Probabilistic Graphical Models

Directed Graphical Modes

D-Separation

Probabilistic Graphical Models

Directed graphical models

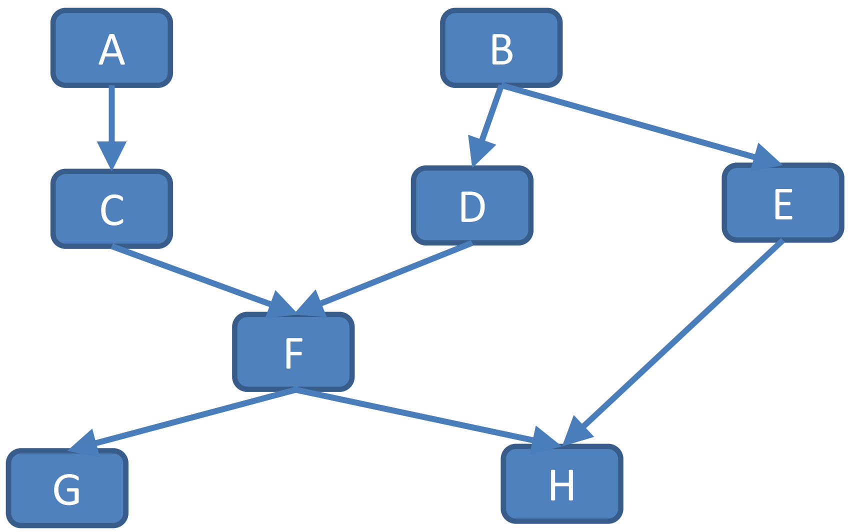
- Bayes Nets
- Conditional dependence

Undirected graphical models

- Markov random fields (MRFs)
- Factor graphs

Directed Graphical Models

A Directed Acyclic Graph



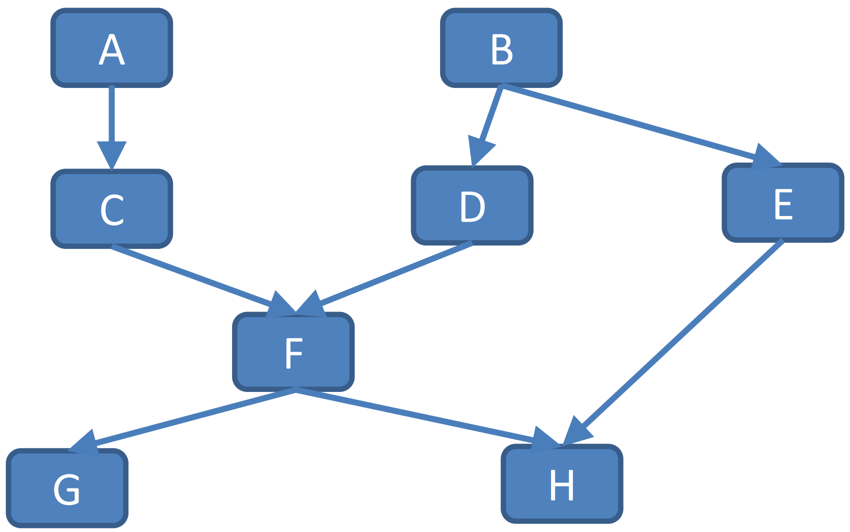
A joint Probability Distribution

$$P(A, B, C, D, E, F, G, H)$$

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

From Markov Condition to Factorization

A Directed Acyclic Graph



A joint Probability Distribution

$$P(A, B, C, D, E, F, G, H)$$

$$P(A, \ldots, H) = \prod_{V \in \{A, \ldots, H\}} P(V | Pa_G(V))$$

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

Factorization Theorem

- Let I(G) be the set of independencies entailed by the Markov Condition

- The Equivalence Theorem

  For a graph G,
  Let D_1 denote the family of all distributions that satisfy I(G),
  Let D_2 denote the family of all distributions that factor according to G (that are Markov to G),

  Then $D_1 \equiv D_2$.

$$P(X) = \prod_{i=1:d} P(X_i | Pa_G(X_i))$$

- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

Summary

Bayesian networks: Graph (DAG)+JPD

JPD factorizes according to the Markov Condition.

Markov Condition implies a factorization (proof) above

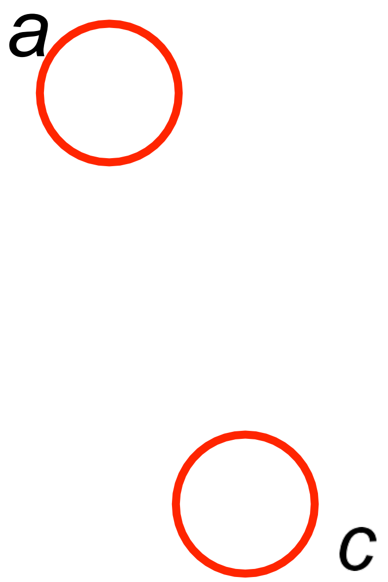
Factorization also implies the Markov condition

Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply like a certain architecture (e.g. a layered graph)
  - …
- Graphs imply some conditional independencies. (wherever you got them)
  - What does this mean?
  - For every distribution that factorizes according to the graph

Independencies.

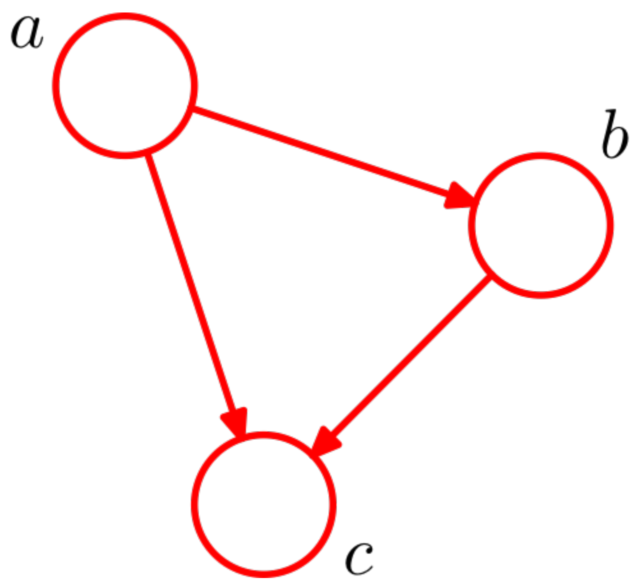
Are a and b independent ($a \perp b$ *)?*



$$\mathbf{p(a,b,c) = p(a)p(b)p(c)}$$

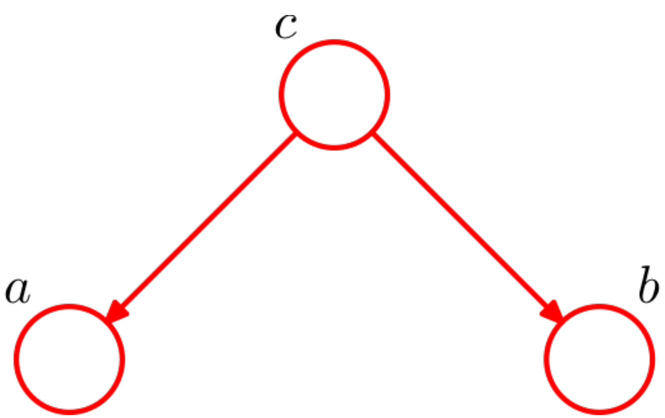
Independencies

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

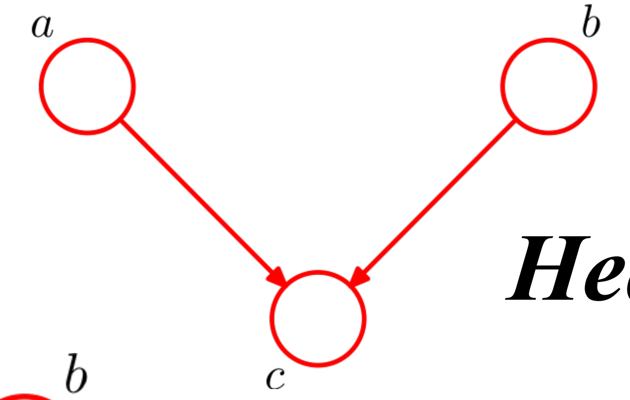


Note there are **no conditional independencies** (fully connected graph)

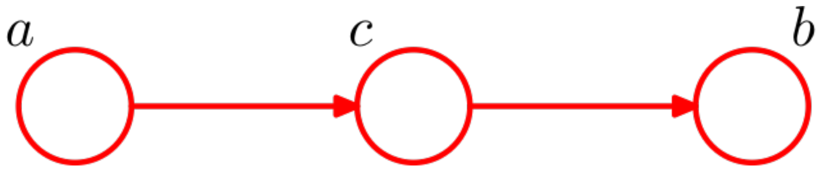
Three interesting cases



Tail-to-tail

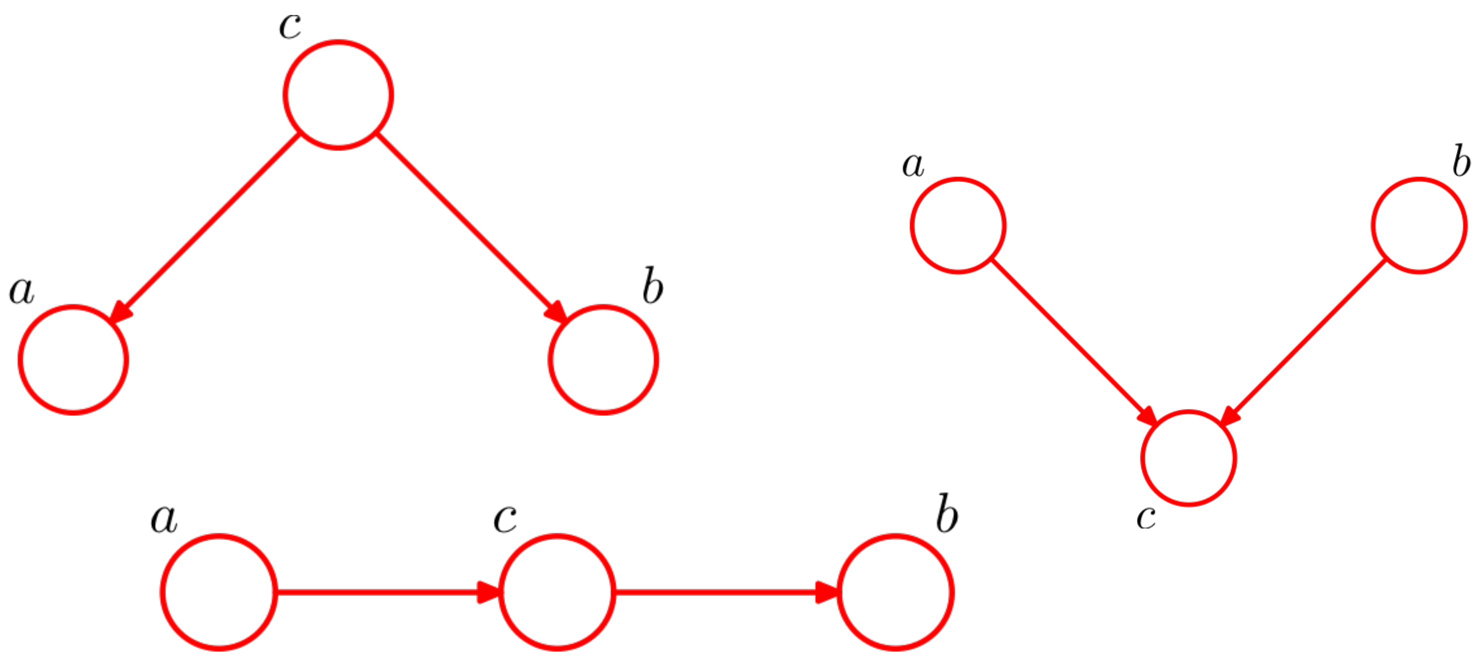


Head-to-head



Head-to-tail

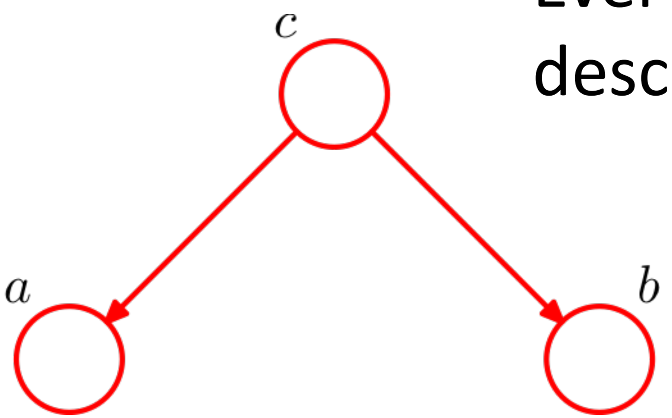
Three interesting cases



For each case, consider two questions:

