DAGs, I-Maps, Factorization, d-Separation, Minimal I-Maps, Bayesian Networks

Slides by Nir Friedman

.

Probability Distributions

- Let $X_1,\ldots,X_n$ be random variables
- Let P be a joint distribution over $X_1,\ldots,X_n$

If the variables are binary, then we need $O(2^n)$ parameters to describe P

Can we do better?

- **Key idea**: use properties of independence

Independent Random Variables

- Two variables X and Y are **independent** if
  - $P(X = x|Y = y) = P(X = x)$ for all values x,y
  - That is, learning the values of Y does not change prediction of X

- If X and Y are independent then
  - $P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$
- In general, if $X_1,\ldots,X_n$ are independent, then
  - $P(X_1,\ldots,X_n) = P(X_1)\ldots P(X_n)$
  - Requires O(n) parameters

Conditional Independence

- Unfortunately, most of random variables of interest are not independent of each other

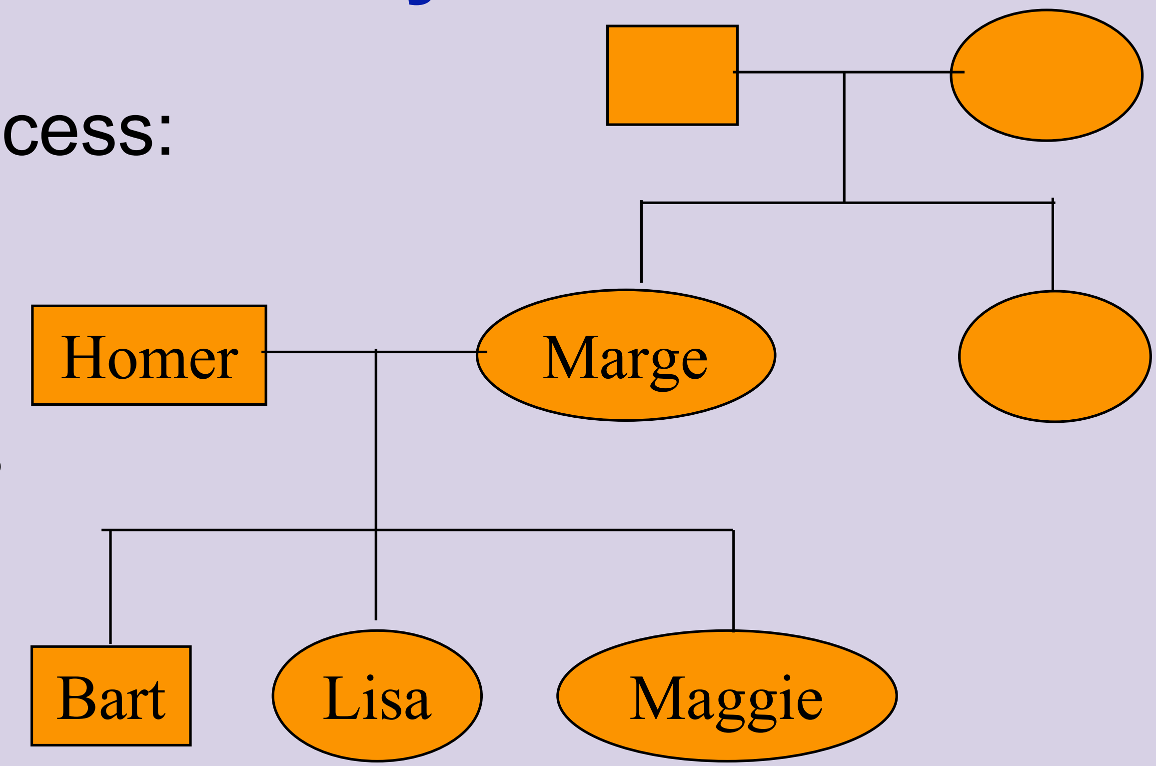
- A more suitable notion is that of **conditional independence**
- Two variables X and Y are **conditionally independent** given Z if
  - $P(X = x | Y = y, Z=z) = P(X = x | Z=z)$ for all values x,y,z
  - That is, learning the values of Y does not change prediction of X once we know the value of Z
  - notation: $Ind(X ; Y | Z)$

