CMPUT463/563
Probabilistic Graphical Models

Representation: Bayesian Network

Lili Mou

Dept. Computing Science, University of Alberta

lmou@ualberta.ca

Consider the environment before printing.
Please print double-sided.

Introduction

Describing a probabilistic distribution

Discrete variable (with finite values)

- One variable: probability table
- Multiple variables: joint probability table
- Consider two variables, each taking value 0 or 1

Joint probability

Row #	X_1	X_2	$P(X_1, X_2)$
1	0	0	π_1
2	0	1	π_2
3	1	0	π_3
4	1	1	$\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$

Introduction

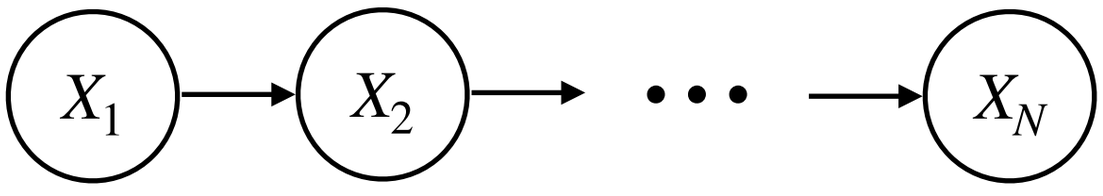
N variables, each taking K values

- No independencies are known
  - $K^N - 1$ free parameters
- All variables are independent
  - $N(K-1)$ free parameters
- What if we know X_i depends on X_{i-1} only for $i = 2, \cdots, N$?

$$p(X_1, \cdots X_N) = p(X_1)p(X_2 \mid X_1) \cdots p(X_n \mid X_n - 1)$$

  - For X_1, we have $K - 1$ parameters
  - For $X_i, i = 2, \cdots, N$, we have $K(K-1)$ parameters

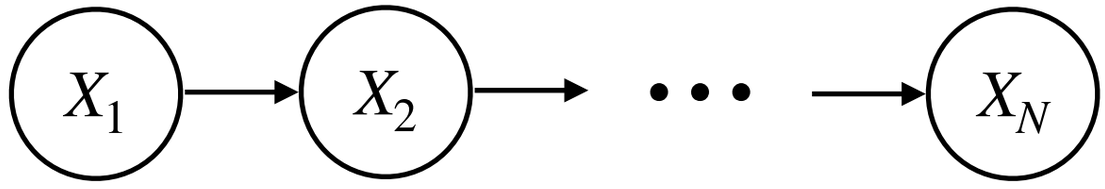
In total, we have $K^2 - 1$ parameters



Key Idea

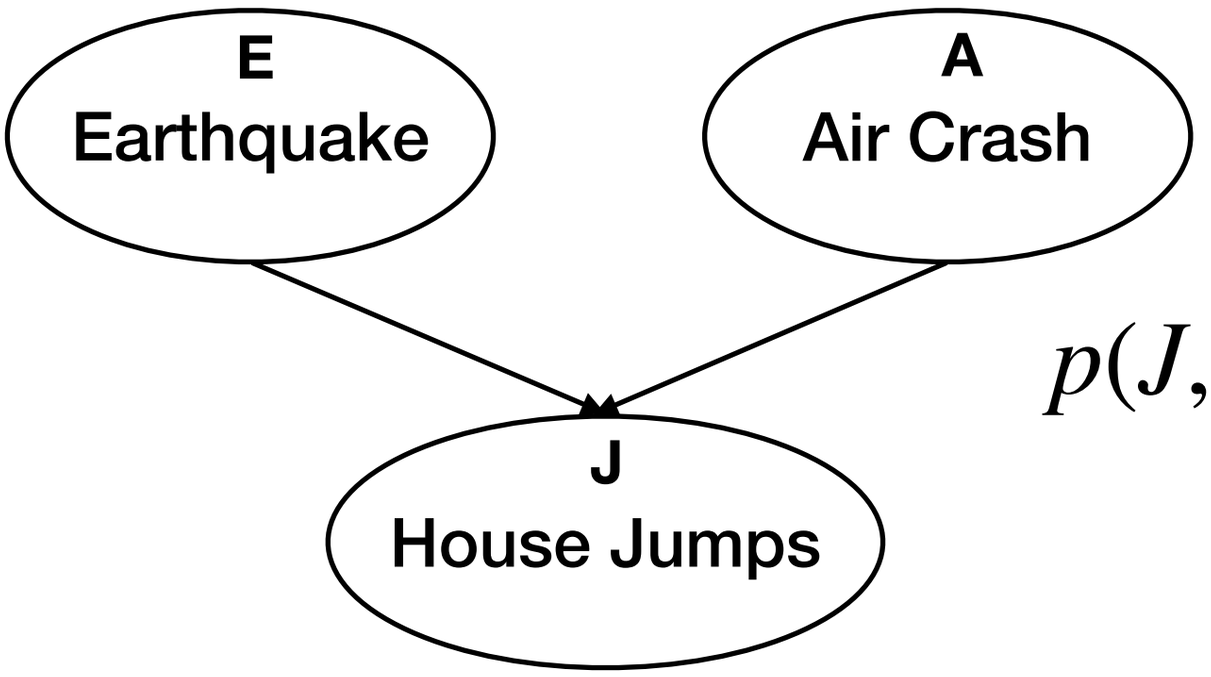
Factorize a joint probability by conditional probabilities