1) Is $a \perp b$?
2) Is $a \perp b \mid c$? (i.e. c is observed)

Case one (tail-to-tail / fork)

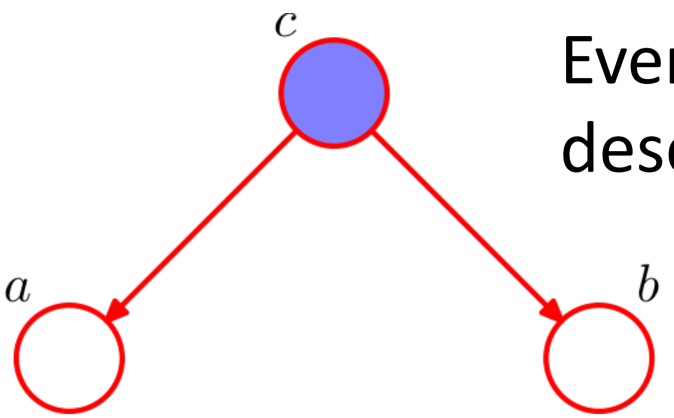


Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

$a \not\perp b$

This graph represents $P(a, b, c) = P(c)P(a|c)P(b|c)$

Case one (tail-to-tail)



Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

$$a \perp b \mid c$$

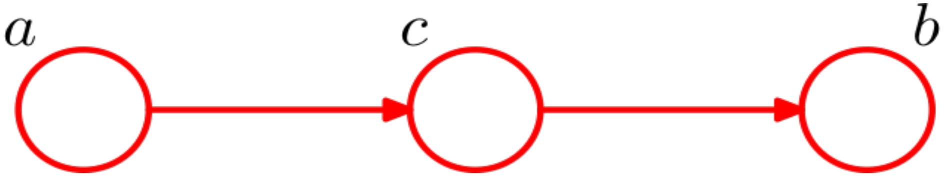
$$p(a,b,c) = p(c)p(a|c)p(b|c) \quad \text{(what the graph represents in general)}$$

$$p(a,b|c) = p(a|c)p(b|c) \quad \text{(with } c \text{ observed)}$$

This is the definition of $a \perp b|c$

Case two (head-to-tail, cascade)

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

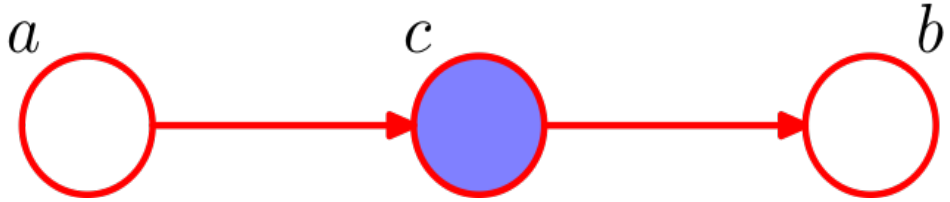


$$a \not\perp\!\!\!\perp b$$

This graph represents $P(a, b, c) = P(a)P(c|a)P(b|c)$

Case two (head-to-tail)

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

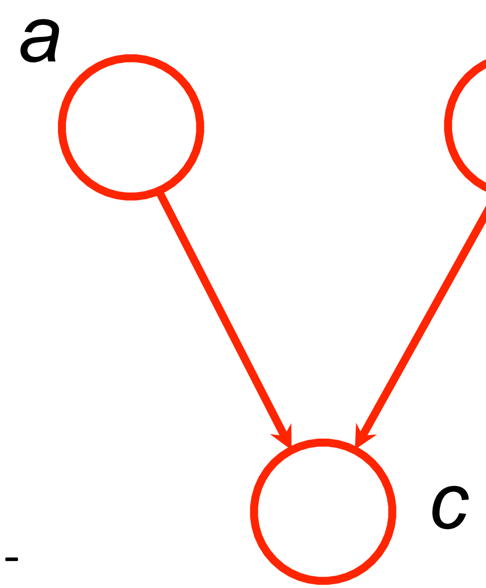


$$a \perp b \mid c$$

$$
\begin{aligned}
p(a,b \mid c) &= \frac{p(a,b,c)}{p(c)} && \text{(definition)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} && \text{(from graph)} \\
&= \frac{p(a)p(a|c)p(c)p(b|c)}{p(a)p(c)} && \text{(Bayes on } p(c|a)) \\
&= p(a|c)p(b|c)
\end{aligned}
$$

Case three (head-to-head)

Are a and b independent ($a \perp b$)?

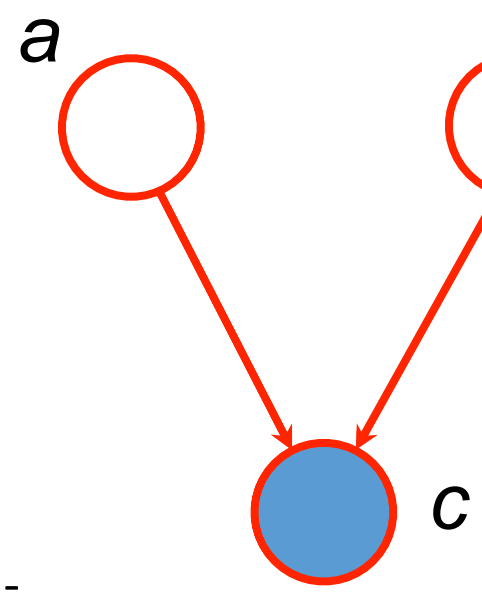


Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

$$p(a, b) = \sum_{c} p(a)p(b)p(c \mid a, b) = p(a)p(b)$$

Case three (head-to-head)

Are a and b conditionally independent $(a \perp b \mid c)$?

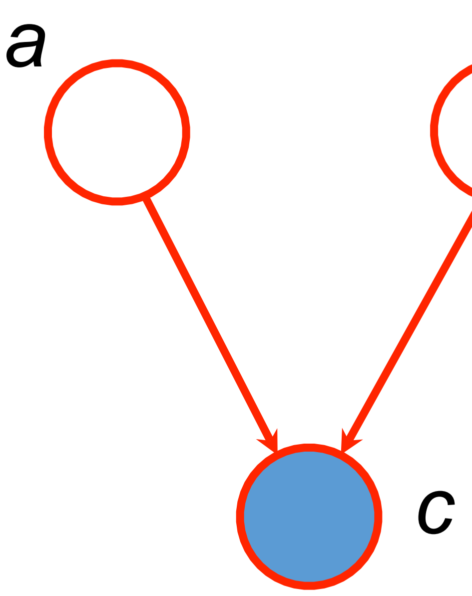


Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

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

Case three (head-to-head)

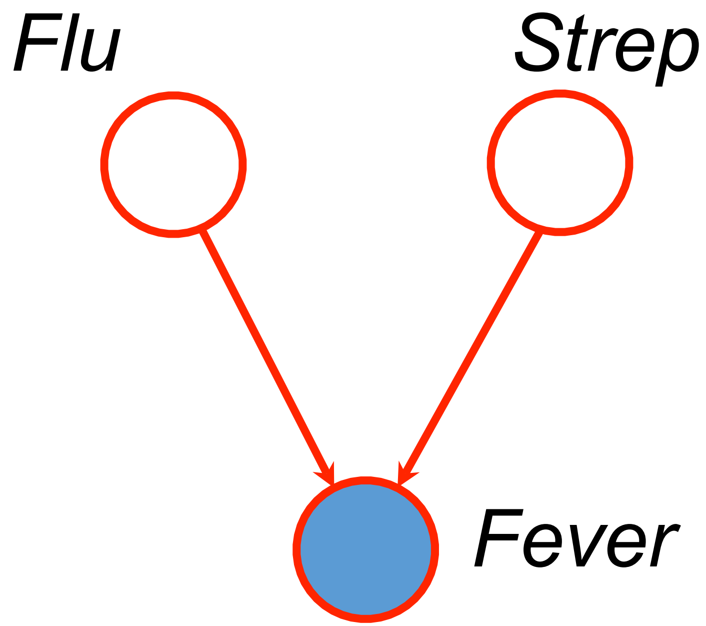
Are a and b conditionally independent ($a \perp b \mid c$)?

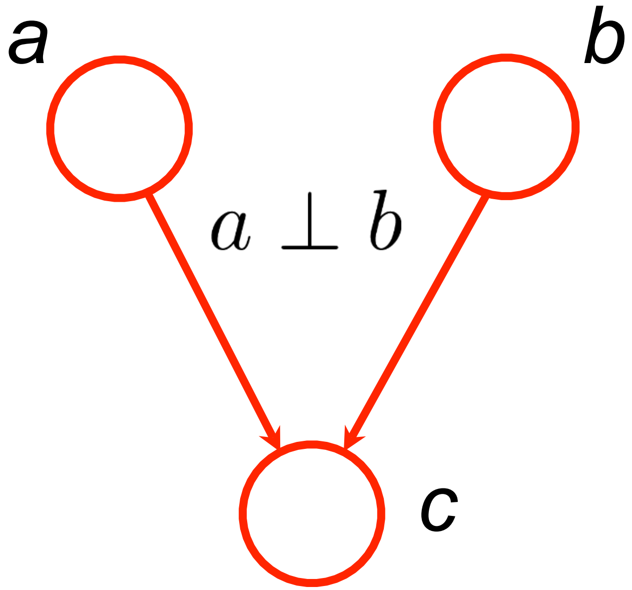


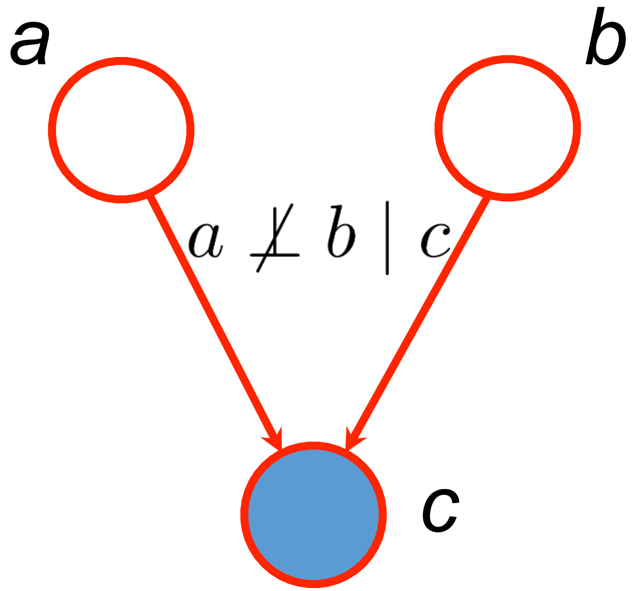
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a|c)p(b|c) \quad \text{(in general)}$$

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

Case three (head-to-head)



