrm{pa}_k \,) = P(x_k \mid x_1, \dots, x_{k-1})$, i.e. $x_k \perp\!\!\!\perp \{x_1, \dots, x_{k-1}\} \setminus \mathrm{pa}_k \mid \mathrm{pa}_k$
  - E.g. $x_6 \perp\!\!\!\perp \{x_1, x_2, x_3, x_5\} \mid x_4$
- Edges oftentimes encode *causal relationships* between the node variables

Bayes net = DAG + Conditional probability table

- $P(x_1, x_2, \ldots, x_K) = \prod_{k=1}^{K} P(x_k \mid \mathrm{pa}_k)$ <- also need to specify each $P(x_k \mid \mathrm{pa}_k)$ respectively
- Aside: $J \perp\!\!\!\perp B, E \mid A$ => the effect of B, E to John's calling is "completely captured" in Alarm status

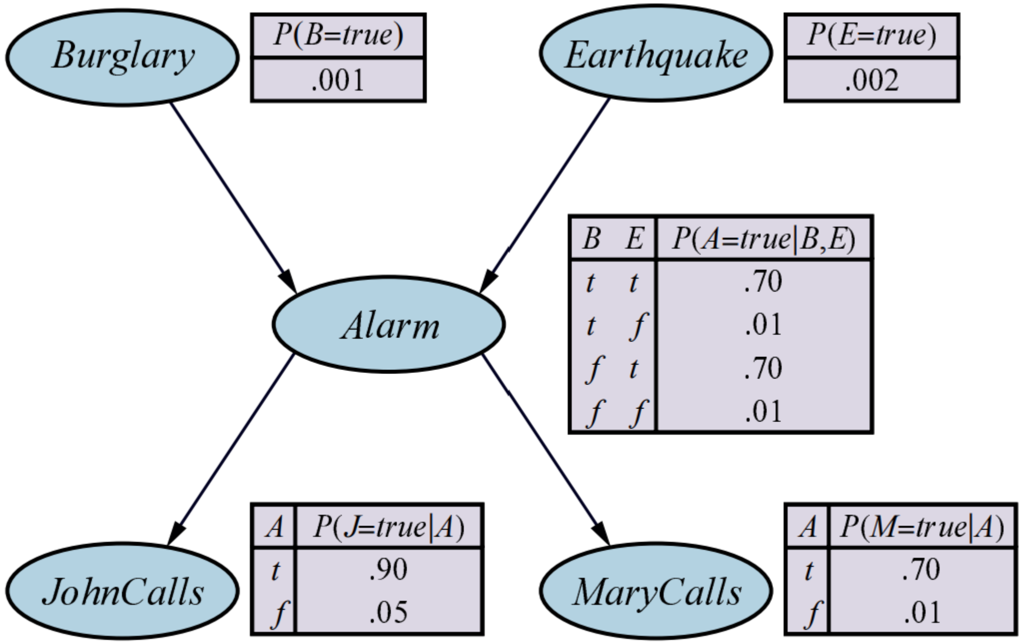


Figure 13.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

http://aima.cs.berkeley.edu/

PGM: parsimonious representation of distributions

- Suppose each $x_1, x_2, \dots, x_K$ take binary values
- Naively representing $P(x_1, x_2, \dots, x_K)$ requires 2^K entries
- With graphical model representation