Example: Family trees

Noisy stochastic process:

Example: Pedigree

- A node represents an individual's genotype

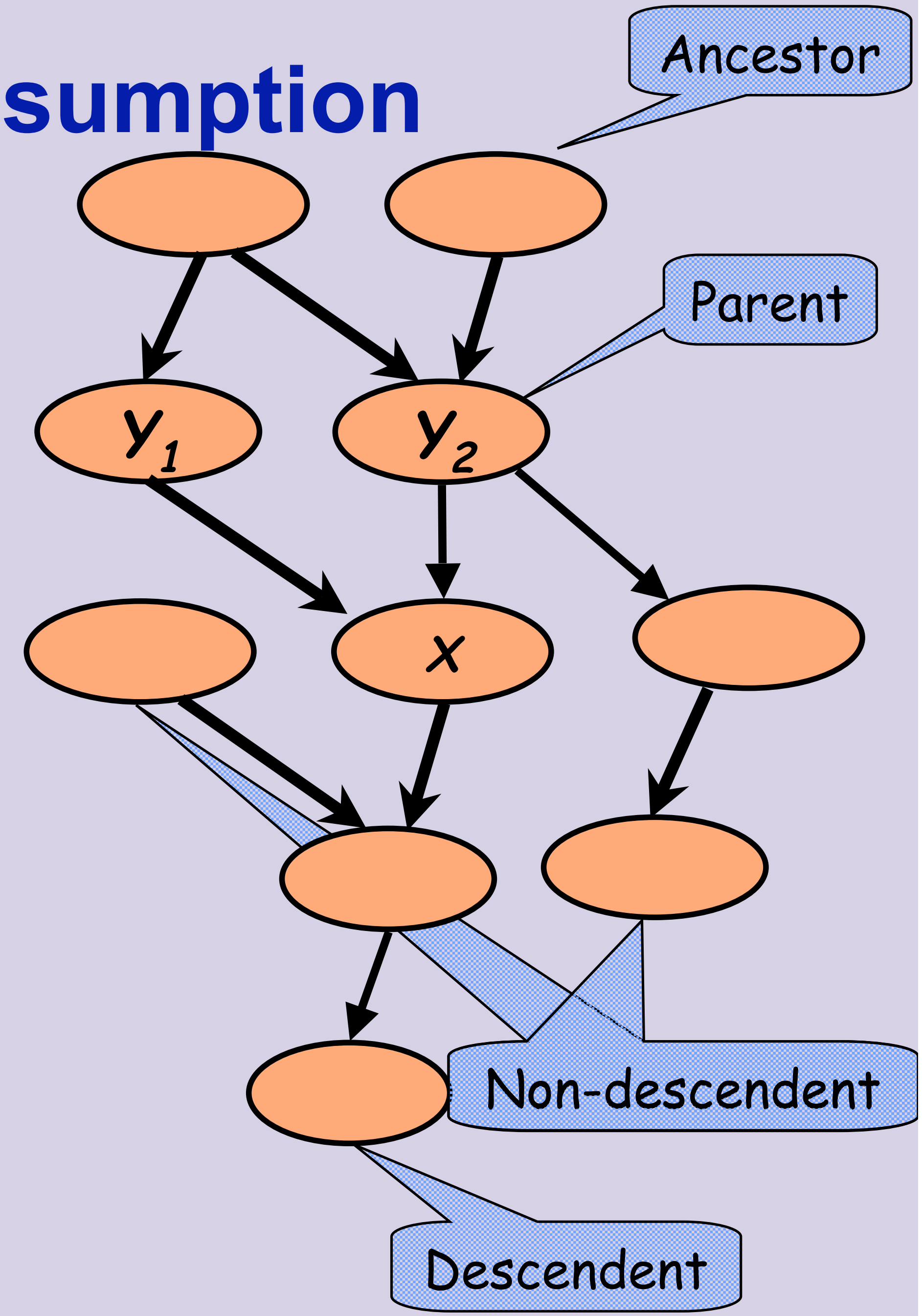


Modeling assumptions:

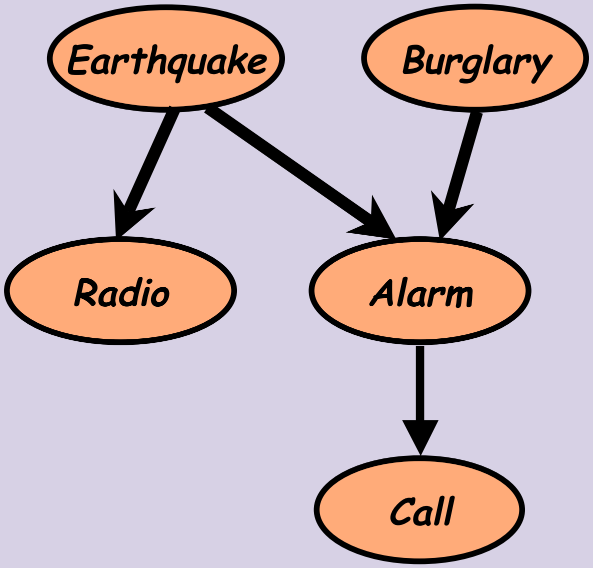
Ancestors can effect descendants' genotype only by passing genetic materials through intermediate generations

Markov Assumption

- We now make this independence assumption more precise for **directed acyclic graphs** (DAGs)
- Each random variable *X*, is independent of its non-descendents, given its parents *Pa(X)*
- Formally,
  Ind(X; NonDesc(X) | Pa(X))



Markov Assumption Example



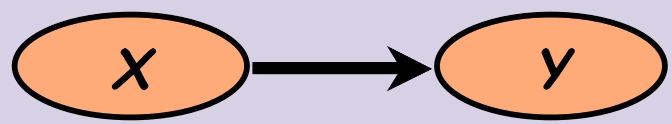
- In this example:
  - *Ind(E; B)*
  - *Ind(B; E, R)*
  - *Ind(R; A, B, C | E)*
  - *Ind(A; R | B,E)*
  - *Ind(C; B, E, R | A)*

I-Maps

- A DAG *G* is an **I-Map** of a distribution *P* if the all Markov assumptions implied by *G* are satisfied by *P*

  (Assuming *G* and *P* both use the same set of random variables)

Examples:



x	y	P(x,y)
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x	y	P(x,y)
0	0	0.2
0	1	0.3
1	0	0.4
1	1	0.1

Factorization

- Given that G is an I-Map of P, can we simplify the representation of P?
- Example:

- Since $Ind(X;Y)$, we have that $P(X|Y) = P(X)$
- Applying the chain rule

$$P(X,Y) = P(X|Y)\,P(Y) = P(X)\,P(Y)$$

- Thus, we have a simpler representation of $P(X,Y)$

Factorization Theorem

Thm: if G is an I-Map of P, then

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa(X_i))$$

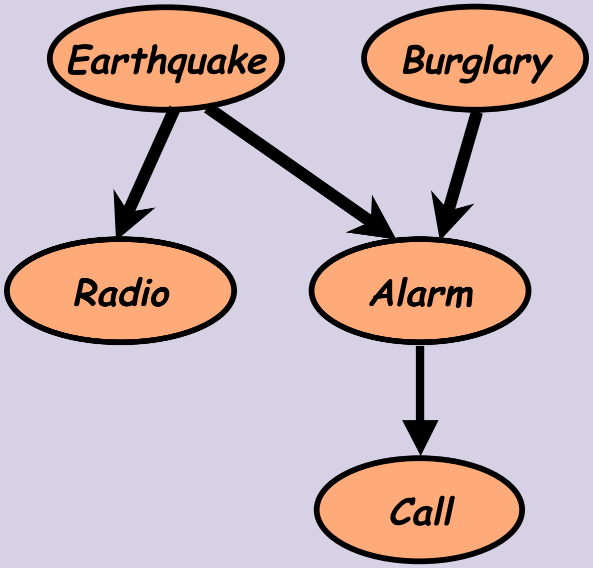
Proof:

- By chain rule: $P(X_1,\ldots,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$
- wlog. $X_1,\ldots,X_n$ is an ordering consistent with G
- From assumption: $Pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$