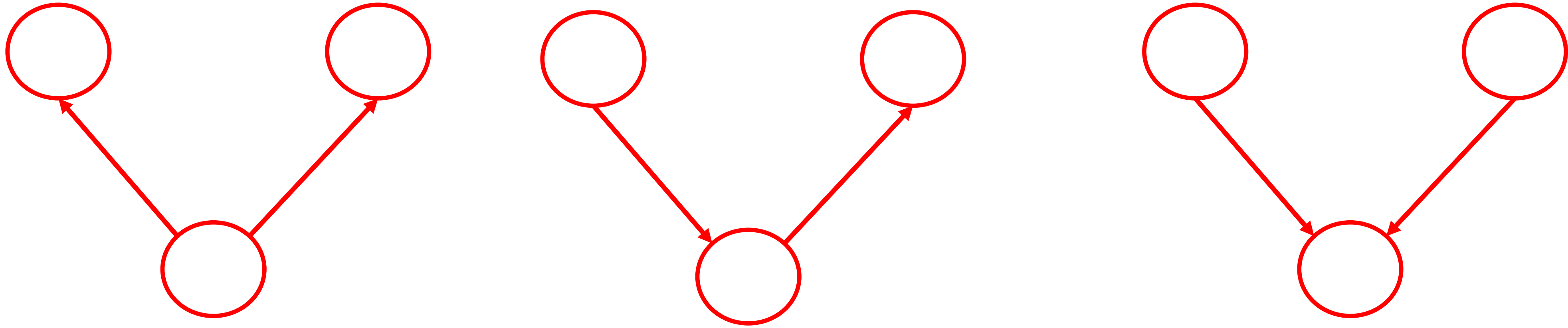




Phenomenon in Bayes networks known as **explaining away**

Three interesting cases

$a \perp b \mid c$ $\qquad$ $a \perp b \mid c$ $\qquad$ $a \perp b$

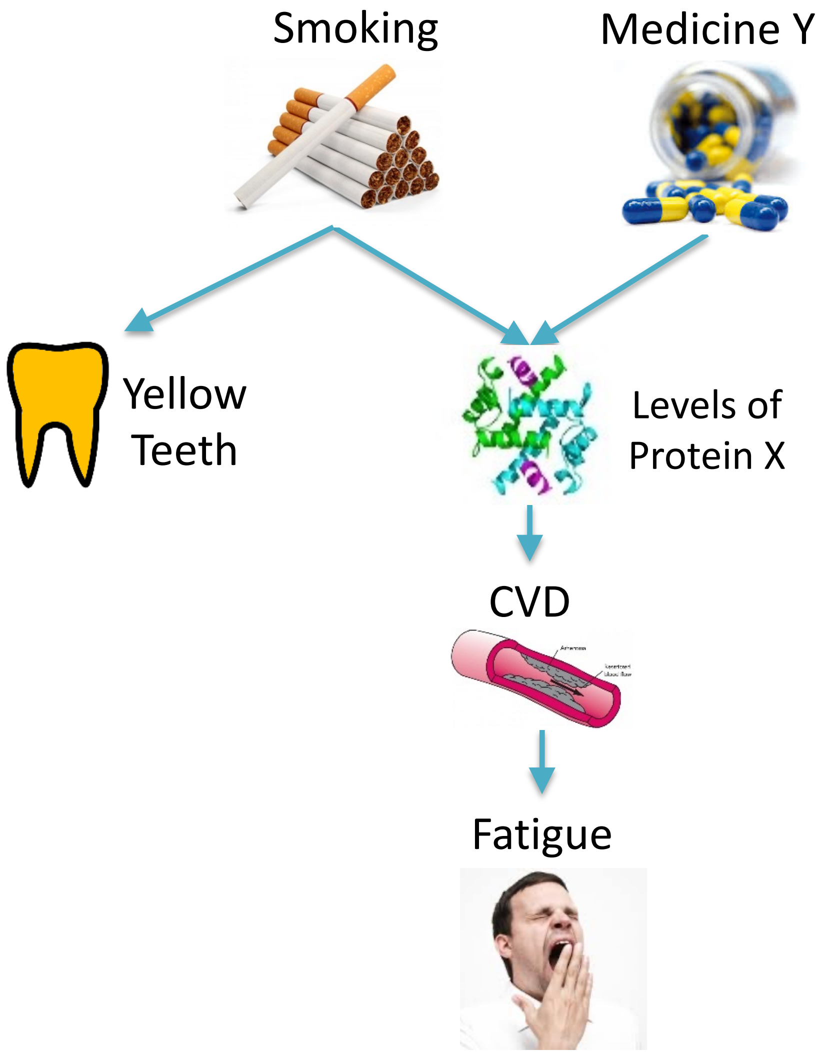


1) Is $a \perp b$?
2) Is $a \perp b \mid c$? (i.e. c is observed)

Quering the Markov Condition

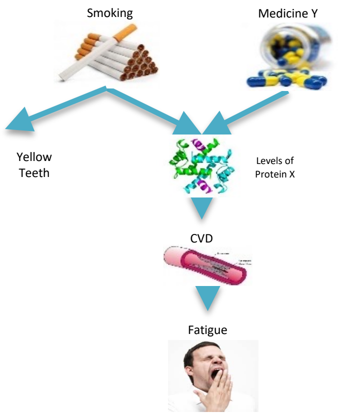
Is Fatigue independent of Smoking given Levels of Protein X?

P(Fatigue | Smoking,Levels of Protein X) ?=?
P(Fatigue | Levels of Protein X)



Do the Math

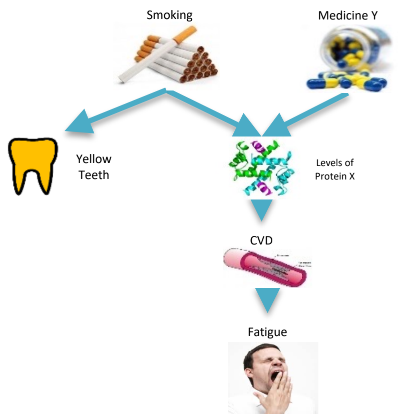
$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$



Quering the Markov Condition

$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$

$$\sum_{cvd} P(Fatigue|Smoking, Levels\ X, CVD)P(CVD|Smoking, Levels\ X)$$

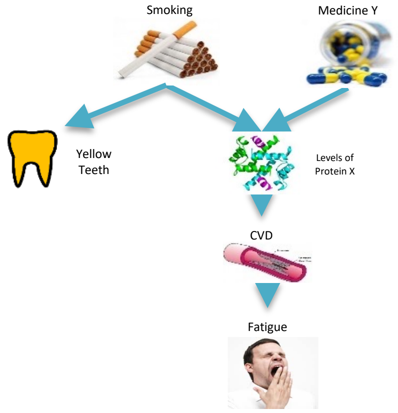


Law of total probability.

Quering the Markov Condition

$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$

$$\sum_{cvd} P(Fatigue|\cancel{Smoking}, Levels\ X, CVD)P(CVD|Smoking, Levels\ X)$$

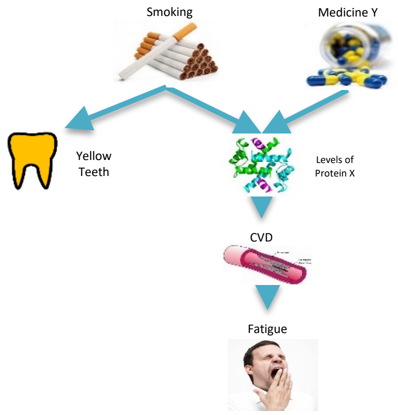


Markov Condition: P(F|S, X, CVD) = P(F|X, CVD).

Do the math

$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$

$$\sum_{cvd} P(Fatigue|\cancel{Smoking}, Levels\ X, CVD)P(CVD|\cancel{Smoking,} Levels\ X)$$



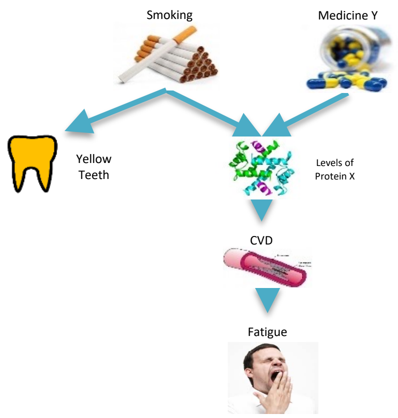
Markov condition: P(CVD|S, X) = P(CVD|X).

Do the math

$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$

$$\sum_{cvd} P(Fatigue|Levels\ X, CVD)P(CVD|Levels\ X) =$$

$$P(\text{Fatigue} \mid \text{Levels of Protein X})$$

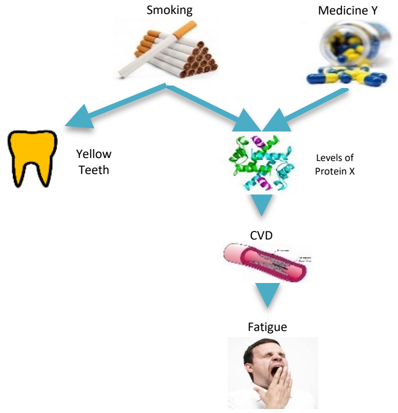


Law of total Probability

Do the math

$$P(\text{Fatigue} \mid \text{Smoking, Levels of Protein X}) =$$

$$P(\text{Fatigue} \mid \text{Levels of Protein X})$$



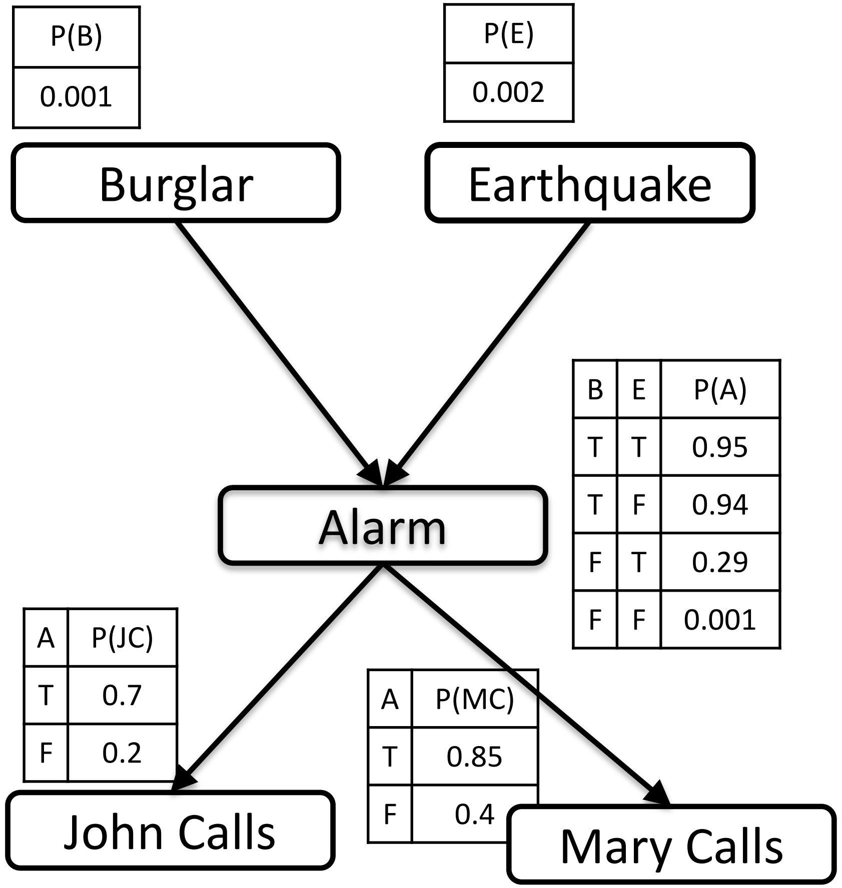
Markov condition dictates some independencies, entails some more.