$$P(x_1, x_2, \dots, x_K) = \prod_{k=1}^{K} P(x_k \mid \mathrm{pa}_k)$$

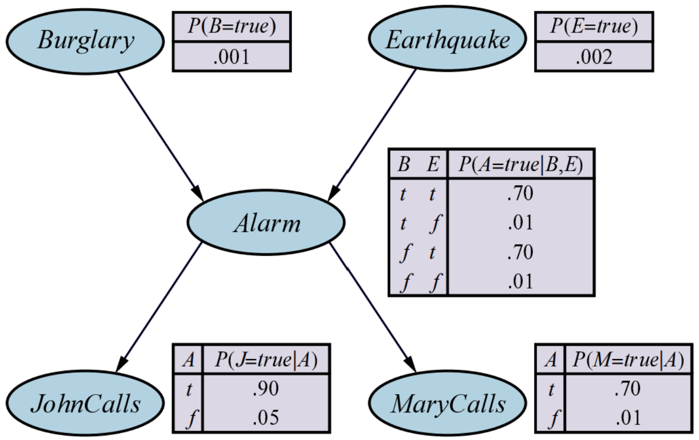


Figure 13.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

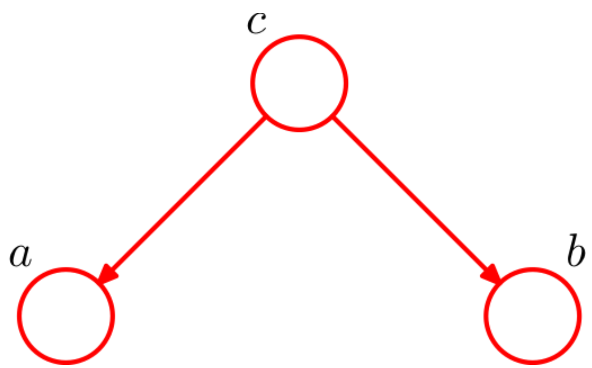
Each $P(x_k \mid \mathrm{pa}_k)$ takes $2^{|\mathrm{pa}_k|+1}$ entries

so total representation complexity $\leq \sum_k 2^{|\mathrm{pa}_k|+1} \leq 2^{O\left(\max_k |\mathrm{pa}_k|\right)}$

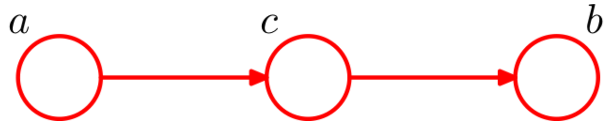
much smaller than 2^K if $\max_k |\mathrm{pa}_k| \ll K$ (we will see that this happens in many natural PGMs)

Three landmark examples

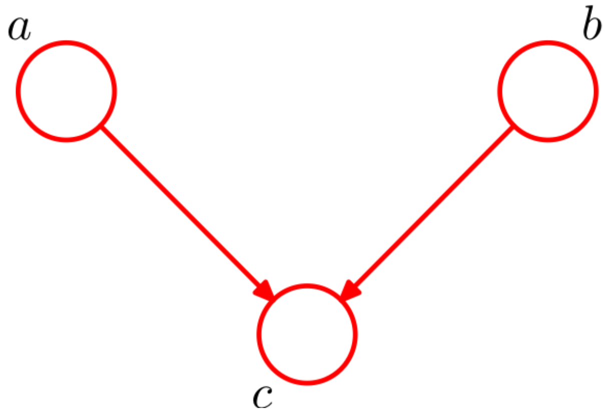
- tail-to-tail



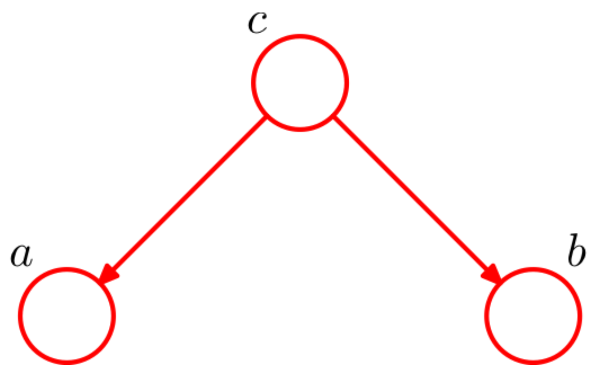
- Head-to-tail



- head-to-head



Ex 1: Tail-to-tail (common cause)