  $\{X_1, \ldots, X_{i-1}\} - Pa(X_i) \subseteq NonDesc(X_i)$

- Since G is an I-Map, $Ind(X_i; NonDesc(X_i) \mid Pa(X_i))$
- Hence, $Ind(X_i; \{X_1, \ldots, X_{i-1}\} - Pa(X_i) \mid Pa(X_i))$
- We conclude, $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Pa(X_i))$

Factorization Example



$$P(C,A,R,E,B) =$$

$$P(B)P(E|B)P(R|E,B)P(A|R,B,E)P(C|A,R,B,E)$$

versus

$$P(C,A,R,E,B) = P(B)\ P(E)\ P(R|E)\ P(A|B,E)\ P(C|A)$$

Consequences

- We can write *P* in terms of “local” conditional probabilities

If *G* is **sparse**,

- that is, $|Pa(X_i)| < k$,

⇒ each conditional probability can be specified compactly

- e.g. for binary variables, these require $O(2^k)$ params.

⇒ representation of P is **compact**

- linear in number of variables

Conditional Independencies

- Let *Markov(G)* be the set of Markov Independencies implied by *G*
- The decomposition theorem shows

$$G \text{ is an I-Map of } P \;\Rightarrow\; P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa_i)$$

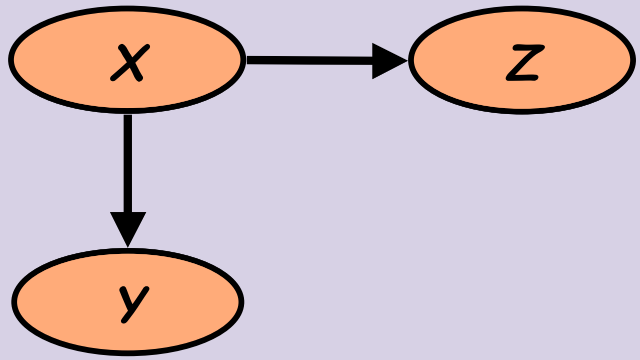
- We can also show the opposite:

Thm:

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa_i) \;\Rightarrow\; G \text{ is an I-Map of } P$$

Proof (Outline)

Example:



$$P(Z \mid X, Y) = \frac{P(X, Y, Z)}{P(X, Y)} = \frac{\cancel{P(X)}\,\cancel{P(Y \mid X)}\,P(Z \mid X)}{\cancel{P(X)}\,\cancel{P(Y \mid X)}}$$

$$= P(Z \mid X)$$

Implied Independencies

- Does a graph *G* imply additional independencies as a consequence of *Markov(G)*
- We can define a **logic** of independence statements
- We already seen some axioms:
  - $Ind(X ; Y \mid Z) \Rightarrow Ind(Y; X \mid Z)$
  - $Ind(X ; Y_1, Y_2 \mid Z) \Rightarrow Ind(X; Y_1 \mid Z)$
- We can continue this list..

d-seperation

- A procedure *d-sep(X; Y | Z, G)* that given a DAG *G*, and sets *X*, *Y*, and *Z* returns either *yes* or *no*

- **Goal**:

  d-sep(X; Y | Z, G) = yes iff *Ind(X;Y|Z)* follows from *Markov(G)*

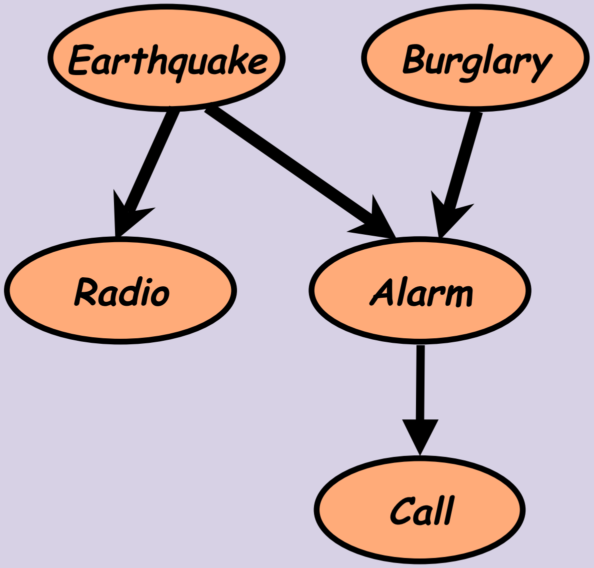
Paths

- **Intuition:** dependency must "flow" along paths in the graph

- A path is a sequence of neighboring variables

Examples:

- $R \leftarrow E \rightarrow A \leftarrow B$
- $C \leftarrow A \leftarrow E \rightarrow R$



Paths blockage

- We want to know when a path is
  - **active** -- creates dependency between end nodes
  - **blocked** -- cannot create dependency end nodes

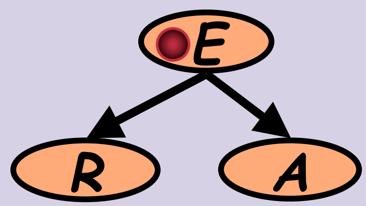
- We want to classify situations in which paths are active given the evidence.

Path Blockage

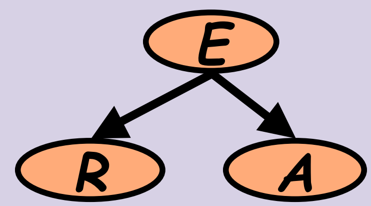
Three cases:

- Common cause
-
-

Blocked



Active

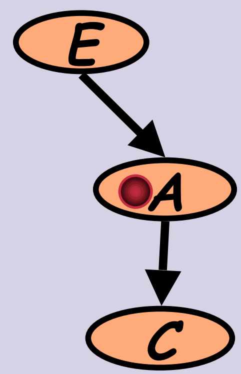


Path Blockage

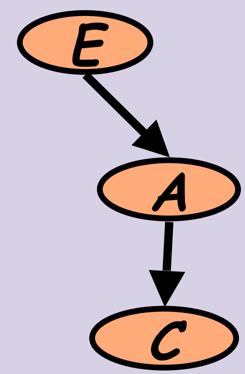
Three cases:

- Common cause
- Intermediate cause
-

Blocked



Active

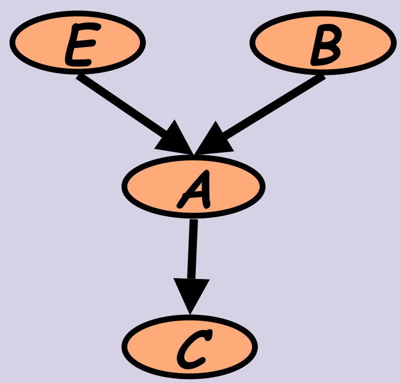


Path Blockage

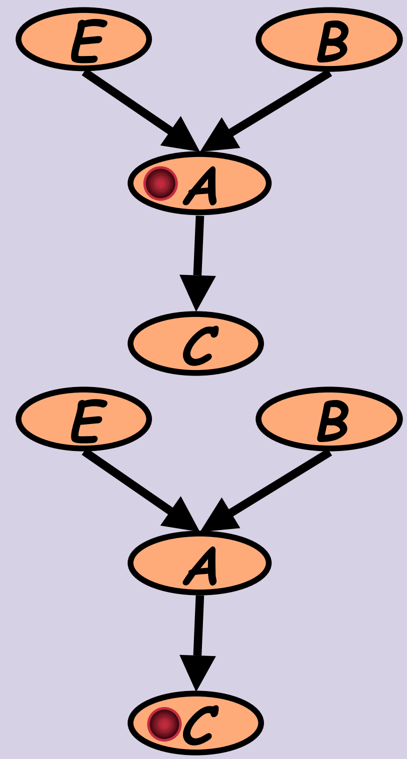
Three cases:

- Common cause
- Intermediate cause
- Common Effect

Blocked



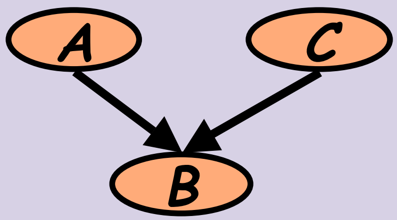
Active



Path Blockage -- General Case

A path is active, given evidence Z, if

- Whenever we have the configuration



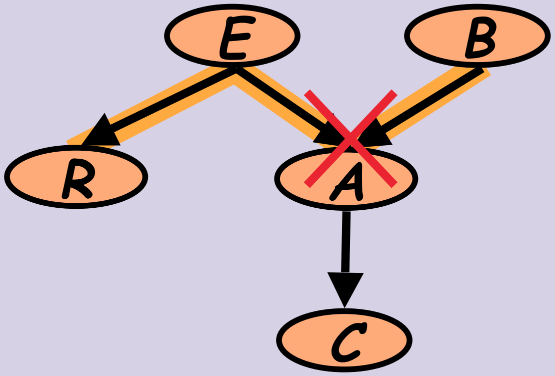
  B or one of its descendents are in Z

- No other nodes in the path are in Z

A path is blocked, given evidence Z, if it is not active.

