Probabilistic Graphical Models

Directed Graphical Modes

Pt 2

Probabilistic Graphical Models

Directed graphical models

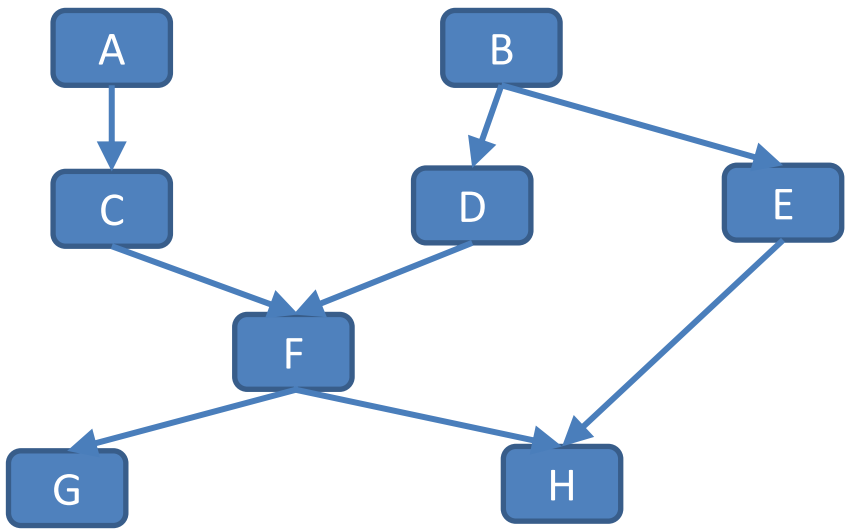
- Bayes Nets
- Conditional dependence

Undirected graphical models

- Markov random fields (MRFs)
- Factor graphs

Directed Graphical Models

A Directed Acyclic Graph



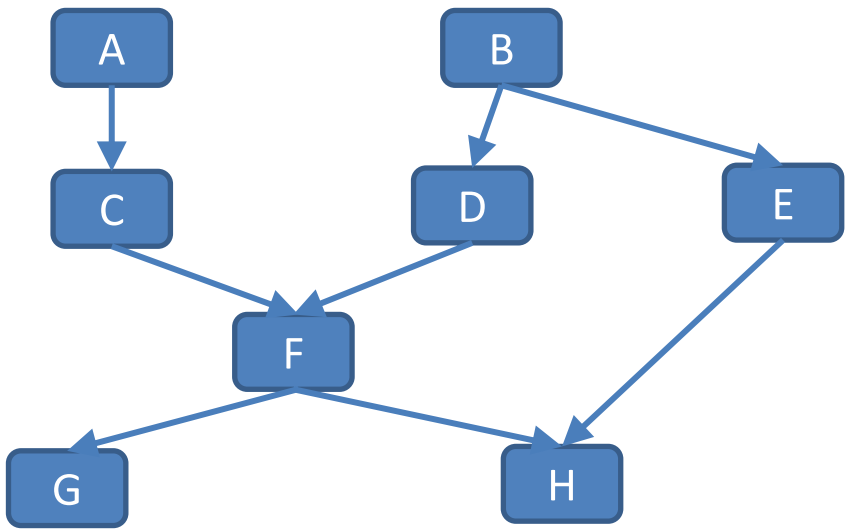
A joint Probability Distribution

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

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

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

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

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

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

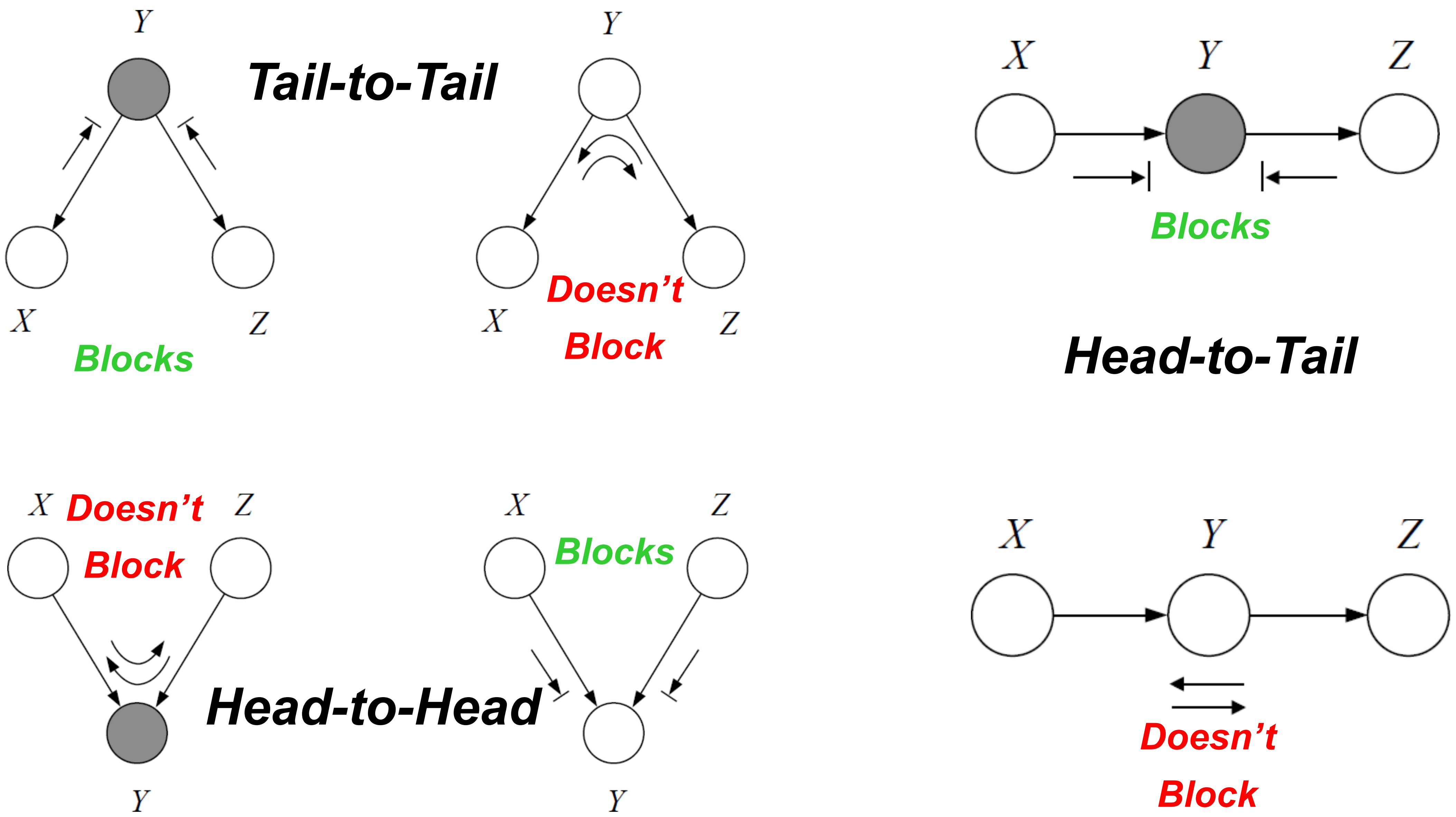
Algorithm to determine all independencies that are entailed by the MC.