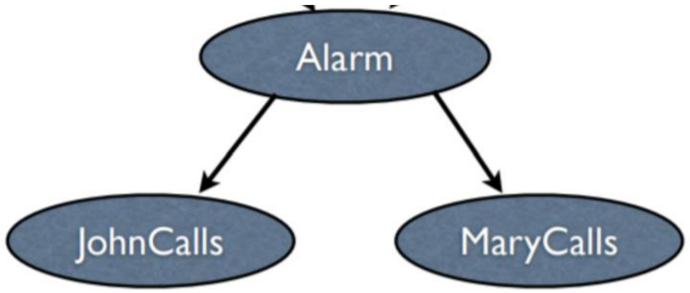


- $P(a, b, c) = P(c)P(\, a \mid c \,)P(b \mid c)$

- $P(a, b) = \sum_c P(c)P(\, a \mid c \,)P(b \mid c)$ and in general it does not factorize

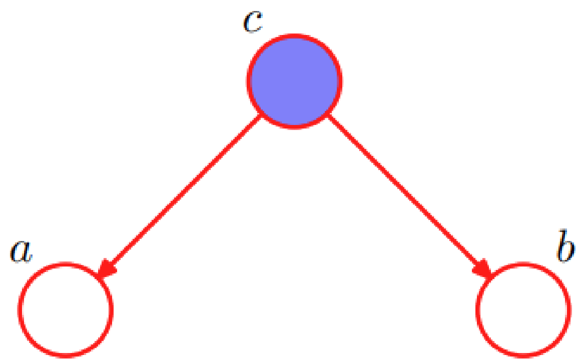
=> It is generally not true that $a \perp\!\!\!\perp b$

(e.g. John's calling is correlated with Mary's calling)



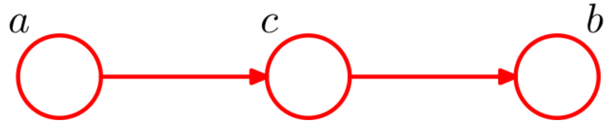
- However, $P(a, b \mid c) = \frac{P(a,b,c)}{P(c)} = P(\, a \mid c \,)P(b \mid c)$

=> $a \perp\!\!\!\perp b \mid c$



Ex 2: head-to-tail

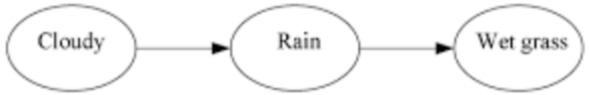
- $P(a, b, c) = P(a)P(c \mid a)P(b \mid c)$



- $P(a, b) = P(a) \sum_c P(c \mid a)P(b \mid c) = P(a) \cdot P(b \mid a)$

=> It is generally not true that $a \perp\!\!\!\perp b$

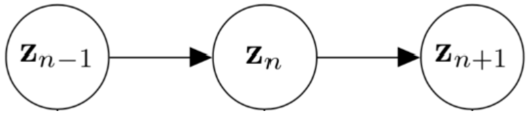
(e.g. “Cloudy” is correlated with “Wet grass”)



- However, $P(a, b \mid c) = \frac{P(a,b,c)}{P(c)} = \frac{P(a)P(c|a)P(b|c)}{P(c)} = P(a \mid c)P(b \mid c)$

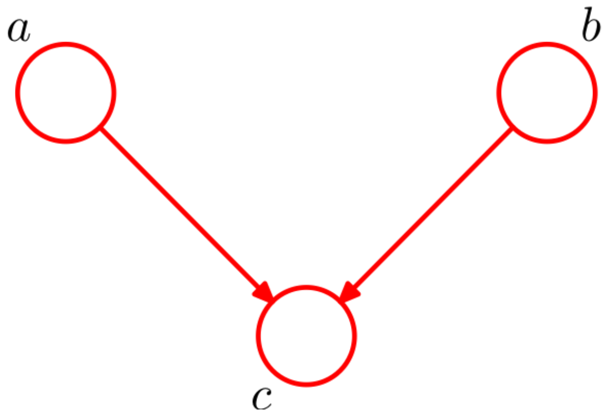
=> $a \perp\!\!\!\perp b \mid c$

- Another important example: Markov chain (for time series data)



https://www.cs.odu.edu/~zeil/cs795ML/webcourse/Slides/graphical/handout.pdf

Ex 3: head-to-head (common effect)