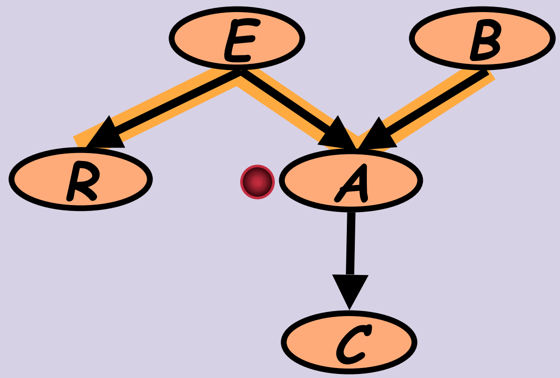
Example

- d-sep(R,B) = yes



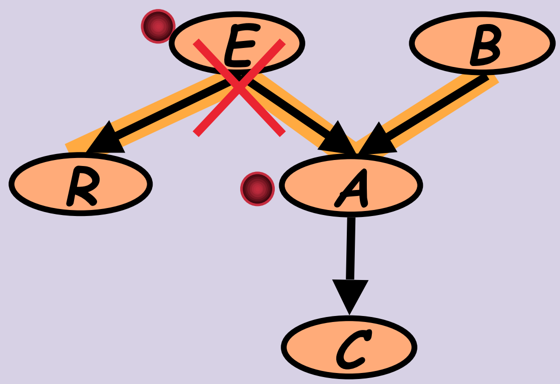
Example

- d-sep(R,B) = yes
- d-sep(R,B|A) = no



Example

- *d-sep(R,B) = yes*
- *d-sep(R,B|A) = no*
- *d-sep(R,B|E,A) = yes*



d-Separation

- X is **d-separated** from Y, given Z, if all paths from a node in X to a node in Y are blocked, given Z.

- Checking d-separation can be done efficiently (linear time in number of edges)
  - Bottom-up phase:
    Mark all nodes whose descendents are in Z
  - X to Y phase:
    Traverse (BFS) all edges on paths from X to Y and check if they are blocked

Soundness

Thm:

- If
  - *G* is an I-Map of *P*
  - *d-sep(X; Y | Z, G) = yes*
- then
  - *P* satisfies *Ind(X; Y | Z)*

Informally,

- Any independence reported by d-separation is satisfied by underlying distribution

Completeness

Thm:

- If *d-sep(X; Y | Z, G) = no*
- then there is a distribution *P* such that
  - *G* is an I-Map of *P*
  - *P* does not satisfy *Ind(X; Y | Z)*