The d-separation criterion

Algorithm to determine all independencies entailed by the Markov Condition.

Paths in the graph represent information flow (or lack thereof)

Example



P(B)
0.001

P(E)
0.002

B	E	P(A)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

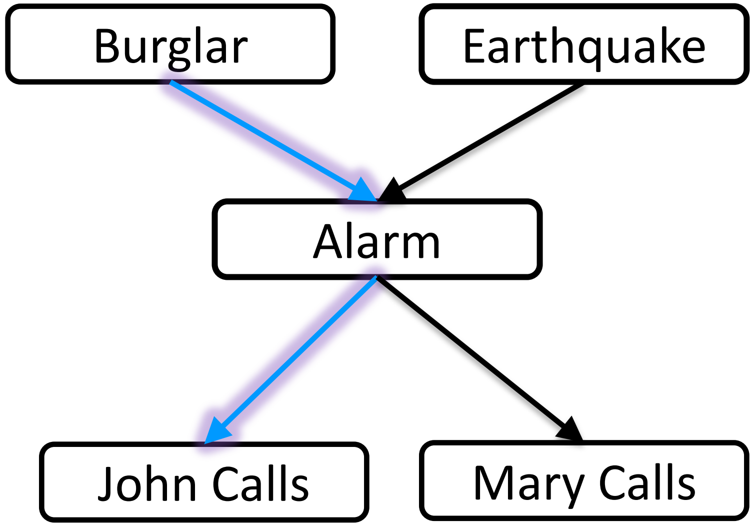
A	P(JC)
T	0.7
F	0.2

A	P(MC)
T	0.85
F	0.4

Example by J. Pearl

You have an installed alarm. Burglars set off the alarm. Earthquakes set of the alarm. When the alarm goes off, one of your neighbors (John or Mary) may call you.

Chains, head-to-tail

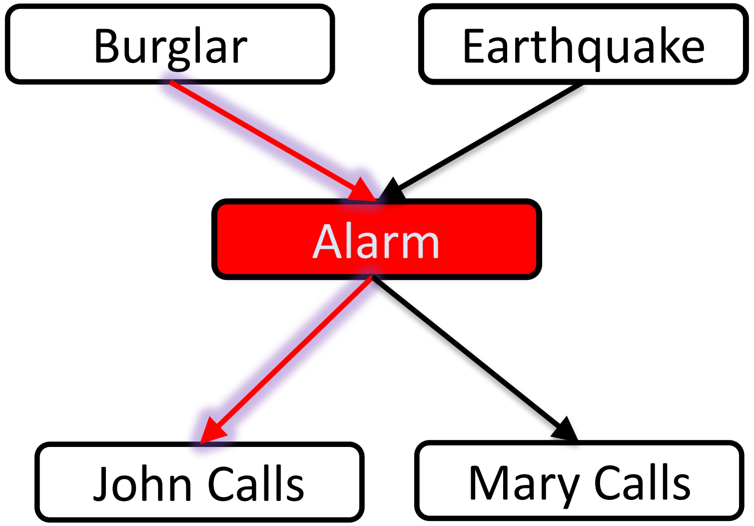


$$\text{Burglar} \not\perp\!\!\!\perp \text{John Calls} \mid \varnothing$$

A **directed path** is an **open path** (it allows information to flow).

When your house is being robbed, the probability of getting a phone call from John is higher than usual (due to the alarm going off).

Chains, head-to-tail

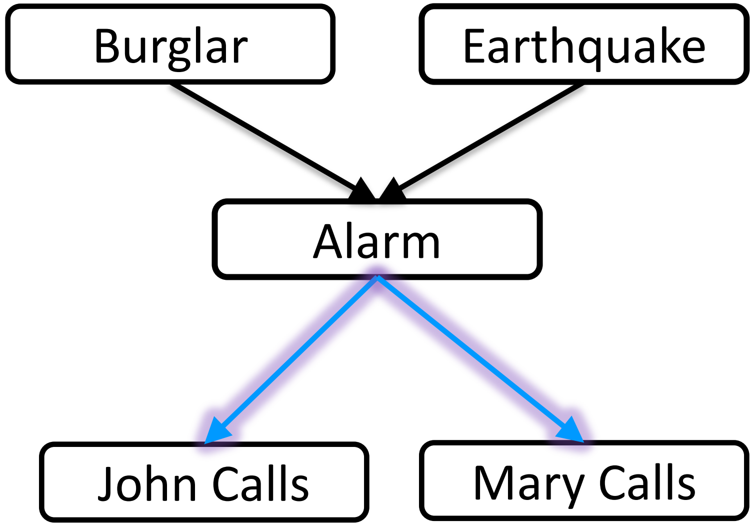


$\text{Burglar} \perp\!\!\!\perp \text{John Calls} \mid \text{Alarm}$

Conditioning on **intermediate** nodes **blocks** the path.

If you know the alarm has gone off, whether this was due or not to a burglary does not change your belief about the probability of getting a phone call from John.

Forks, tail-to-tail

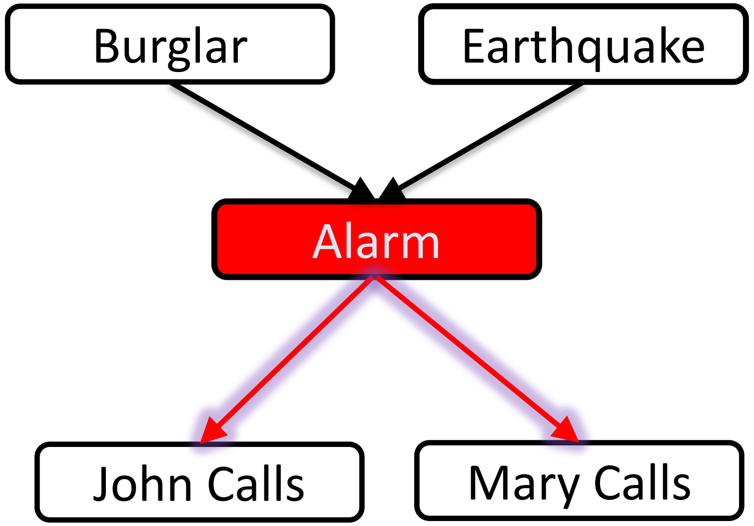


A path that goes through a **common parent** is an **open path** (it allows information to flow).

John Calls $\not\perp\!\!\!\perp$ Mary Calls $| \emptyset$

When John calls you, the probability that Mary will also call you is higher than usual (because John usually calls when the alarm has gone off).

Forks, tail-to-tail

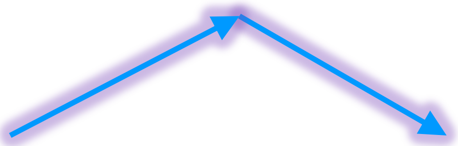


Conditioning on **common** parents **blocks** the path.

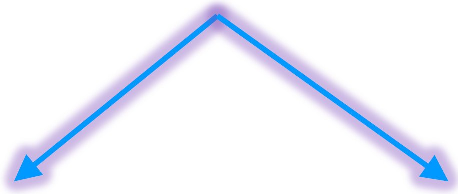
John Calls $\perp\!\!\!\perp$ Mary Calls | Alarm

If you know the alarm has gone off, whether John has called does not change your belief that Mary will call.

Chains and forks



Chains

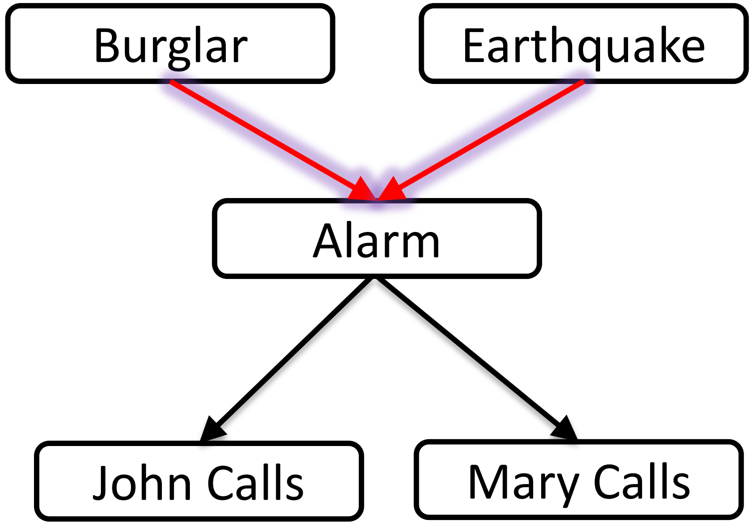


Forks

Chains and forks are open

Conditioning on the middle node blocks them

Colliders, head-to-head

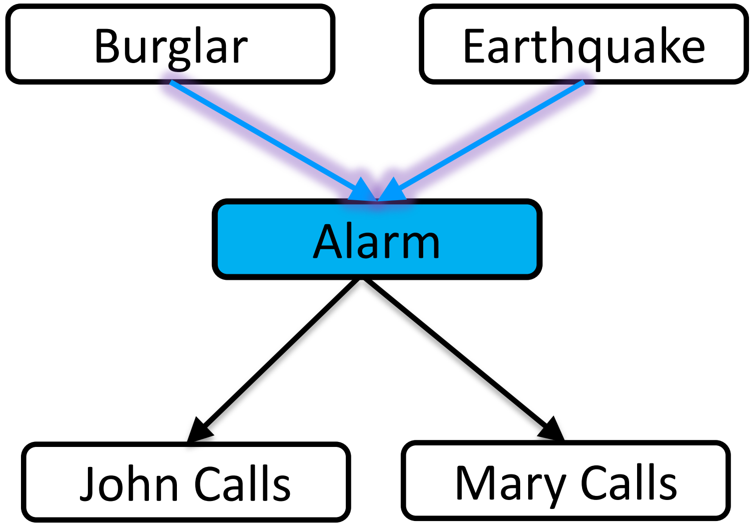


A path that goes through a **common child** is a **blocked** path (no information flows).

$\text{Burglar} \perp\!\!\!\perp \text{Earthquake} \mid \emptyset$

Burglaries and earthquakes are not correlated.

Colliders, head-to-head

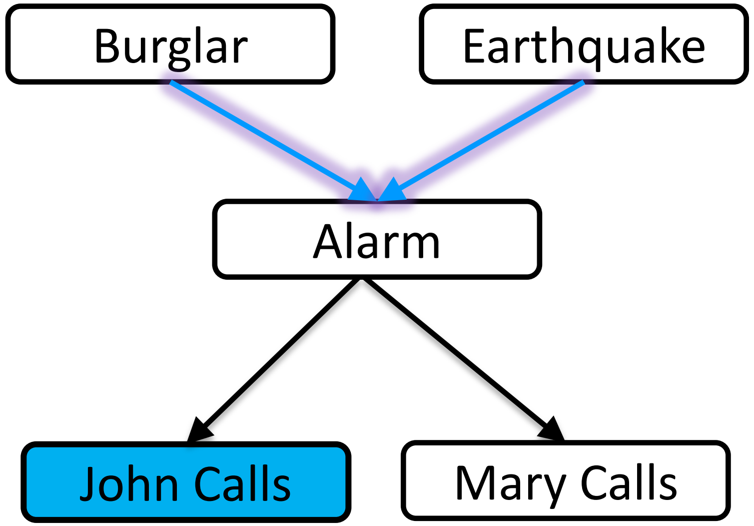


However, conditioning on **common children** **opens** the path (information flows through the path.)

$$\text{Burglar} \not\perp\!\!\!\perp \text{Earthquake} \mid \text{Alarm}$$

Your alarm has gone off. If there is also an earthquake, you think there is probably no burglary.

Colliders, head-to-head

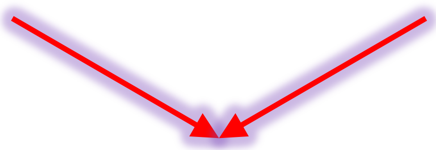


However, conditioning on common child OR descendant opens the path (information flows through the path).