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

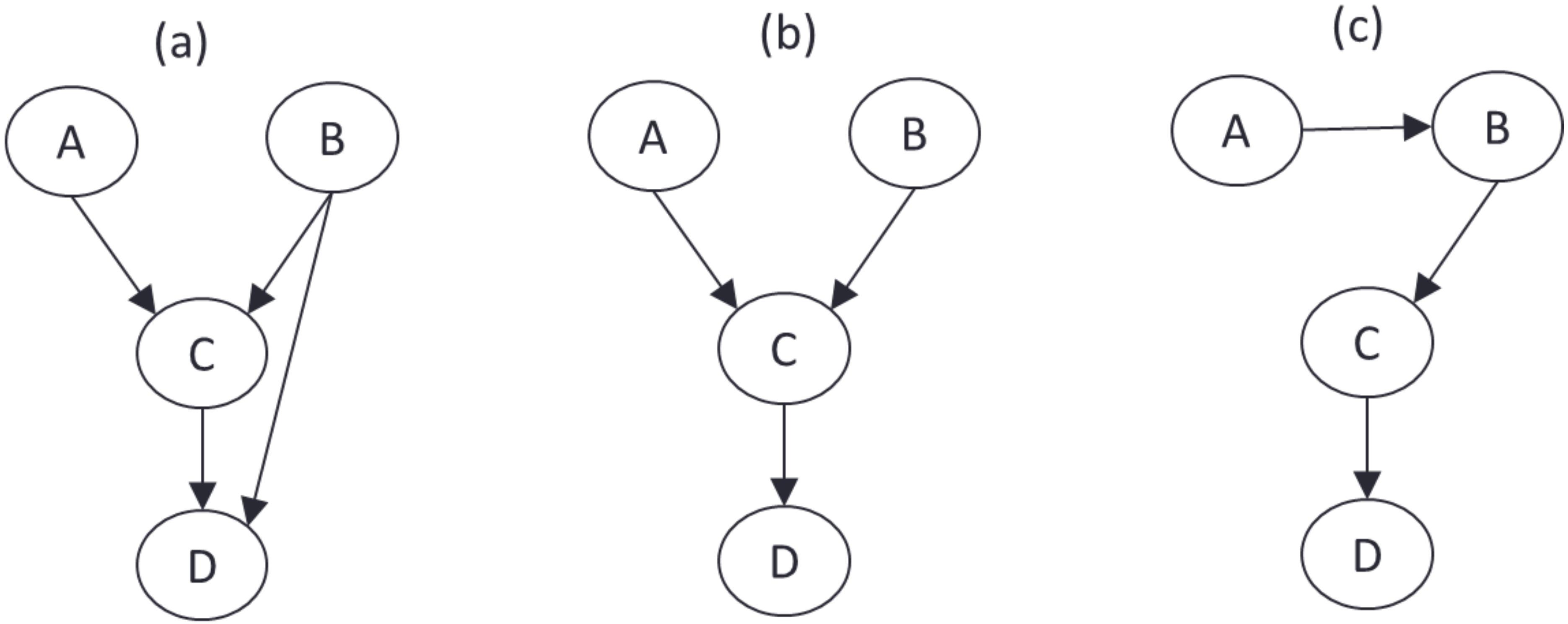
- Open paths are called d-connecting paths (given a set of variables).
- If no open path exists, the endpoints are d-separated (given the set of variables).
- Otherwise, the endpoints are d-connected (given the set of variables)

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 $\boldsymbol{Z}$.

Bayes Ball Algorithm



Practice: Find d-separations



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)

Markov Condition entails some independencies $\equiv$ Factoriazation
$P(X_1, \dots, X_n) = \prod P(X_i | Pa(X_i))$

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

If $I(G) \subseteq I(P)$ then G is an I-Map for P

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):

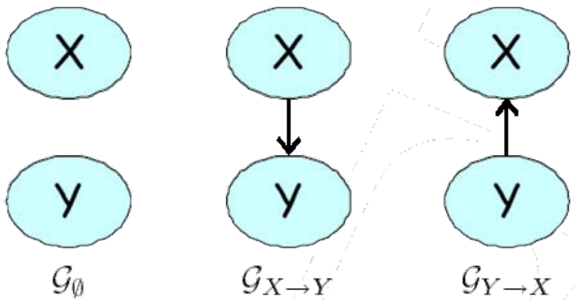
$$I(G) \subseteq I(P)$$

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

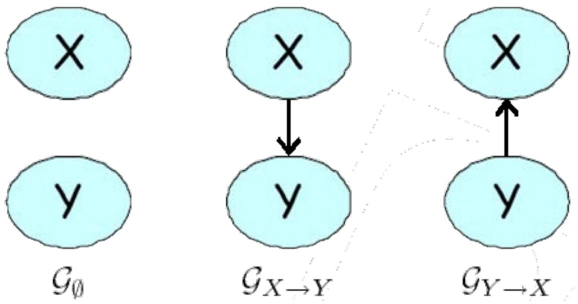
X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

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

The complete graph is trivially an I-map for any distribution

Minimal 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).

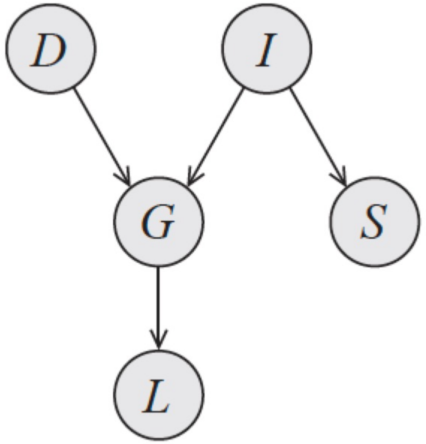
- Defn: Let K be *any graph object* associated with a set of independencies I(K). We say that K is an *I-map* for a set of independencies I, if $I(K) \subseteq I$

- A graph G is a minimal I-map for I if
  - it is an I-map for I,
  - The removal of even a single edge from G renders it not an I-map.

Constructing minimal I-maps

Algorithm 3.2 Procedure to build a minimal I-map given an ordering

Procedure Build-Minimal-I-Map (
  $X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
  $\mathcal{I}$ // Set of independencies
)
  Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
  for $i = 1, \ldots, n$
    $U \leftarrow \{X_1, \ldots, X_{i-1}\}$ // U is the current candidate for parents of X_i
    for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$
      if $U' \subset U$ **and** $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in \mathcal{I}$ **then**
        $U \leftarrow U'$
      // At this stage U is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
      // Now set U to be the parents of X_i
    for $X_j \in U$
      Add $X_j \rightarrow X_i$ to $\mathcal{G}$
  return $\mathcal{G}$



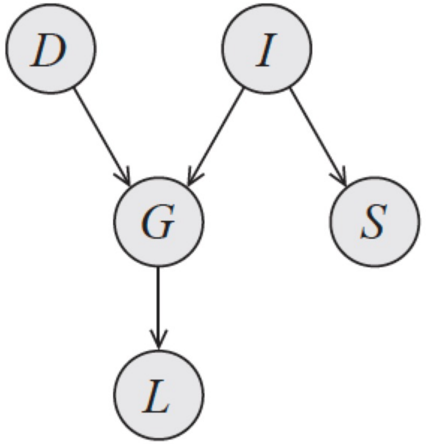
L; S; G; I; D;

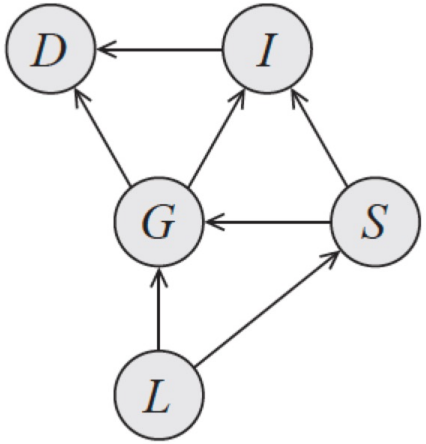
L; D; S; I; G

Constructing minimal I-maps

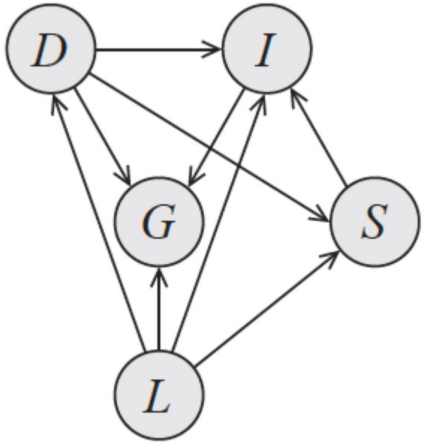
Algorithm 3.2 Procedure to build a minimal I-map given an ordering

Procedure Build-Minimal-I-Map (
$X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
$\mathcal{I}$ // Set of independencies
)
Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
for $i = 1, \ldots, n$
$\boldsymbol{U} \leftarrow \{X_1, \ldots, X_{i-1}\}$ // $\boldsymbol{U}$ is the current candidate for parents of X_i
for $\boldsymbol{U}' \subseteq \{X_1, \ldots, X_{i-1}\}$
if $\boldsymbol{U}' \subset \boldsymbol{U}$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U}' \mid \boldsymbol{U}') \in \mathcal{I}$ **then**
$\boldsymbol{U} \leftarrow \boldsymbol{U}'$
// At this stage $\boldsymbol{U}$ is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U} \mid \boldsymbol{U})$
// Now set $\boldsymbol{U}$ to be the parents of X_i
for $X_j \in \boldsymbol{U}$
Add $X_j \rightarrow X_i$ to $\mathcal{G}$
return $\mathcal{G}$