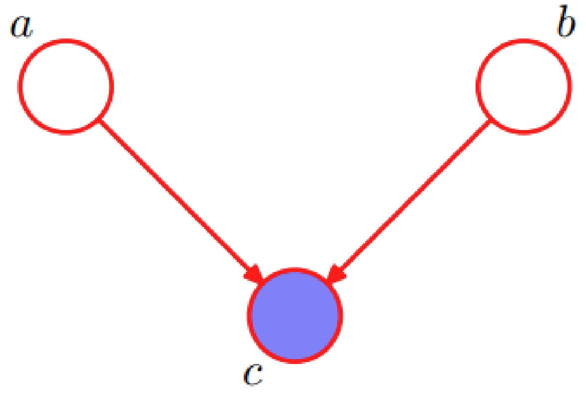
- $P(a,b,c) = P(a)P(b)P(c \mid a,b)$

- $P(a,b) = \sum_c P(a)P(b)P(c \mid a,b) = P(a)\,P(b)$

=> $a \perp\!\!\!\perp b$

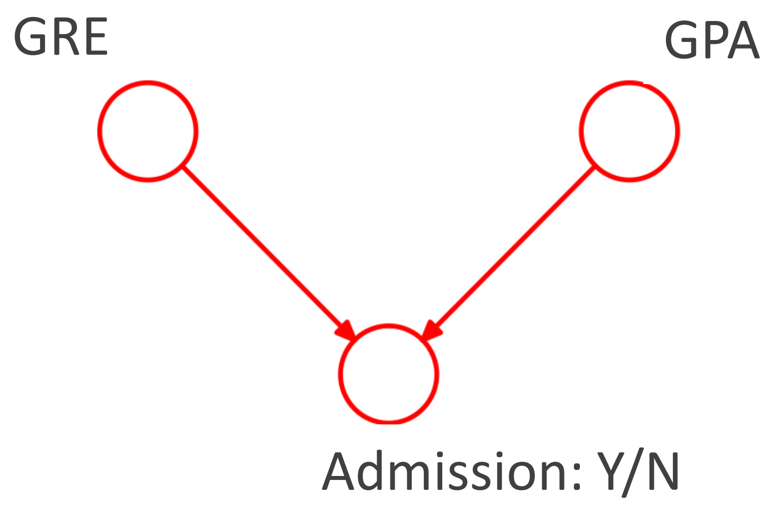
- However, $P(a,b \mid c) = \frac{P(a,b,c)}{P(c)} = \frac{P(a)P(b)P(c|a,b)}{P(c)}$ does not necessarily factorize

=> It is generally not true that $a \perp\!\!\!\perp b \mid c$

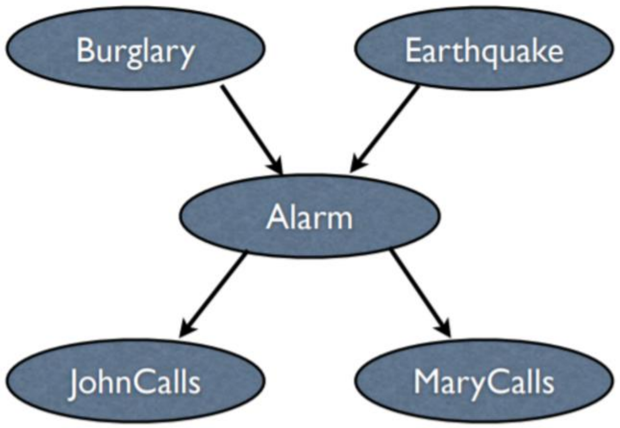


Ex 3: head-to-head (cont'd)