$\text{Burglar} \not\perp\!\!\!\perp \text{Earthquake} \mid \text{Call}$

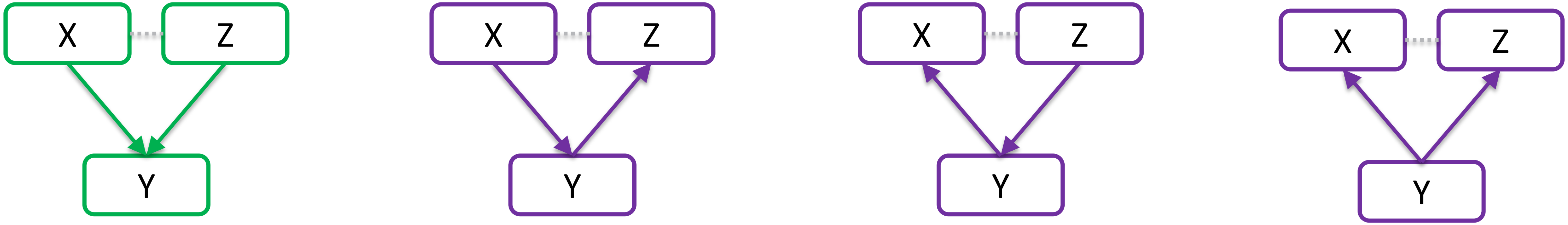
John is calling. If there is also an earthquake, you think there is probably no burglary.

Colliders, head-to-head



Colliders are blocked
Conditioning on (a descendant of) the middle node opens them

(non) colliders



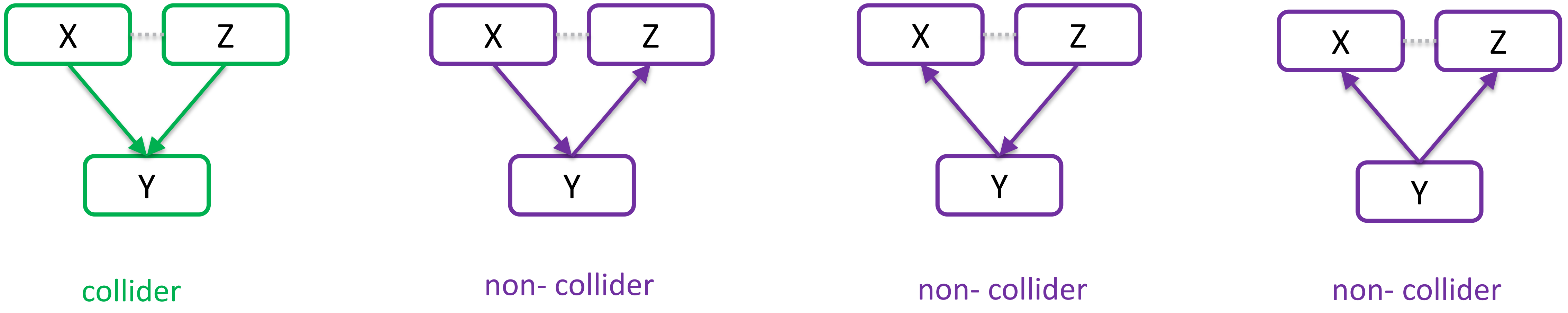
collider | non- collider | non- collider | non- collider

For a triple X-Y-Z:

If both edges are into Y, the triplet (and Y) is a collider.
Otherwise the triplet (and Y) is a non-collider.

The term is used to denote both the triplet and the middle node!

(non) colliders



A collider triplet is a blocked path

=> conditioning on Y (or a descendant of Y) opens the path.

A non-collider triplet is an open path

=> conditioning on Y blocks the path.

The d-separation criterion

Open (d-connecting) paths :
A path is d-connecting given **Z** iff
every collider on the path is in **Z** or has a descendant in **Z**
AND
every non-collider on the path is not in **Z.**

Otherwise, the path is blocked (d-separating).

The same path can be d-connecting given $\boldsymbol{Z_1}$, d-separating given $\boldsymbol{Z_2}$

The d-separation criterion

Algorithm to determine all independencies that are entailed by the MC.

Conditional independencies in the joint distribution can be decided based on the absence of open paths in the graph:

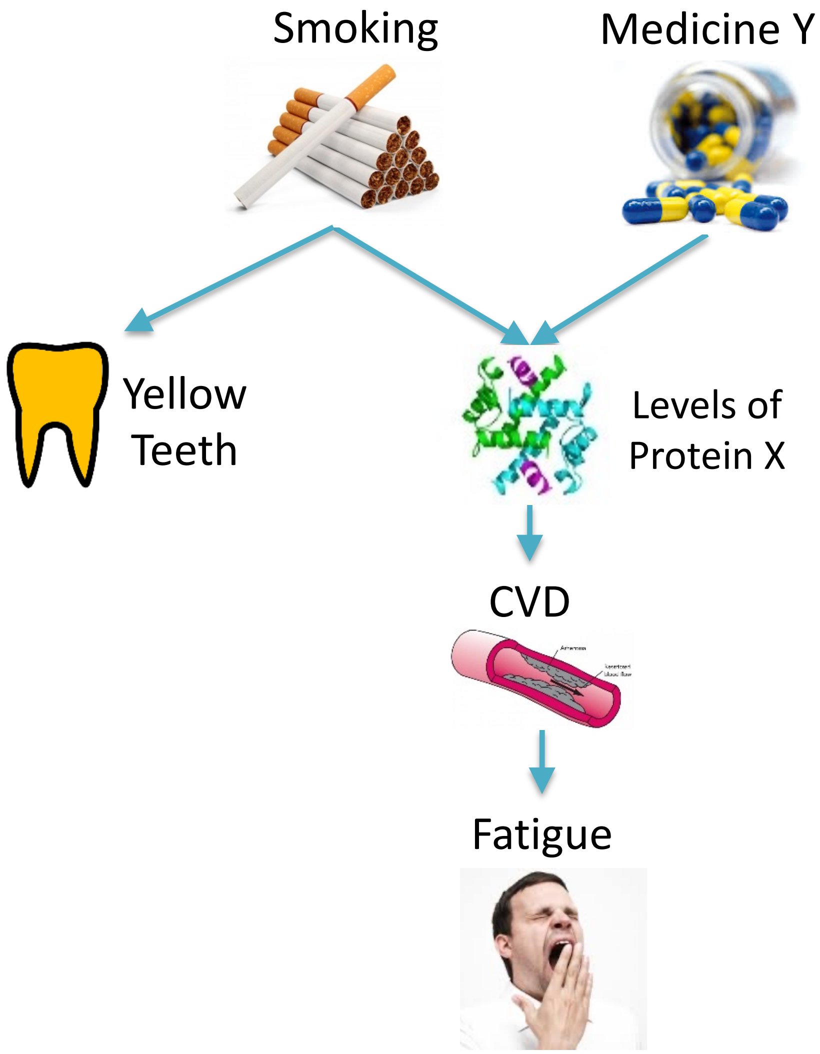
Open paths are called d-connecting paths (given a set of variables).
If no open path exists, the endpoints are d-separated (given the set of variables).
Otherwise, the endpoints are d-connected (given the set of variables)

Notation: $dsep(A, B\,|\boldsymbol{Z})$: A and B are d-separated given $\boldsymbol{Z}$.
$dcon(A, B\,|\boldsymbol{Z})$: A and B are d-connected given $\boldsymbol{Z}$.

The d-separation criterion

To find if $dsep(X, Y|\mathbf{Z})$ in the graph:

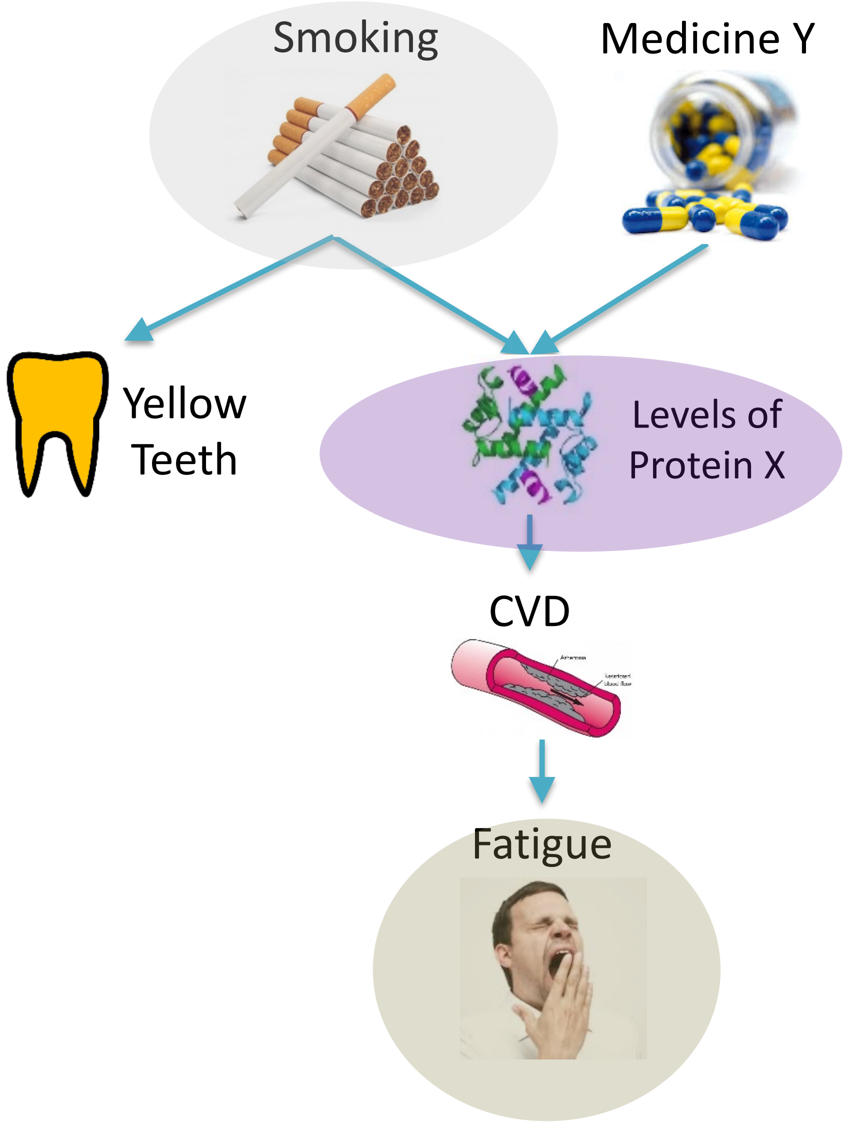
1. Find the paths from X to Y (ignoring orientations).
2. If there exists no open path given **Z**, then $dsep(X, Y|\mathbf{Z})$.



In BNs:
$dsep(X, Y|\mathbf{Z})$ in $G \Rightarrow X \perp\!\!\!\perp Y|\mathbf{Z}$ in P

The d-separation criterion

Is Fatigue independent of Smoking given Levels of Protein X?