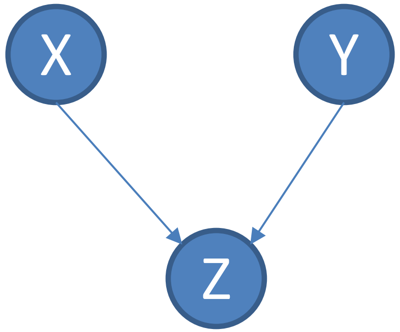
L; S; G; I; D;



L; D; S; I; G

Find a minimal I-map

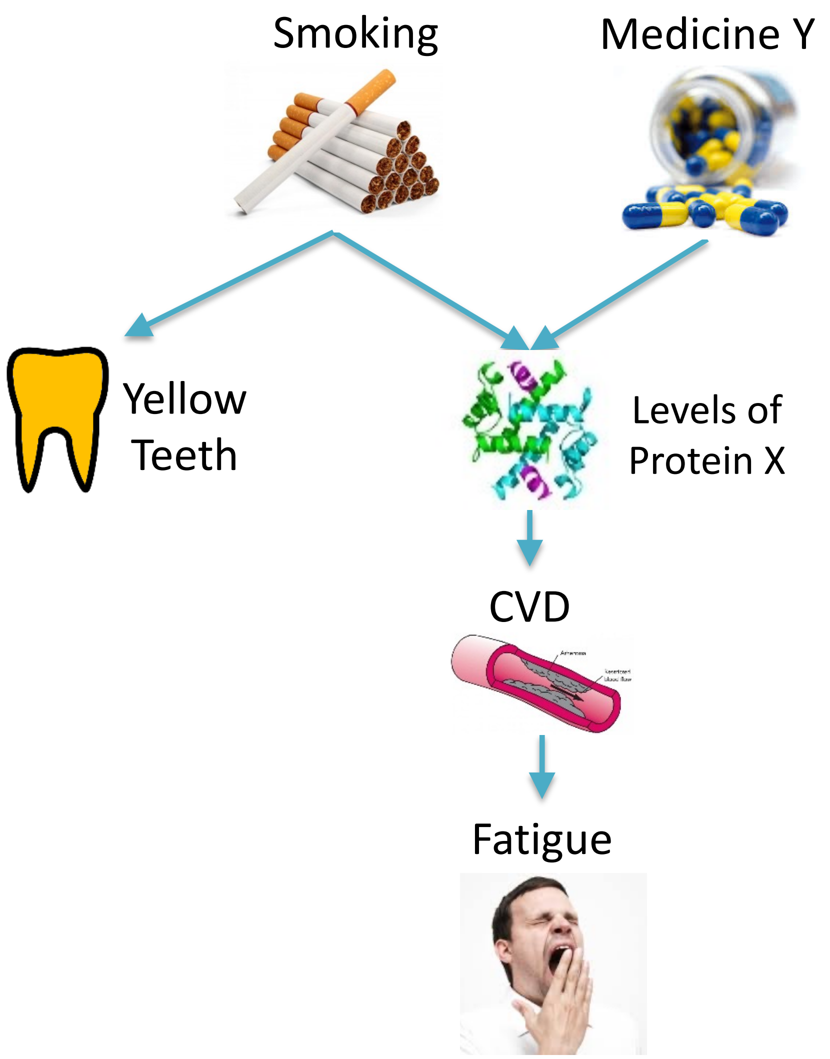
$$P(x,y,z) = \begin{cases} 1/12 & x \oplus y \oplus z = \text{ false} \\ 1/6 & x \oplus y \oplus z = \text{ true} \end{cases}$$



d-connection and conditional dependencies

You want to know if $\mathrm{A} \perp\!\!\!\perp \mathrm{B} | \mathbf{Z}$ in the JPD:

1. Find the paths from A to B in the graph (ignoring orientations).
2. If there exists no open path given $\boldsymbol{Z}$, then $\mathrm{A} \perp\!\!\!\perp \mathrm{B} | \mathbf{Z}$.
3. Else?



Faithfulness

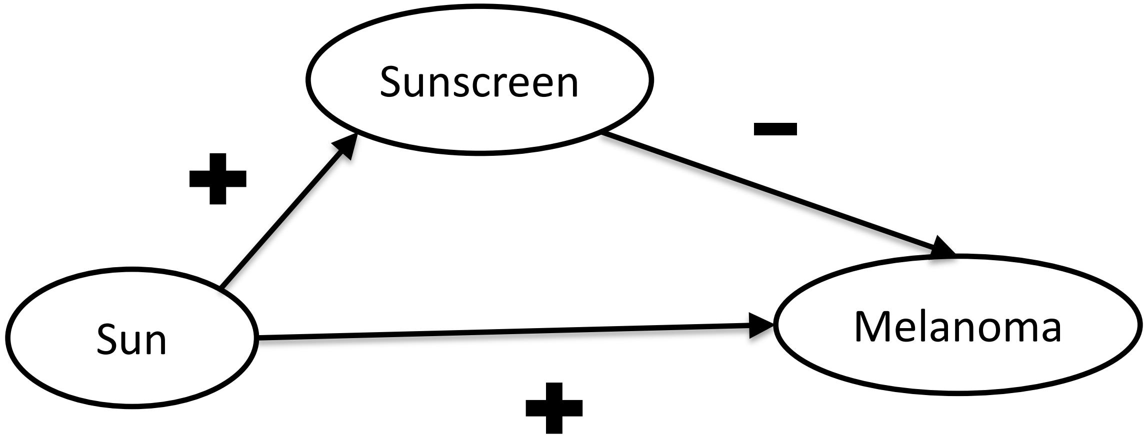
Faithfulness Condition:
Independences stem **only** from the causal structure, **not the parameterization** of the distribution.
We say that the graph and the distribution are **faithful to each other**.

MC $DSep(A, B|\mathbf{Z})$ in $G \Rightarrow A \perp\!\!\!\perp B|\mathbf{Z}$ in J

MC+FAITHFULNESS $DSep(A, B|\mathbf{Z})$ in $G \Leftrightarrow A \perp\!\!\!\perp B|\mathbf{Z}$ in J

Independencies stem from the causal structure, are not accidental properties of the parameters

Faithfulness



The parameters do not cancel each other out!

Faithfulness

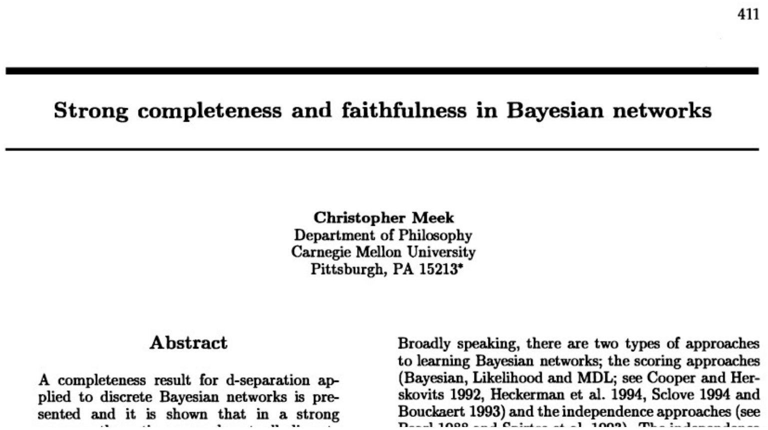
Is it realistic?

Assume you are given a graph and you select the parameters of the conditional probability tables randomly following a Dirichlet distribution.

The probability you get a non-faithful BN is zero (Lebesque measure is zero).

Helpful to devise efficient asymptotically correct methods.