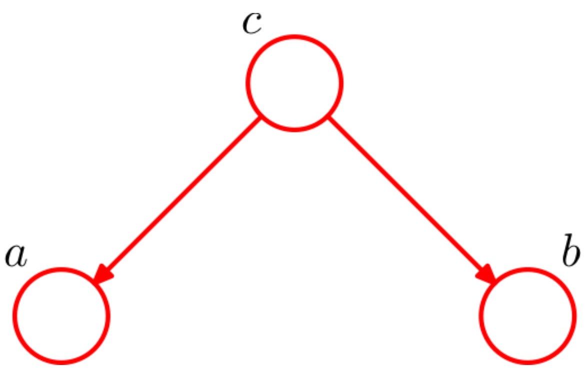
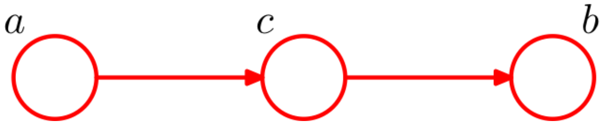
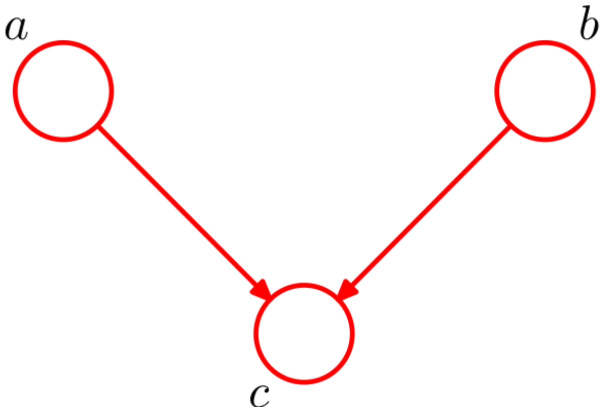
- If you pick an applicant randomly, the GRE and GPA is independent (according to our model)
- However, if you randomly pick an applicant who was accepted, then the low GRE may indicate that she had a high GPA.
  - Otherwise the student would have been rejected.
- This is called the **explain-away** phenomenon.



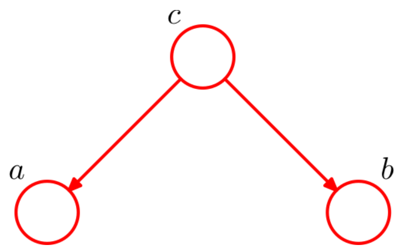
- Another example:
  - B and E are dependent, conditioned on A
  - It is also true that B and E are dependent, conditioned on *descendants of* A (e.g. J)

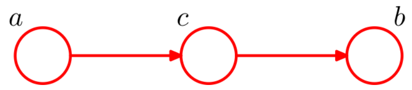


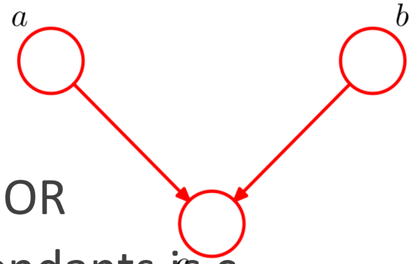
Summary

	$a \perp\!\!\!\perp b$?	$a \perp\!\!\!\perp b \mid c$?
• tail-to-tail 	No	Yes
• Head-to-tail 	No	Yes
• head-to-head 	Yes	No