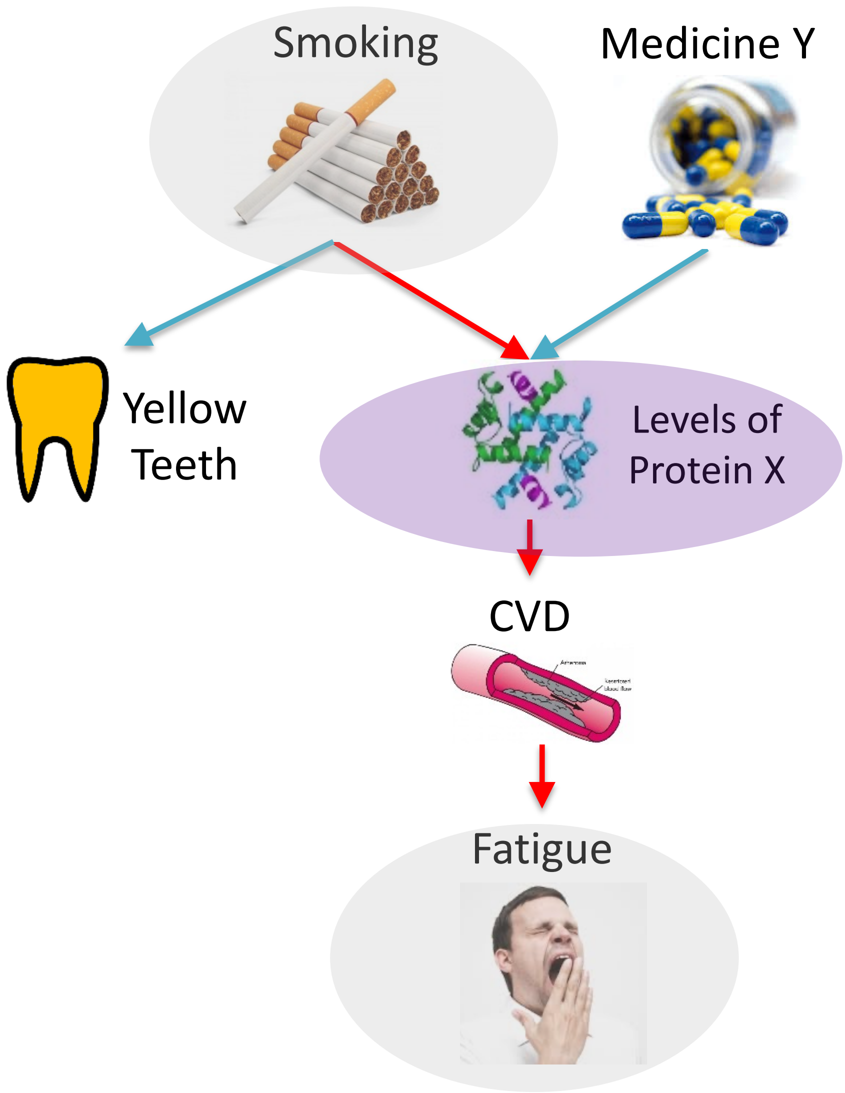


The d-separation criterion

Is Fatigue independent of Smoking given Levels of Protein X?

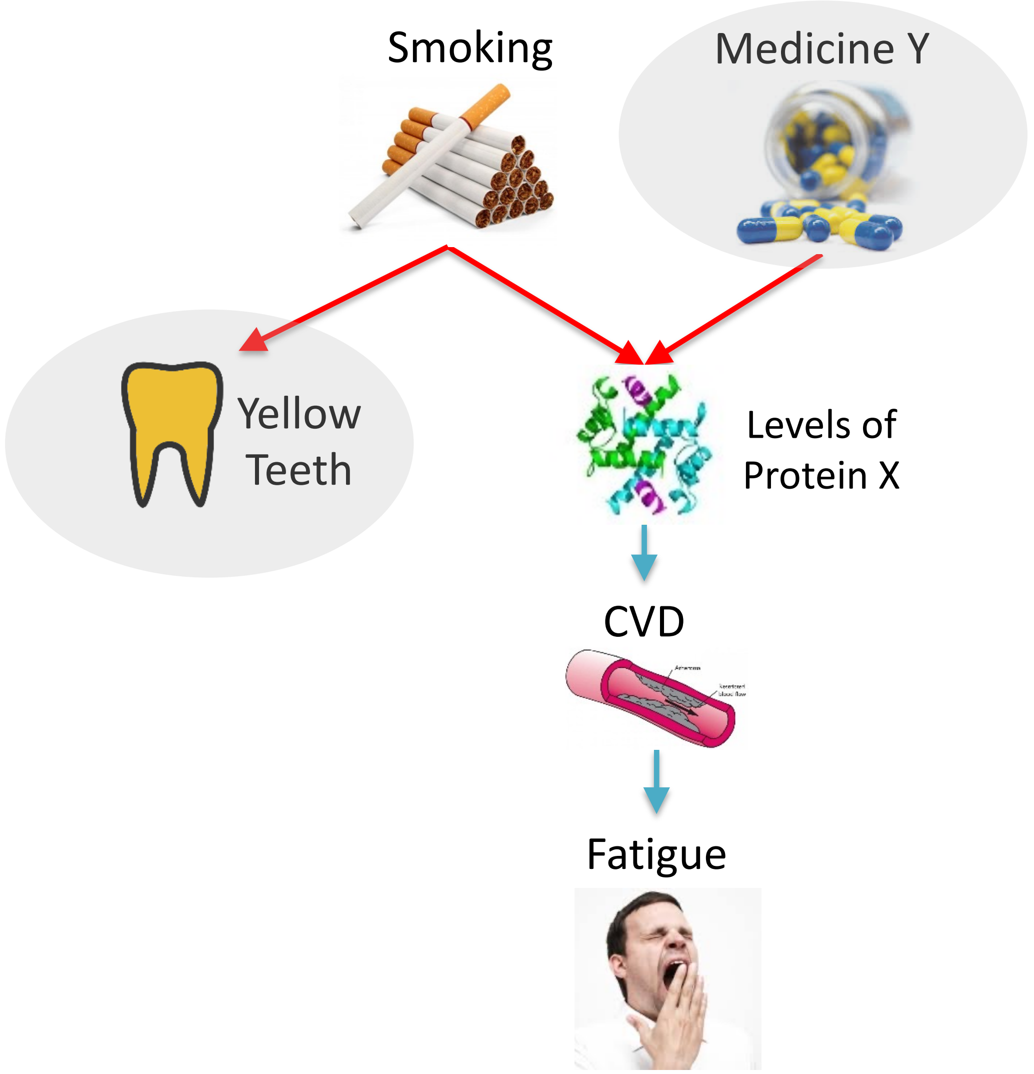
One possible path:
Smoking → Levels of Protein X → CVD → Fatigue

Path is blocked given Levels of Protein X



The d-separation criterion

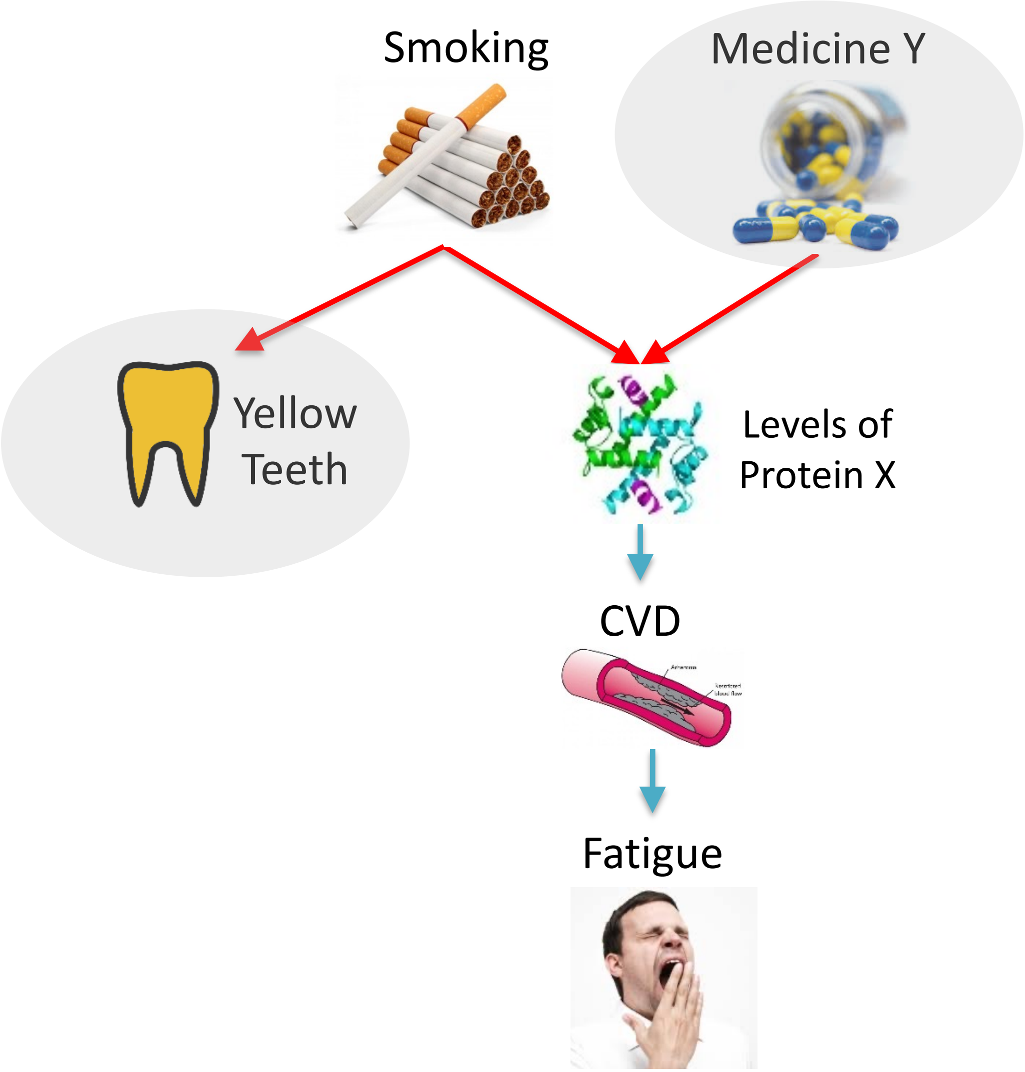
Are Yellow Teeth independent of Medicine Y?



Path is blocked given the empty set (goes through a collider)

The d-separation criterion

Are Yellow Teeth independent of Medicine Y?

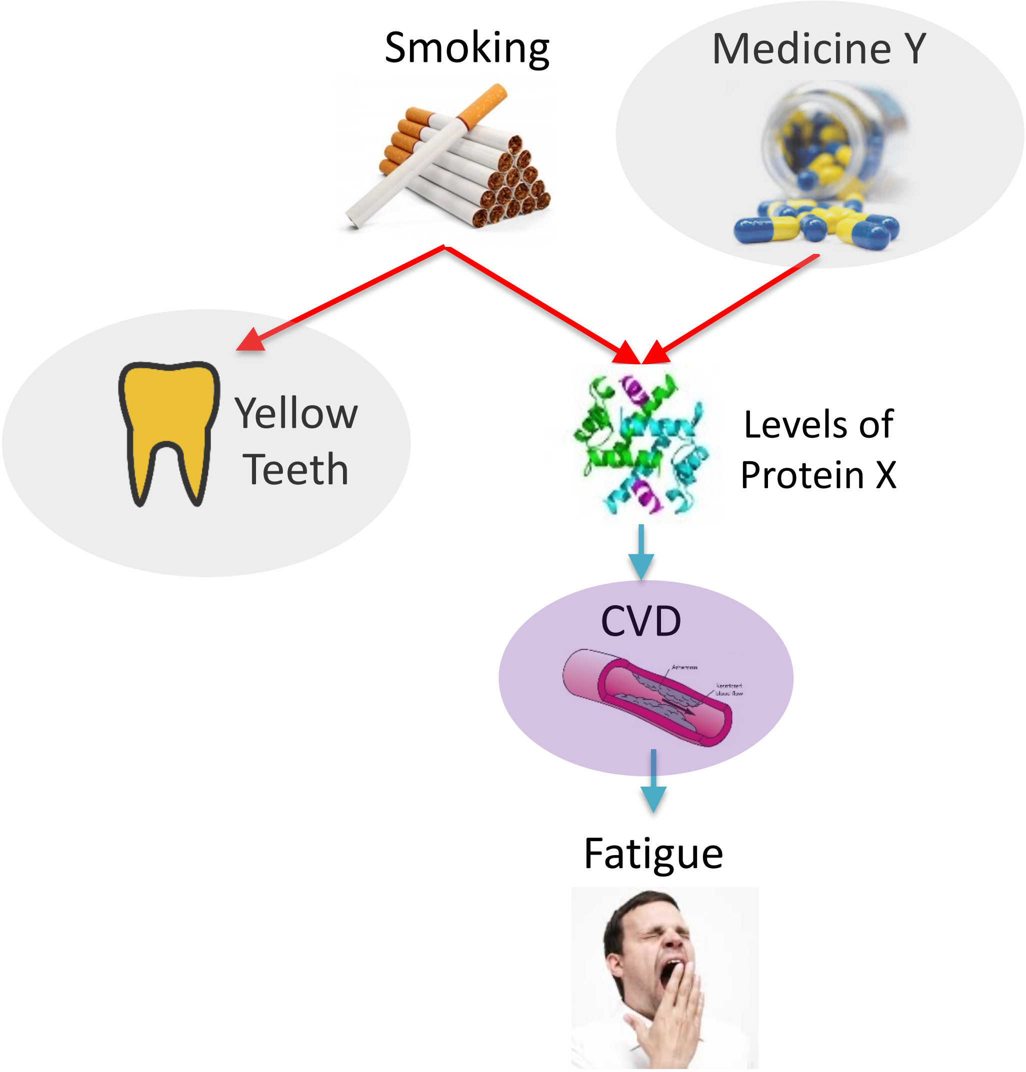


One possible path:
Yellow Teeth →Smoking → Levels of Protein X ← Medicine Y

Path is blocked given the empty set (goes through a collider)