[Meek. C. UAI 1995]



411

Strong completeness and faithfulness in Bayesian networks

Christopher Meek
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213*

Abstract

A completeness result for d-separation applied to discrete Bayesian networks is presented and it is shown that in a strong

Broadly speaking, there are two types of approaches to learning Bayesian networks; the scoring approaches (Bayesian, Likelihood and MDL; see Cooper and Herskovits 1992, Heckerman et al. 1994, Sclove 1994 and Bouckaert 1993) and the independence approaches (see

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):

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

Faithfulness

Is it realistic?

Probable causes of non-faithfulness:

- Too low associations are not detectable for finite samples.
- Too high correlations (determinism or close-to-determinism).
- Natural selection may be biasing towards creating non-faithful distributions in systems in nature (e.g.. cells)!
- Not all joint probability distributions have a faithful representation.

The probability of getting an almost non-faithful distribution is non-zero.

Soundness and completeness

D-separation is sound and "complete" w.r.t. BN factorization law

Soundness:

Theorem: If a distribution P factorizes according to G, then $I(G) \subseteq I(P)$.

"Completeness":

"Claim": For any distribution P that factorizes over G, if $(X \perp Y \mid Z) \in I(P)$ then $d\text{-}sep_G(X; Y \mid Z)$.

Contrapositive of the completeness statement

- "If X and Y are *not d-separated* given Z in G, then X and Y are *dependent in all* distributions P that factorize over G."
- Is this true?

Soundness and completeness

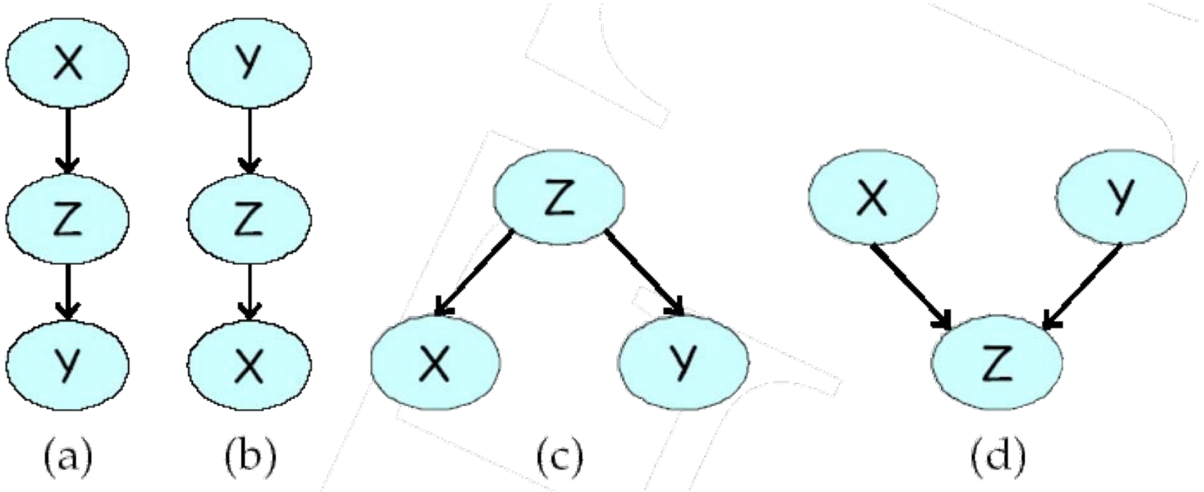
- Contrapositive of the completeness statement
  - "If X and Y are *not d-separated* given Z in G, then X and Y are *dependent in all* distributions P that factorize over G."
  - Is this true?
- No. Even if a distribution factorizes over G, it can still contain additional independencies that are not reflected in the structure
  - Example: graph A->B, for actually independent A and B (the independence can be captured by some subtle way of parameterization)

A	b^0	b^1
a^0	0.4	0.6
a^1	0.4	0.6

- Thm: Let G be a BN graph. If X and Y are not d-*separated* given Z in G, then X and Y are *dependent in some* distribution P that factorizes over G.

Uniqueness of BN

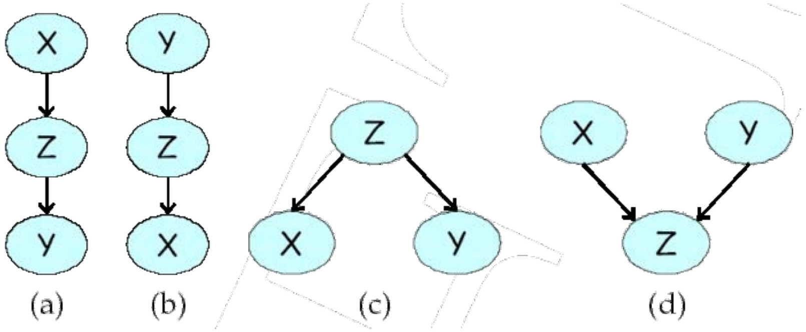
- Very different BN graphs can actually be equivalent, in that they encode precisely the same set of conditional independence assertions.



$(X \perp Y \mid Z)$.

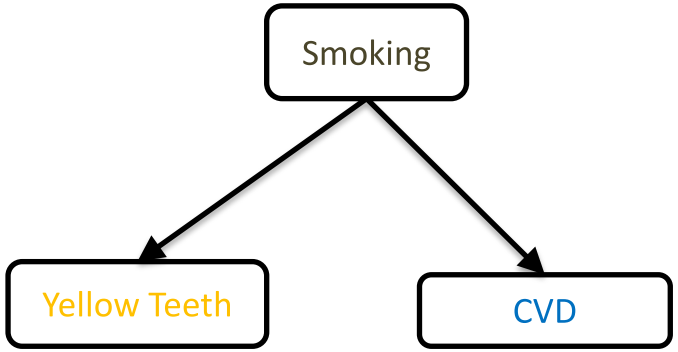
Markov Equivalence

- Defn : Two BN graphs G_1 and G_1 over X are *I-equivalent or Markov Equivalent* if $I(G_1) = I(G_2)$.
  - The set of all graphs over X is partitioned into a set of mutually exclusive and exhaustive *I-equivalence classes*, which are the set of equivalence classes induced by the I-equivalence relation.

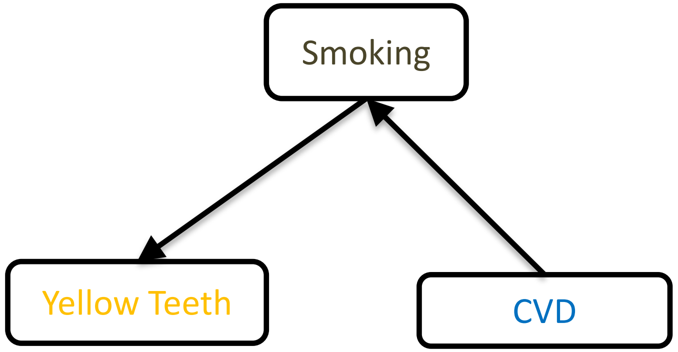


- Any distribution P that can be factorized over graphs (a), (b), or (c) can be factorized over the other.
- Furthermore, there is no intrinsic property of P that would allow us associate it with one graph rather than an equivalent one.
- This observation has important implications with respect to our ability to determine the directionality of influence.