D-separation







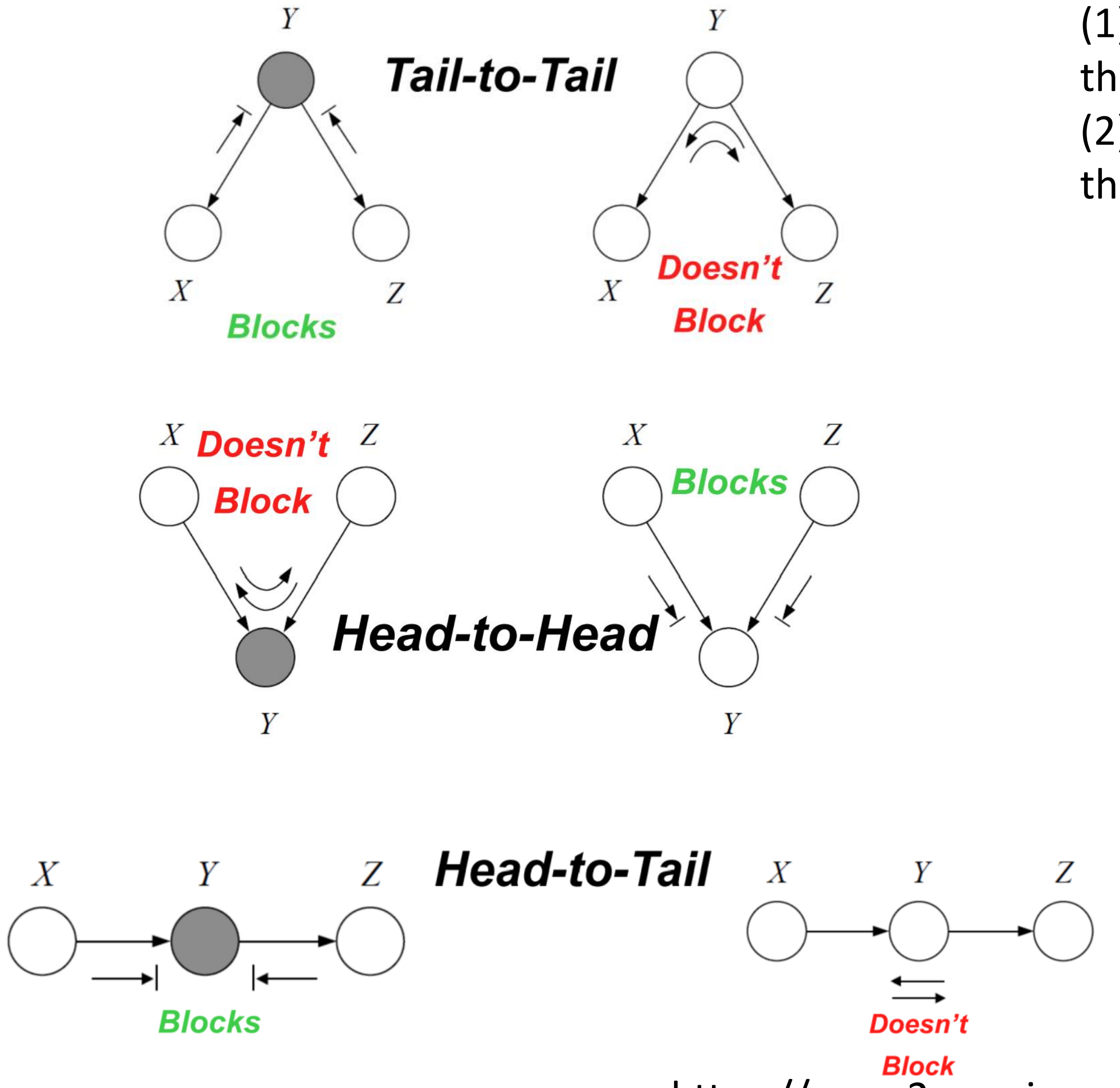
- Systematic Rules for determining conditional independence given a directed acyclic graph.
- Answer questions of the form: Is a ⫫ b | c true or false ?