The d-separation criterion

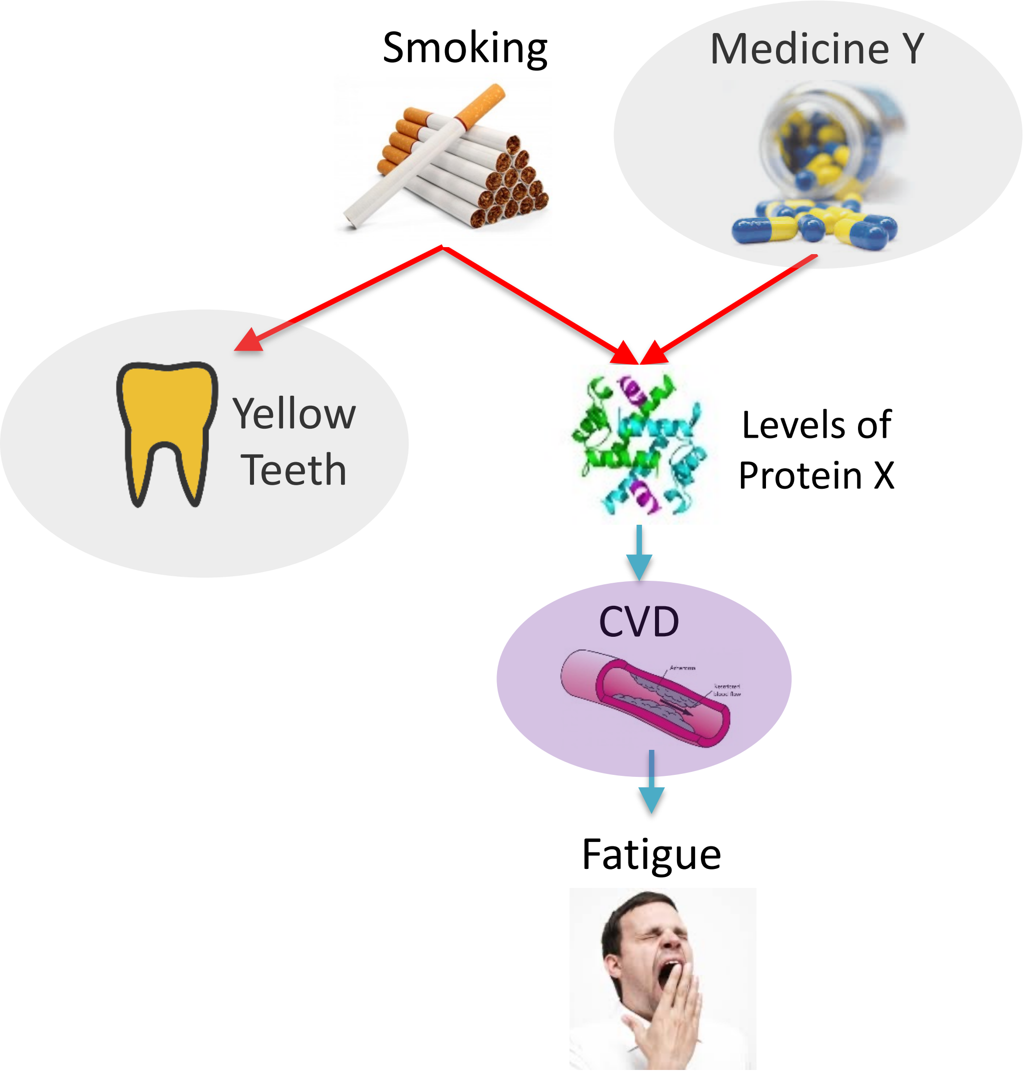
Are Yellow Teeth independent of Medicine Y given CVD?



Path is open given the CVD (conditioning on the descendant of a collider)

The d-separation criterion

Are Yellow Teeth independent of Medicine Y given CVD?



One possible path:
Yellow Teeth →Smoking → Levels of Protein X ← Medicine Y

Path is open given the CVD (conditioning on the descendant of a collider)

D-separation: overview

If you know the graph, you can use d-separation to answer any query for conditional independencies:

-Is X and Y independent given $\boldsymbol{Z}$?

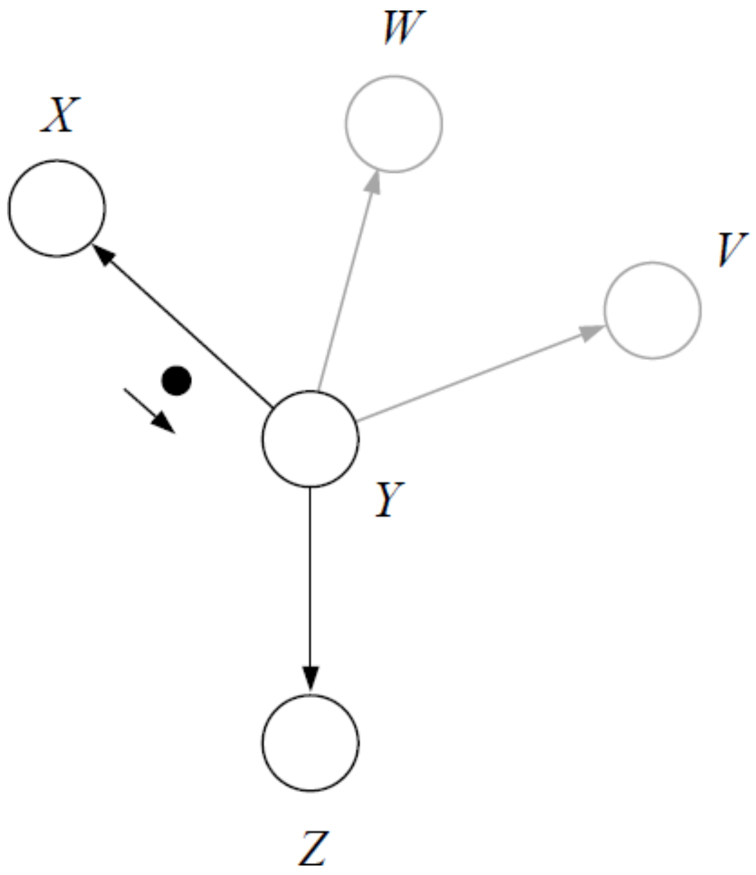
-Look at the graph

    if you cannot find an open path, they are independent

Bayes Ball Algorithm

[Source: Michael I Jordan]

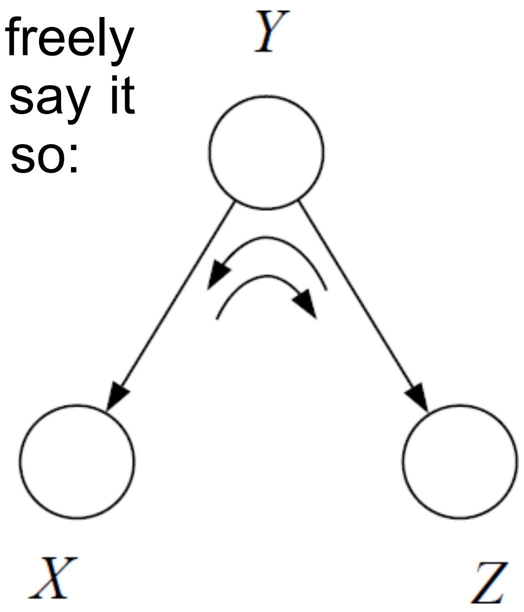
To test if $X \perp Z \mid Y$ imagine rolling a "ball" from X towards Z



The ball follows rules defined by the canonical 3-node subgraphs we've discussed

The ball passes freely from X-to-Z, we say it **does not block** so:

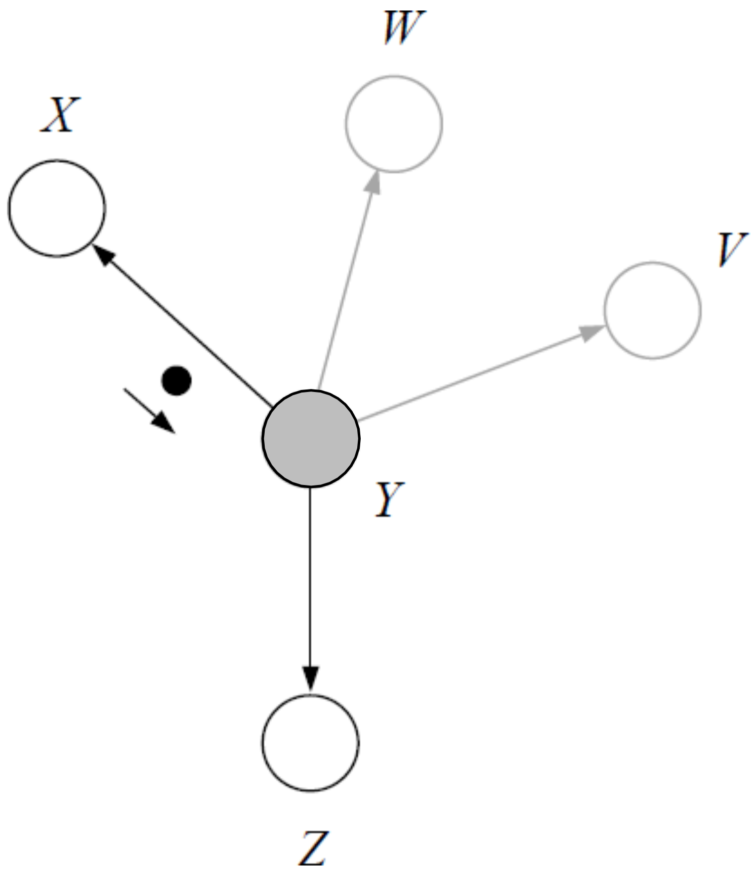
$$X \not\perp Z$$



Bayes Ball Algorithm

[Source: Michael I Jordan]