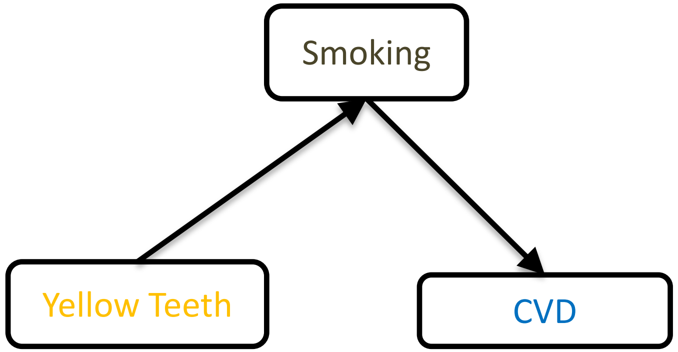
Markov Equivalence



CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking



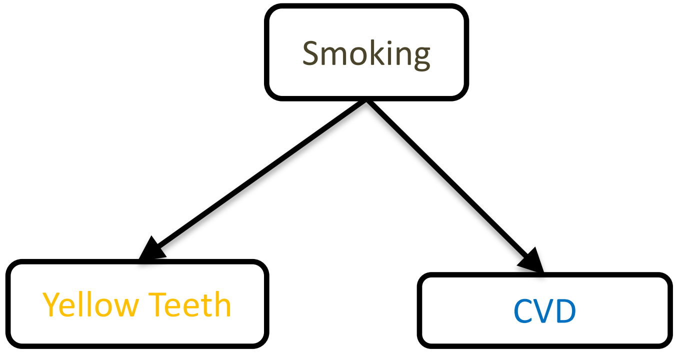
CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking



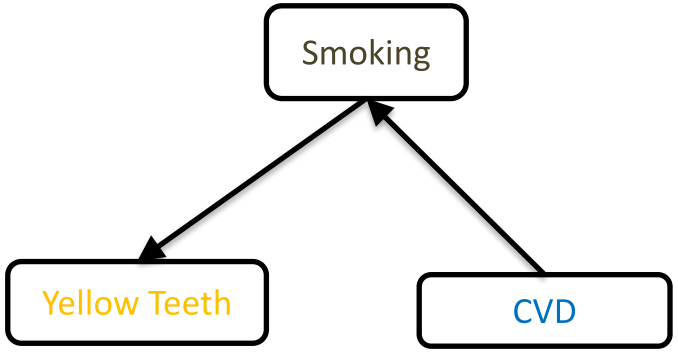
CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

Markov Condition entails the same conditional independence for all three graphs.

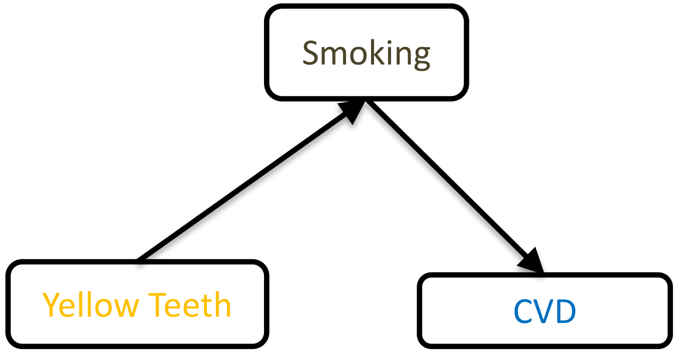
Markov Equivalence



$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



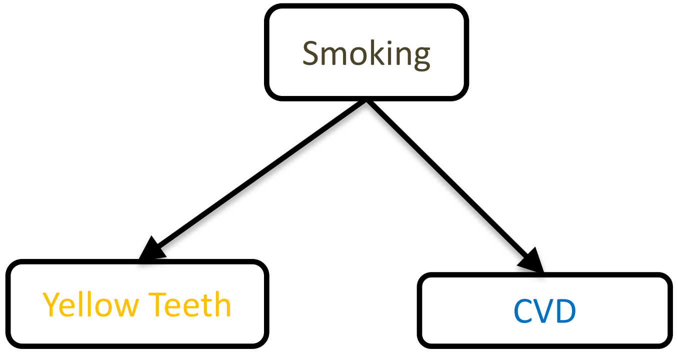
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



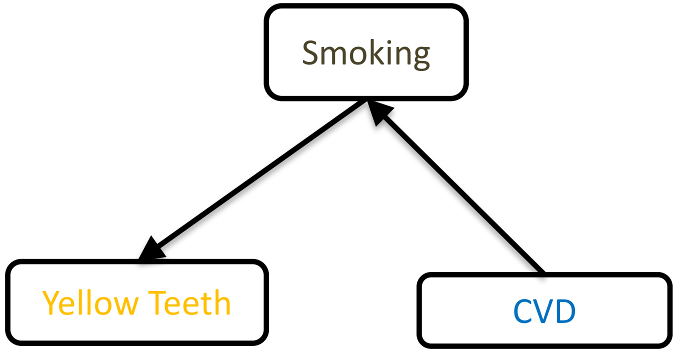
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

- The graphs are called **Markov Equivalent.**
- All Markov equivalent graphs denote a **Markov equivalence class (MEC).**
- **We use [G] to denote the MEC of G.**

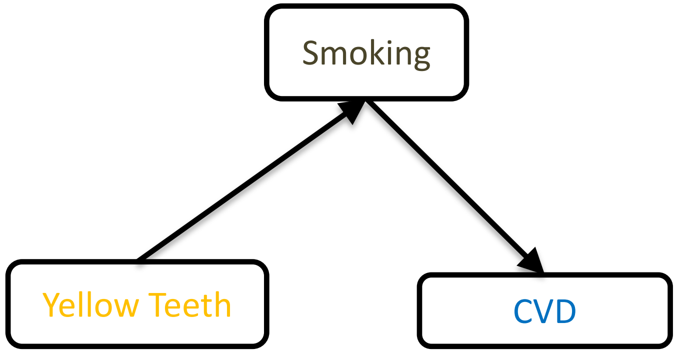
Markov Equivalence



$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



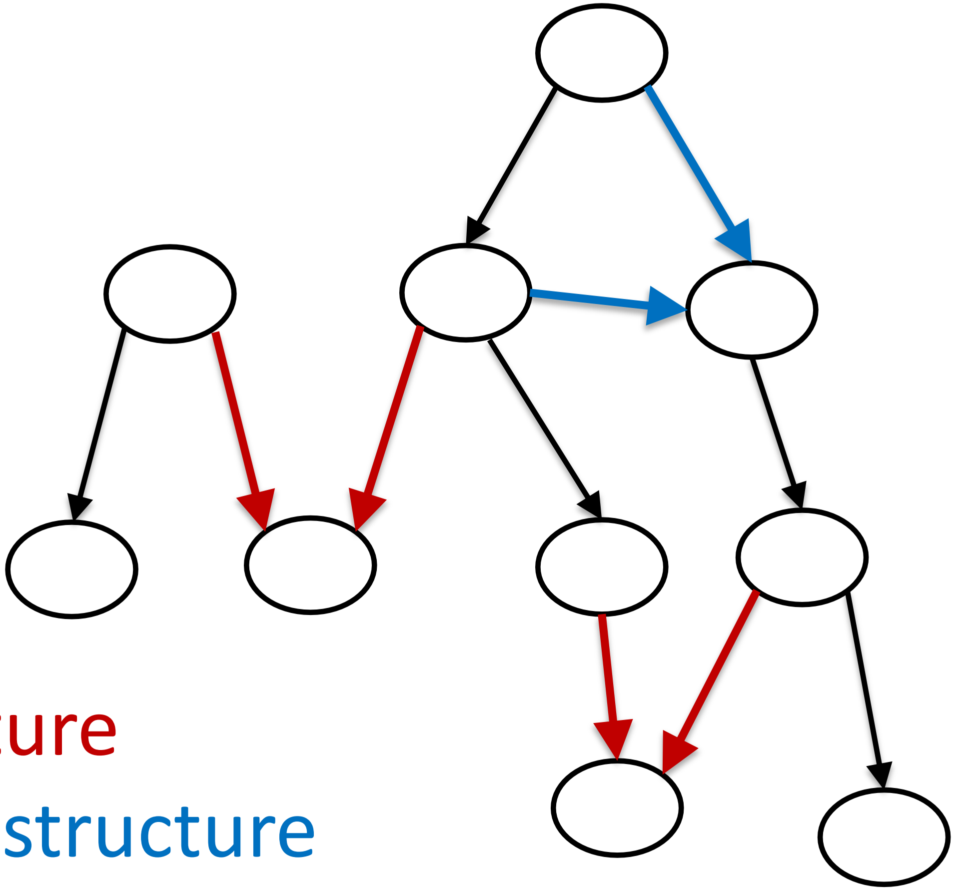
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