- [Def] b is a **descendent** of a if there exists a <u>directed path</u> from a to b.
  - => a is a descendent of a by definition.
- [Def] An <u>undirected path</u> p from a to b is **blocked given** c if it includes a node:
  - (a) the arrows on p meet either head-to-tail or tail-to-tail at the node, and the node is c, OR
  - (b) the arrows meet head-to-head at the node, and neither the node nor any of its descendants is c

  "Conditioned on c being observed, information can flow from a to b through p"

- [Def] (D-separation)
  a is **d-separated from** b **given** c if every <u>undirected path</u> between a and b is blocked given c.
- [Thm] If a is **d-separated from** b **given** c, then $a \perp\!\!\!\perp b \mid c$.

Blockage: pictorial illustration



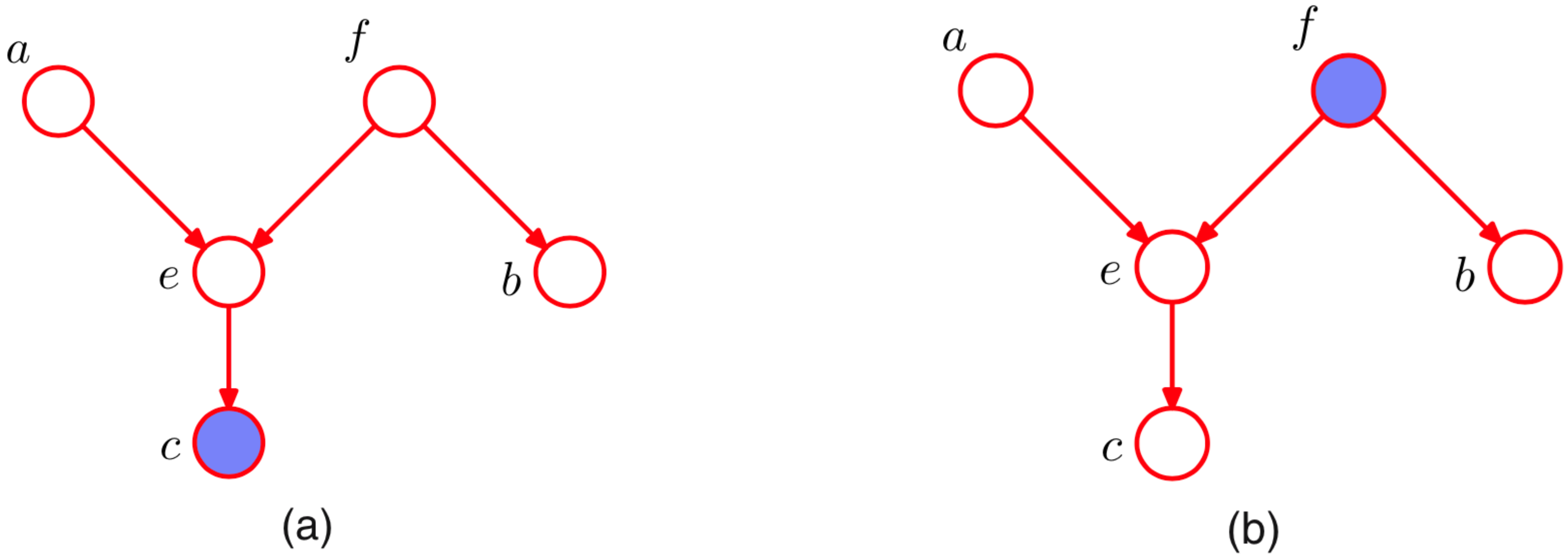
An undirected path p is *blocked* given c if it includes a node:
(1) the arrows on p meet either head-to-tail or tail-to-tail at the node, and the node is c, or
(2) the arrows meet head-to-head at the node, and neither the node nor any of its descendant is c

https://www2.cs.arizona.edu/~pachecoj/courses/csc535_fall20/lectures/pgms.pdf

D-separation examples

An undirected path p is *blocked* given c if it includes a node:
(1) the arrows on p meet either head-to-tail or tail-to-tail at the node, and the node is c, or
(2) the arrows meet head-to-head at the node, and neither the node nor any of its descendant is c

- Let path $p = a - e - f - b$
- In (a): p is not blocked given c => Not necessarily true that $a \perp\!\!\!\perp b \mid c$
- In (b): p is blocked given f => $a \perp\!\!\!\perp b \mid f$
- Is p blocked given Ø?