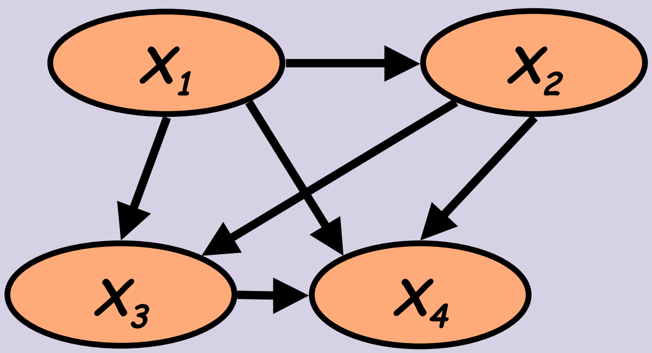
Informally,

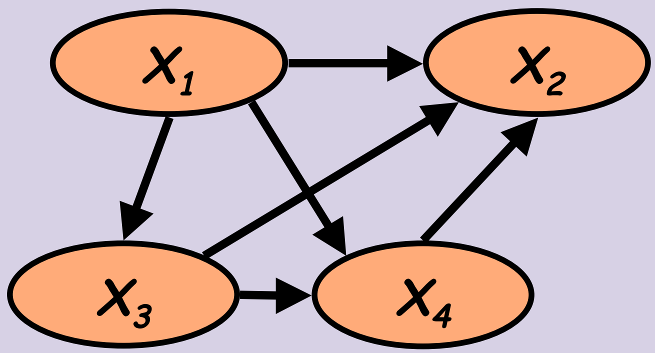
- Any independence not reported by d-separation might be violated by the by the underlying distribution
- We cannot determine this by examining the graph structure alone

I-Maps revisited

- The fact that G is I-Map of P might not be that useful

- For example, **complete** DAGs
  - A DAG is G is complete is we cannot add an arc without creating a cycle





- These DAGs do not imply any independencies
- Thus, they are I-Maps of any distribution

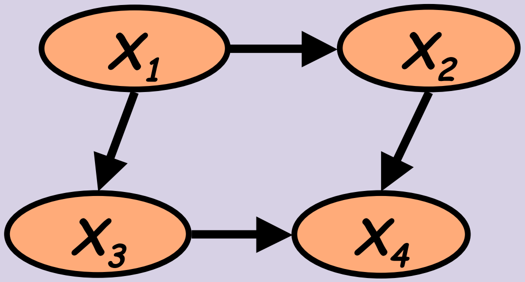
Minimal I-Maps

A DAG G is a **minimal I-Map** of P if

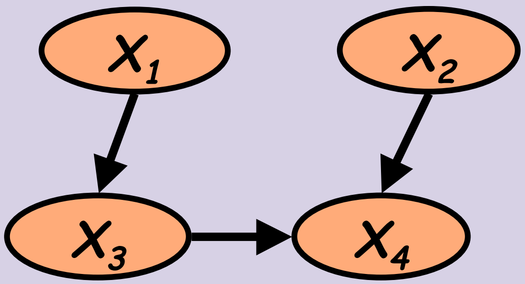
- G is an I-Map of P
- If $G' \subset G$, then G' is not an I-Map of P

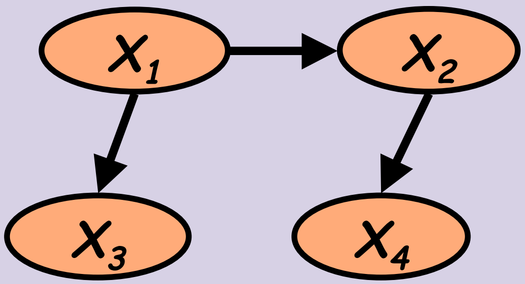
Removing any arc from G introduces (conditional) independencies that do not hold in P

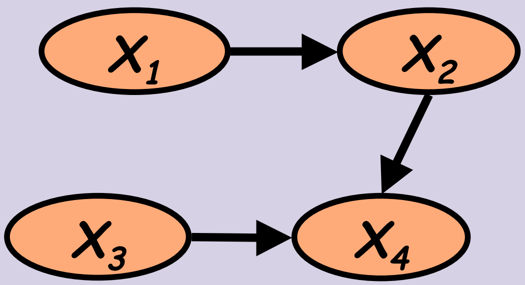
Minimal I-Map Example

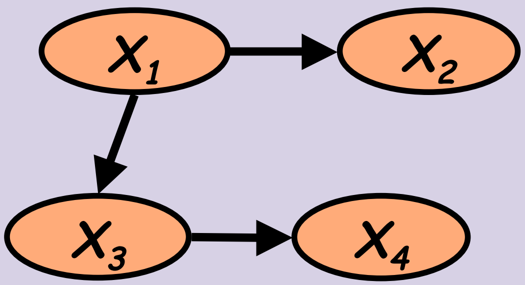
- If  is a minimal I-Map

- Then, these are **not** I-Maps:









Constructing minimal I-Maps

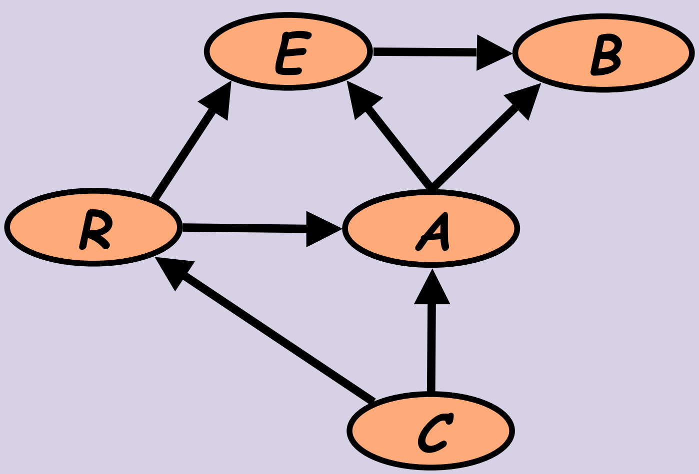
The factorization theorem suggests an algorithm

- Fix an ordering $X_1,\ldots,X_n$
- For each i,
  - select Pa_i to be a minimal subset of $\{X_1,\ldots,X_{i-1}\}$, such that $Ind(X_i\,;\{X_1,\ldots,X_{i-1}\} - Pa_i \mid Pa_i)$
- Clearly, the resulting graph is a minimal I-Map.

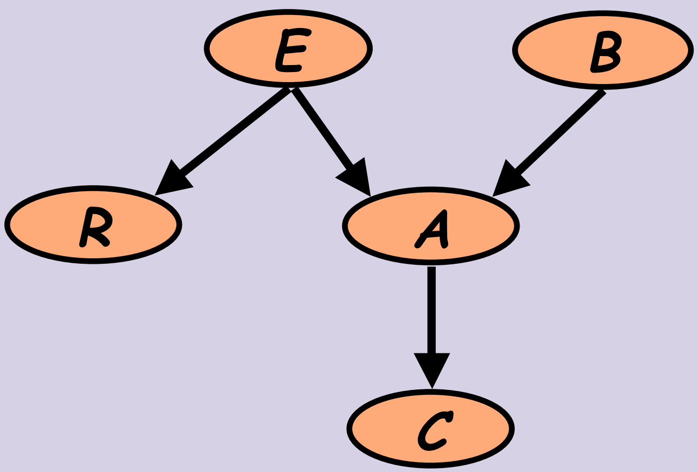
Non-uniqueness of minimal I-Map

- Unfortunately, there may be several minimal I-Maps for the same distribution
  - Applying I-Map construction procedure with different orders can lead to different structures

Order: *C, R, A, E, B*



Original I-Map



P-Maps

- A DAG G is P-Map (**perfect map**) of a distribution P if
  - $Ind(X; Y \mid Z)$ if and only if

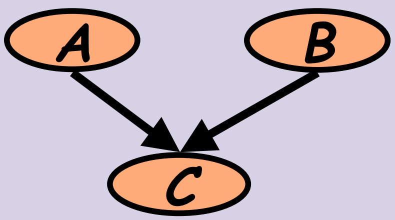
$$d\text{-}sep(X; Y \mid Z, G) = yes$$

Notes:

- A P-Map captures all the independencies in the distribution
- P-Maps are unique, up to DAG equivalence

P-Maps

- Unfortunately, some distributions do not have a P-Map
- Example: $$P(A,B,C) = \begin{cases} 1/12 & \text{if } A \oplus B \oplus C = 0 \\ 1/6 & \text{if } A \oplus B \oplus C = 1 \end{cases}$$
- A minimal I-Map:



- This is not a P-Map since *Ind(A;C)* but *d-sep(A;C) = no*

Bayesian Networks

- A Bayesian network specifies a probability distribution via two components:
  - A DAG G
  - A collection of conditional probability distributions $P(X_i|Pa_i)$
- The joint distribution P is defined by the factorization

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa_i)$$

- Additional requirement: G is a minimal I-Map of P

Summary

- We explored DAGs as a representation of conditional independencies:
  - Markov independencies of a DAG
  - Tight correspondence between *Markov(G)* and the factorization defined by *G*
  - d-separation, a sound & complete procedure for computing the consequences of the independencies
  - Notion of minimal I-Map
  - P-Maps
- This theory is the basis of Bayesian networks