Markov Equivalent Graphs share

- the same skeleton (adjacencies).
- the same unshielded colliders

Characterization of the Markov Equivalence Class

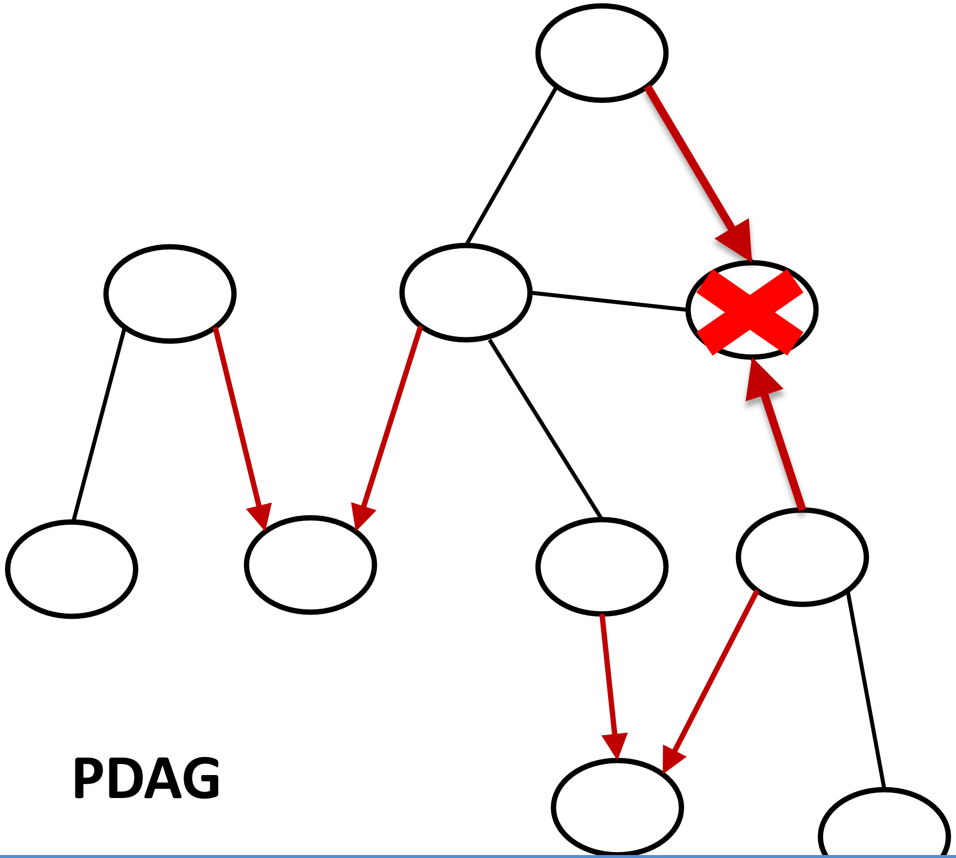
- Unshielded collider: A collider (X-Y-Z) where the endpoints (X, Z) are NOT adjacent.
- AKA v-structure.



- v-structure
- not a v-structure

Pattern DAGs

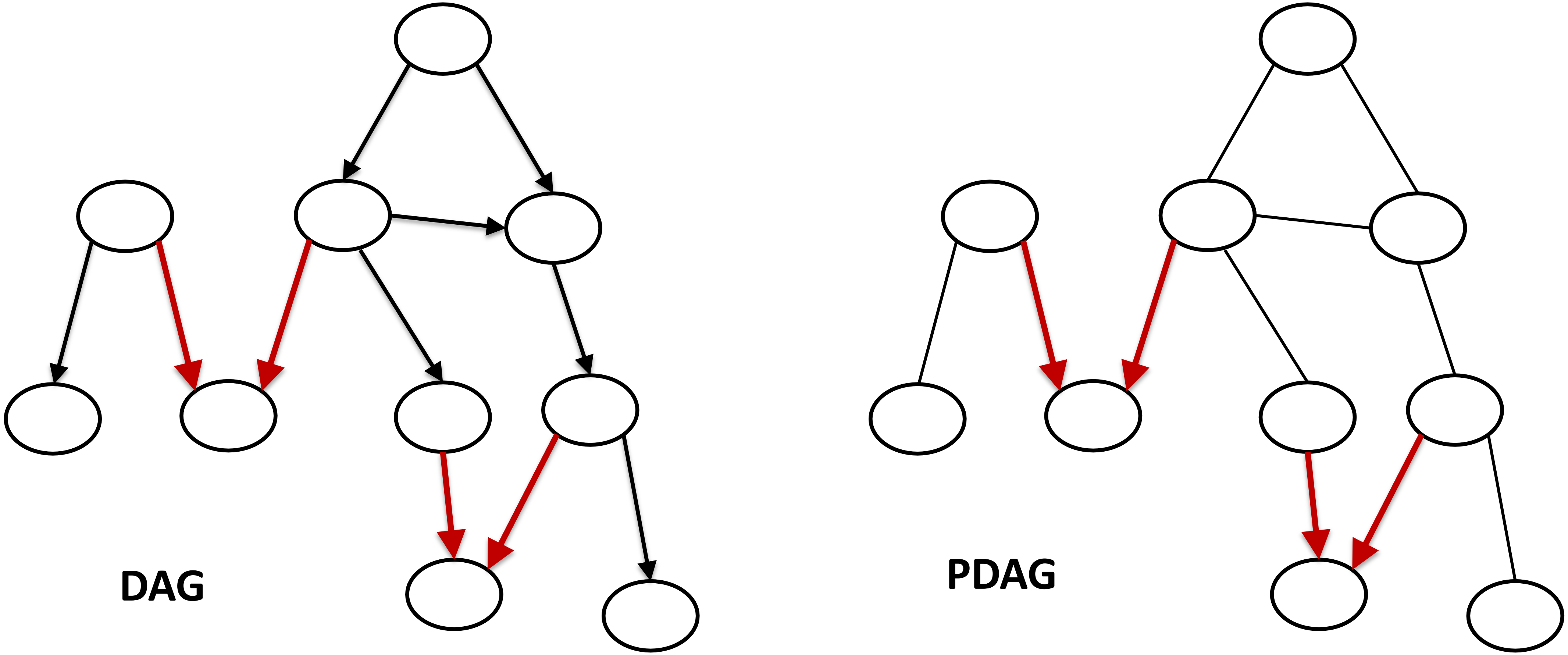
- Represents a class of Markov Equivalent DAGs.
- Has the same edges as every DAG in the class.
- Has only orientations (arrows) shared by all the DAGs in the class.
- Orient the PDAG as a DAG without creating a new collider or directed cycle!



PDAG

Not all configurations are possible!

Pattern DAGs

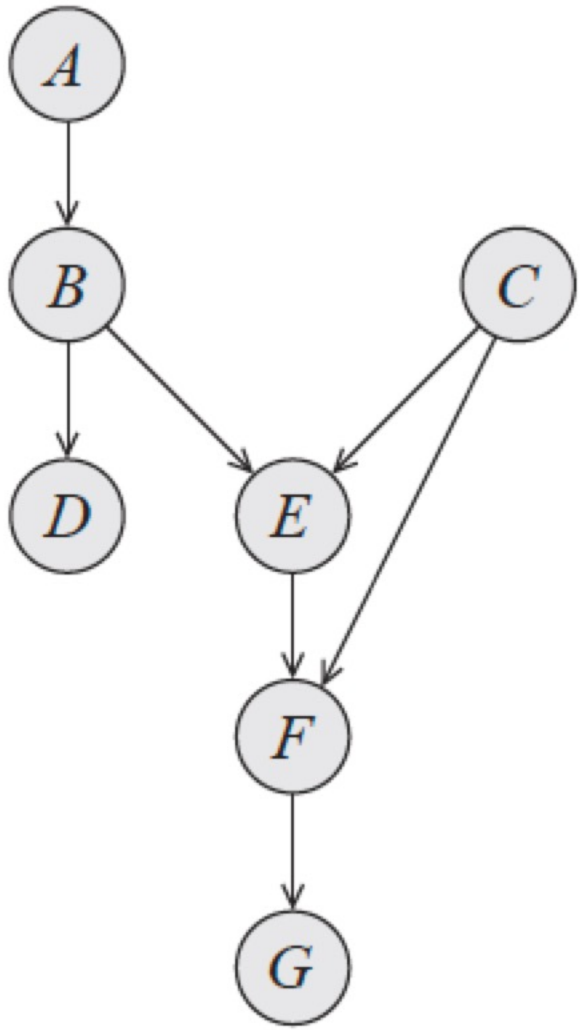


- You can still "read" all conditional independencies entailed by the Causal Markov Condition in the graph using d-separation.