D-separation: general definition for node sets

- Q: Is A ⫫ B | C true or false ?
  - Each of A,B,C is a **set** of random variables

- [Def] An undirected path p from a to b is **blocked given** C if it includes a node:
  - (a) the arrows on p meet either head-to-tail or tail-to-tail at the node, and the node is in C
  - (b) the arrows meet head-to-head at the node, and neither the node nor any of its descendants is in C

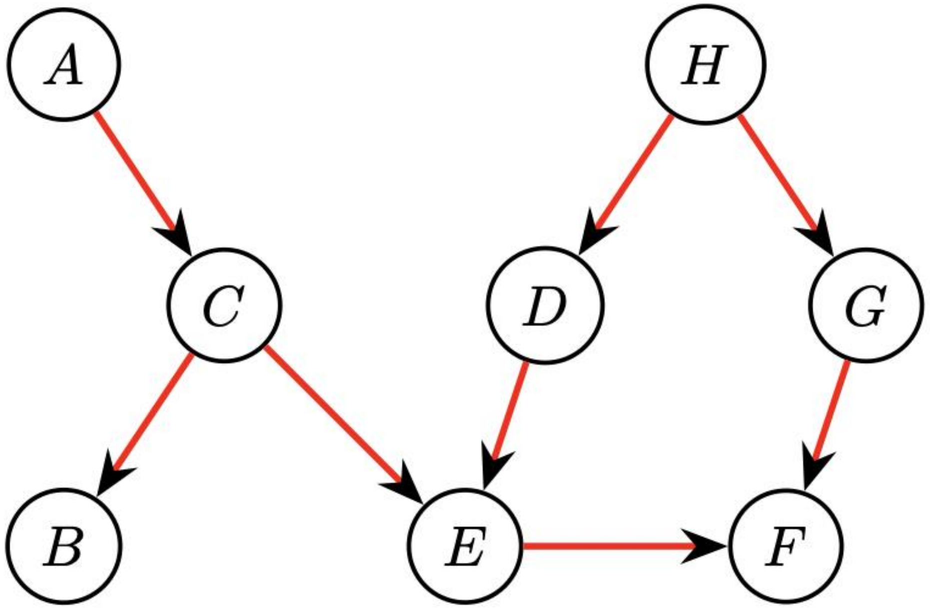
- [Def] (D-separation)
  A is **d-separated from** B **given** C if every undirected path between $a \in A$ and $b \in B$ is blocked given C.
- [Thm] If A is **d-separated from** B **given** C, then $A \perp\!\!\!\perp B \mid C$.

D-separation: an exercise

- Is $G \perp\!\!\!\perp A$ - equivalently, $G \perp\!\!\!\perp A \mid \emptyset$?
- Yes, all paths from A to G are blocked by E

- Is $E \perp\!\!\!\perp H \mid \{D, G\}$?
- Yes, E-D-H is blocked by D; E-F-G-H blocked by F (or G)

- Is $E \perp\!\!\!\perp H \mid \{C, D, F\}$?
- No, although E-D-H is blocked by D, E-F-G-H is not blocked



Next lecture (10/31)

- Markov models; Hidden Markov models (HMMs)

- Assigned reading: Prof. Jason Pacheco's PGM slides: https://www2.cs.arizona.edu/~pachecoj/courses/csc535_fall20/lectures/pgms.pdf

- Additional reading: Bishop, "Pattern Recognition and Machine Learning", Section 8.1-8.2