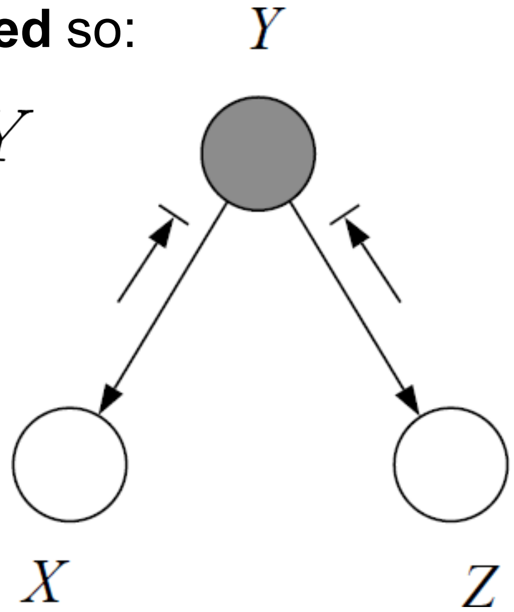
To test if $X \perp Z \mid Y$ imagine rolling a "ball" from X towards Z



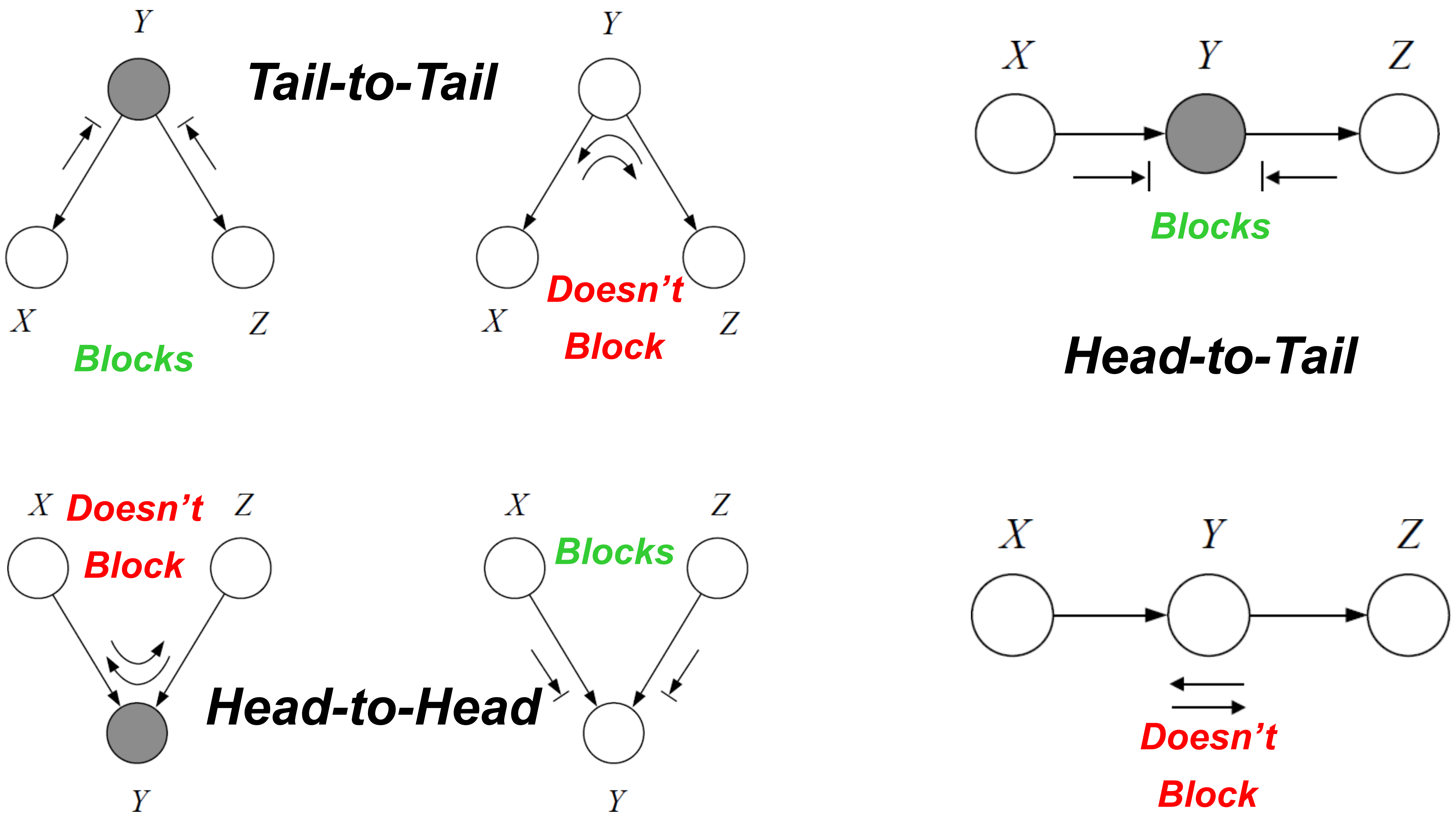
The ball follows rules defined by the canonical 3-node subgraphs we've discussed

The ball **is blocked** so:

$$X \perp Z \mid Y$$

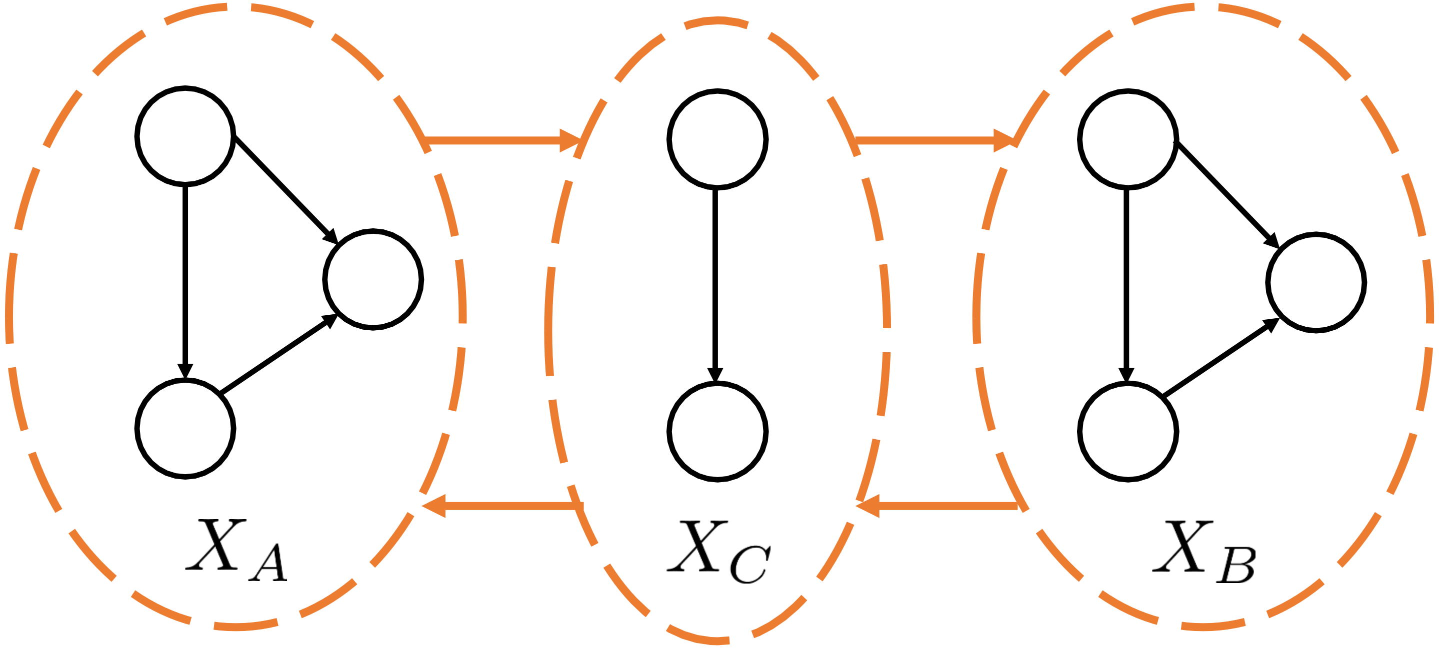


Bayes Ball Algorithm



Bayes Ball Algorithm

To test if $X_A \perp X_B \mid X_C$ roll ball from *every node in* X_A …



If *any* ball reaches *any node* in X_B then

$$X_A \not\perp X_B \mid X_C$$

Otherwise:

$$X_A \perp X_B \mid X_C$$

Tests for property of *directed separation* (d-separation): if X_C separates / blocks X_A from X_B then $X_A \perp X_B \mid X_C$

Summary

BN: DAG + Distribution

The distribution factorizes acording to the graph based on the Markov condition: Every variable is independent from its non-descendants (in the graph) based on its parents (in the graph)
$P(X_1, \dots, X_n) = \prod P(X_i | Pa(X_i))$

Markov Condition entails some independencies, not all of them are straightforward.

D-separation allows us to read the independencies from the graph.

I-map

- Defn : Let P be a distribution over $\boldsymbol{X}$. We define I(P) to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P (however how we set the parameter-values).

- Defn : Let K be *any graph object* associated with a set of independencies I(K). We say that K is an *I-map* for a set of independencies I, if $I(K) \subseteq I$

- We now say that G is an I-map for P if G is an I-map for I(P):

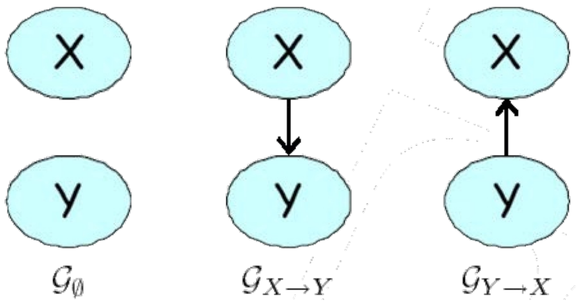
$$I(G) \subseteq I(P)$$

Facts about I-map

- For G to be an I-map of P, it is necessary that G does not mislead us regarding independencies in P:

  any independence that G asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in G

- Example:



P_1

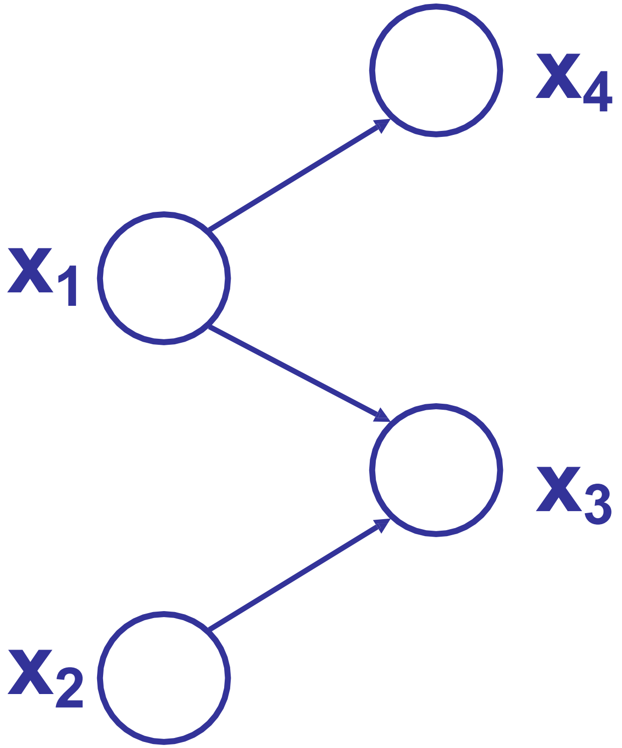
X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

P_2

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

Example:

- Complete the I(G) of this graph:



Perfect map

- Defn : Let P be a distribution over $\boldsymbol{X}$. We define I(P) to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P (however how we set the parameter-values).

- Defn : Let K be *any graph object* associated with a set of independencies I(K). We say that K is an *I-map* for a set of independencies I, if $I(K) \subseteq I$

- We now say that G is an I-map for P if G is a perfect map for I(P):

$$I(G) = I(P)$$