Practice

How many networks are equivalent to the simple directed chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$

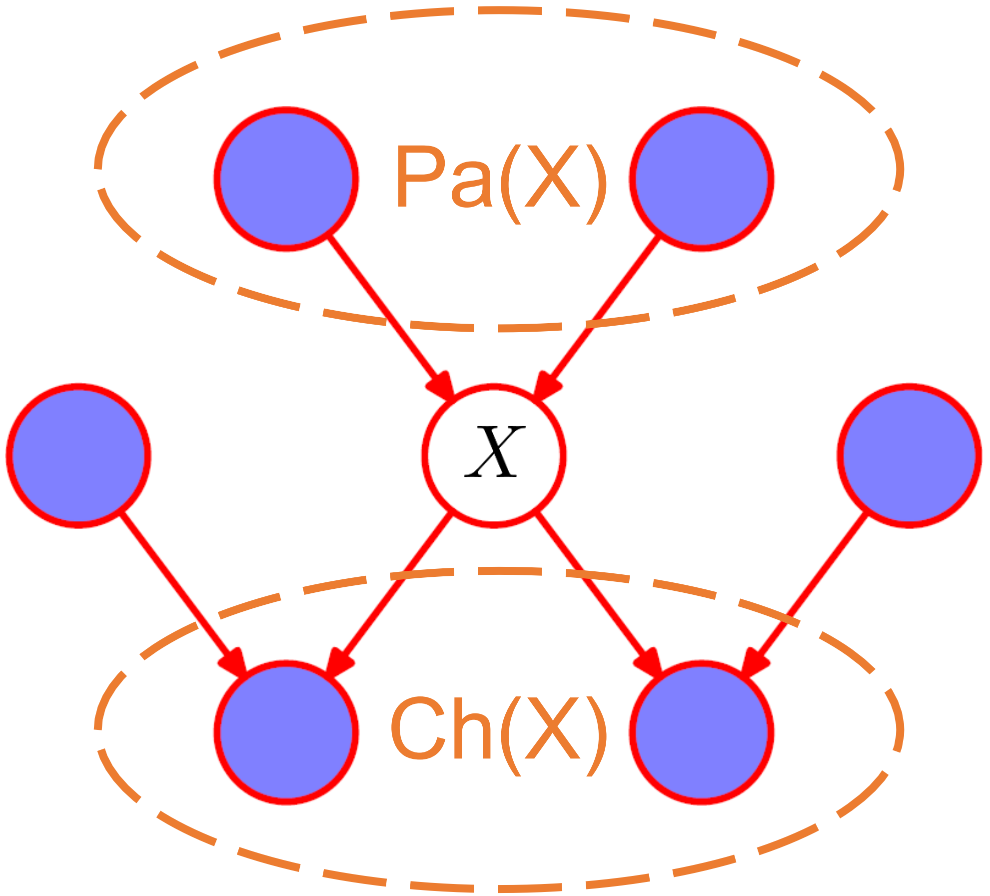
Practice



Which edges can you reverse?

Markov Blanket

X *conditionally independent of all other nodes, given its Markov blanket*



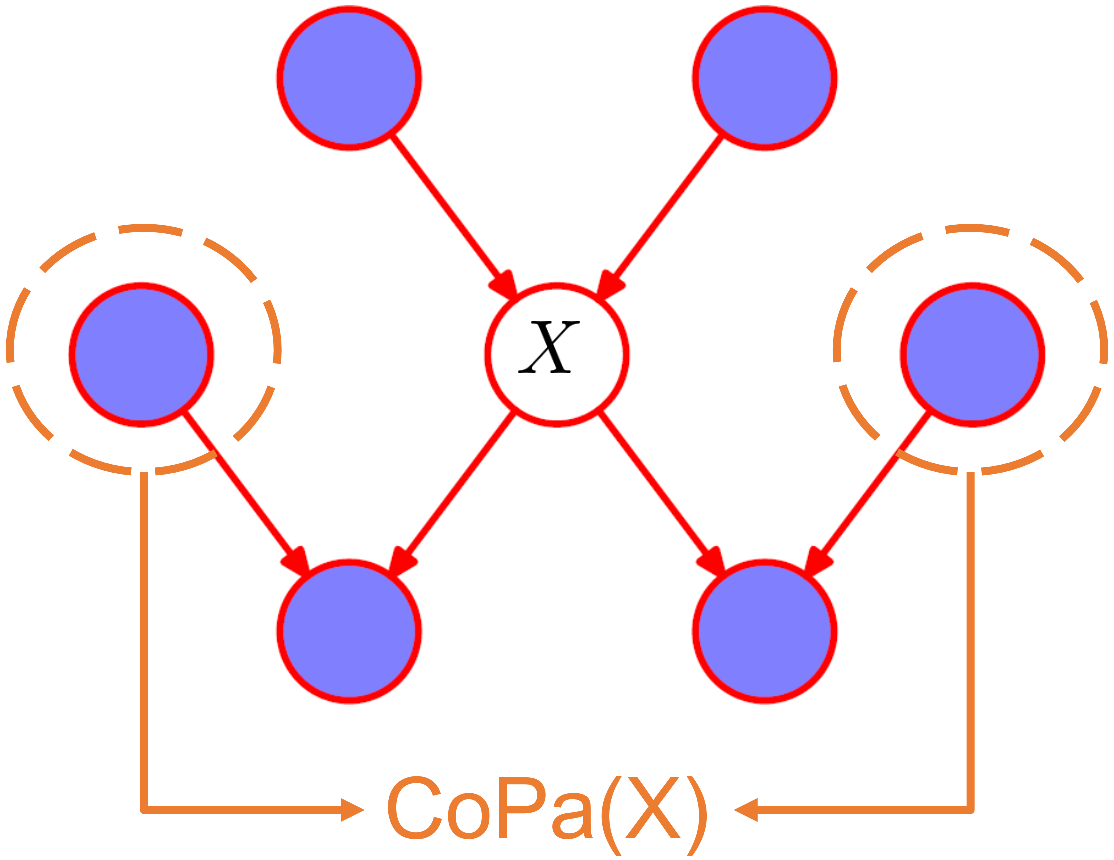
Markov Boundary

X conditionally independent of all other nodes, given its Markov blanket

The minimal Markov blanket is the Markov boundary

Q: Why co-parents?

A: Explaining away



Markov Blanket

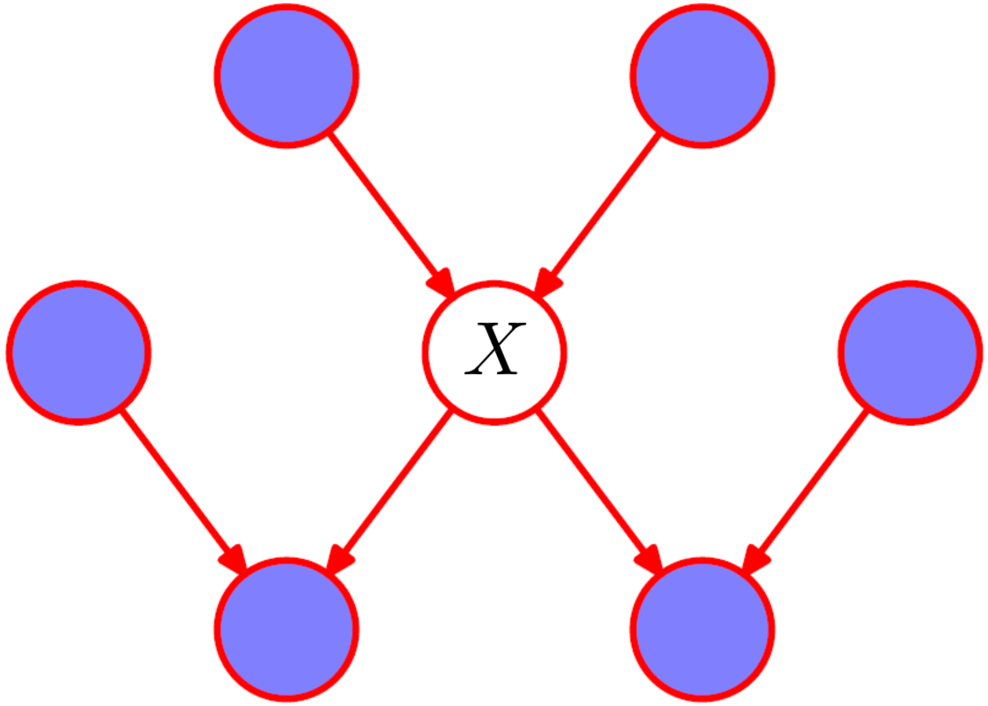
X conditionally independent of all other nodes, given its Markov blanket

Definition *A RV X with distribution p(x) that is Markov w.r.t. graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has a* **Markov blanket** *given by:*

$$\text{Mb}(X) = \text{Pa}(X) \cup \text{Ch}(X) \cup \text{CoPa}(X)$$

For any $Y \notin \text{Mb}(X)$:

$$X \perp Y \mid \text{Mb}(X)$$



Markov boundaries are used to simplify inference and distribute computation (e.g. Gibbs sampler, variational inference, etc.)

Markov Blanket

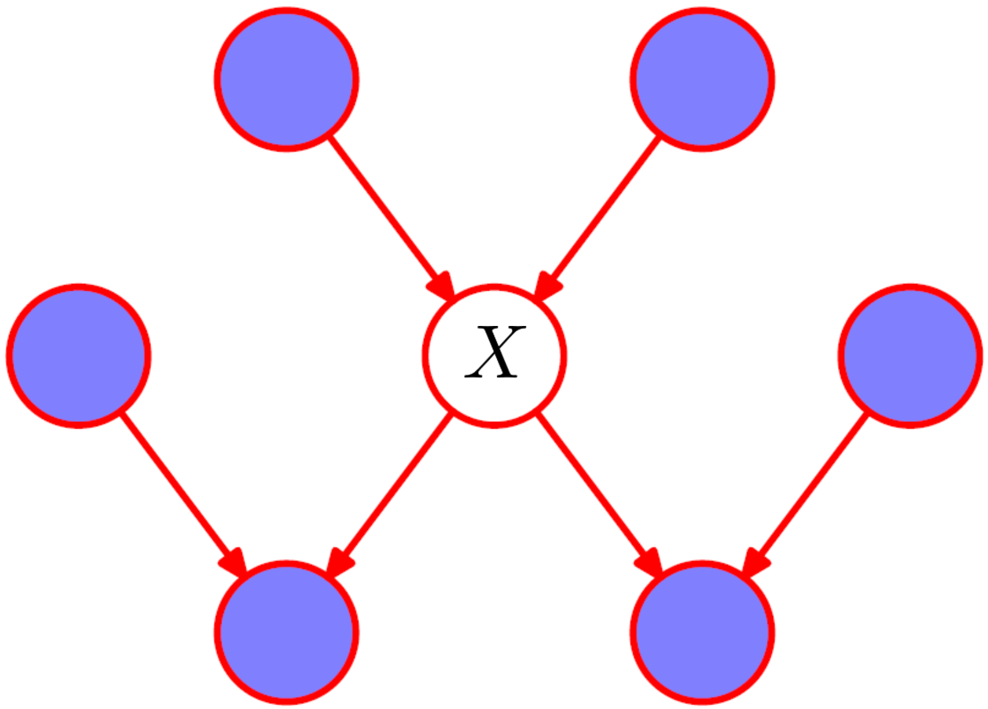
X conditionally independent of all other nodes, given its Markov blanket

Definition *A RV X with distribution p(x) that is Markov w.r.t. graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has a* **Markov blanket** *given by:*

$$\text{Mb}(X) = \text{Pa}(X) \cup \text{Ch}(X) \cup \text{CoPa}(X)$$

For any $Y \notin \text{Mb}(X)$:

$$X \perp Y \mid \text{Mb}(X)$$



They also lead to the optimal prediction of X (for a proper scoring rule)

Practice

(a)	$P(t^1 \mid d^1)$	$P(t^1)$
(b)	$P(d^1 \mid t^0)$	$P(d^1)$
(c)	$P(h^1 \mid e^1, f^1)$	$P(h^1 \mid e^1)$
(d)	$P(c^1 \mid f^0)$	$P(c^1)$
(e)	$P(c^1 \mid h^0)$	$P(c^1)$
(f)	$P(c^1 \mid h^0, f^0)$	$P(c^1 \mid h^0)$
(g)	$P(d^1 \mid h^1, e^0)$	$P(d^1 \mid h^1)$
(h)	$P(d^1 \mid e^1, f^0, w^1)$	$P(d^1 \mid e^1, f^0)$
(i)	$P(t^1 \mid w^1, f^0)$	$P(t^1 \mid w^1)$

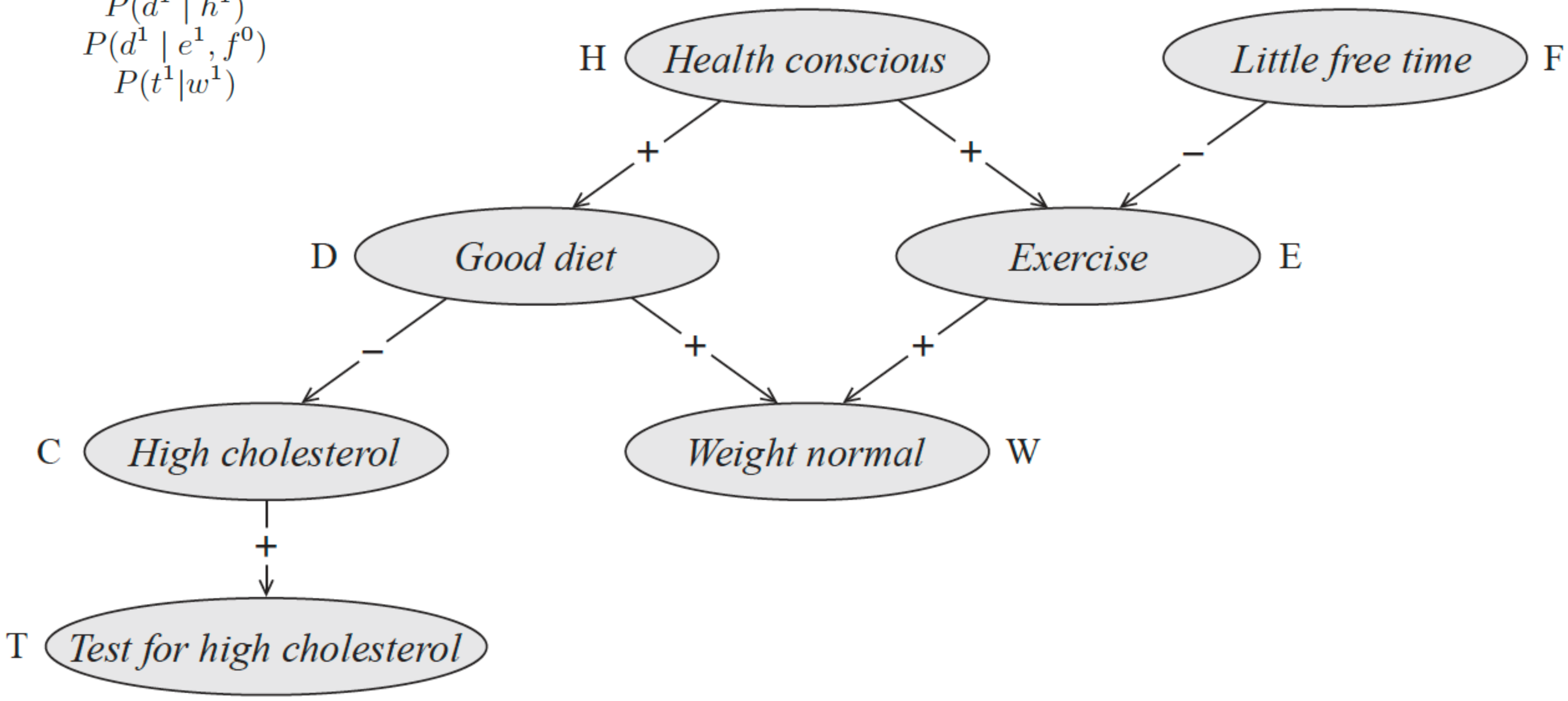


Figure 3.14 A Bayesian network with qualitative influences