- Chain



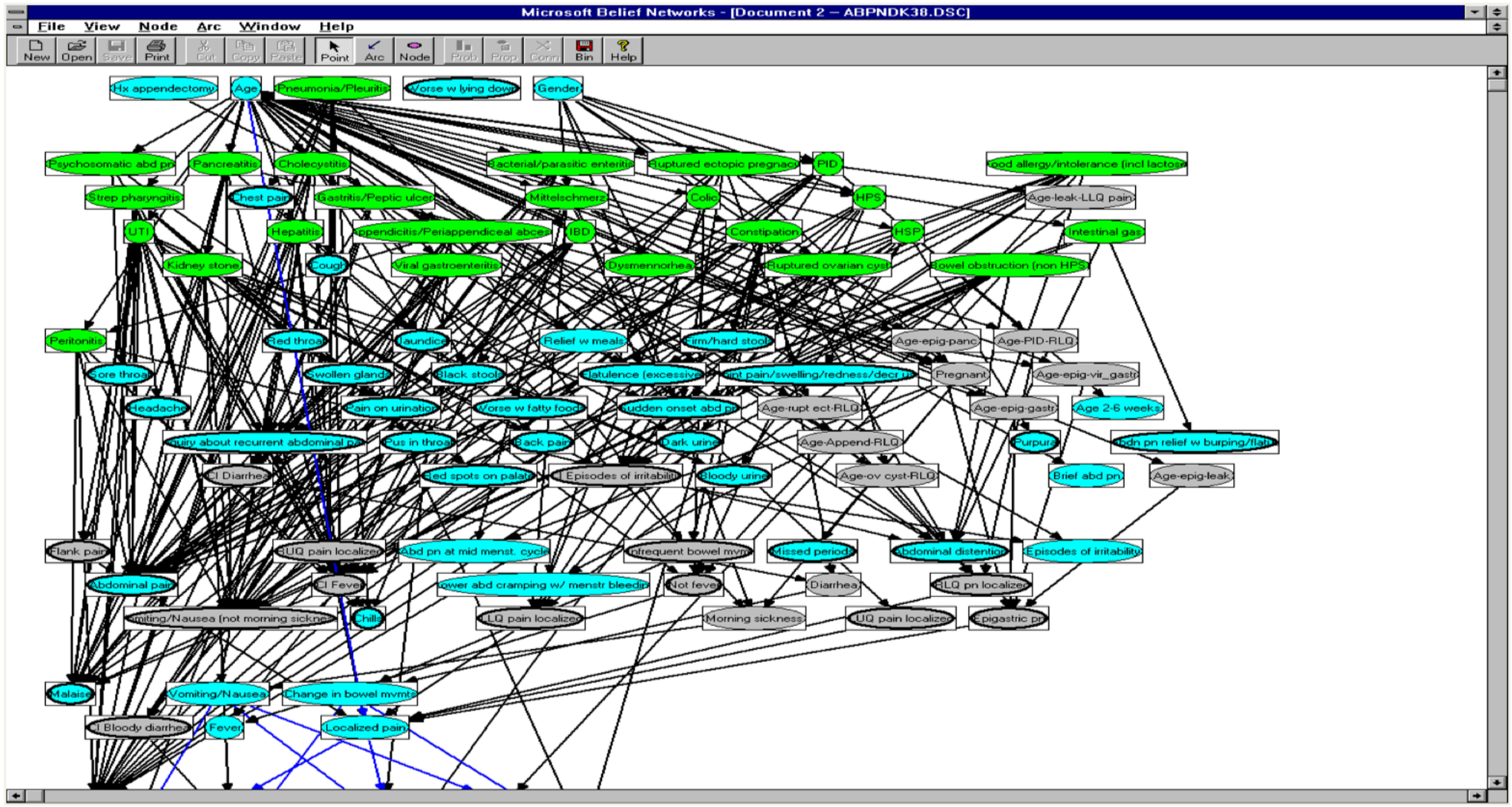
$$p(X_1, \cdots X_n) = p(X_1)p(X_2 \mid X_1) \cdots p(X_n \mid X_n - 1)$$

- Tree



$$p(J, E, A) = p(E)P(A)P(J \mid E, A)$$

Medical Diagnosis (Microsoft)



Thanks to: Eric Horvitz, Microsoft Research

Daphne Koller

Original source

Taken from Coursera PGM course

Bayesian Network (intuition)

- Interested in random variables $X_1, \cdots, X_n$
- Each variable X_i has its parent node(s), roughly speaking "cause(s)," denoted by Par(X_i)
- Model the conditional probability of a variable X_i given its parent(s)
- Joint probability = $\prod$ conditional probabilities

Bayesian Network (definition)

- Let $X_1, \cdots, X_N$ be a set of variables
- A Bayesian network is a **directed acyclic graph** (DAG)

  $G = \langle V, E \rangle$, where

  - $V = \{X_1, \cdots, X_N\}$
  - $E = \{X_i \to X_j : X_i \text{ is a parent of } X_j\}$

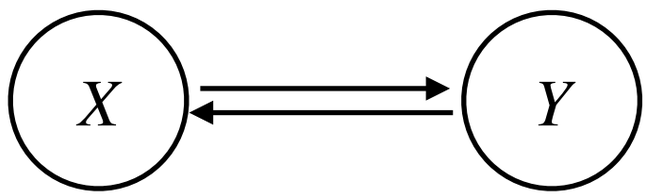
- Joint probability

$$p(X_1, \cdots, X_N) = \prod_{i=1}^{N} p(X_i \,|\, \mathrm{Par}(X_i))$$

Note: If a node X_i has no parent, then $\mathrm{Par}(X_i) = \varnothing$.
$p(X_i \,|\, \mathrm{Par}(X_i))$ is simply $p(X_i)$

BN must be acyclic

- $p(X \mid Y)p(Y \mid X)$ is not $p(X, Y)$ in general

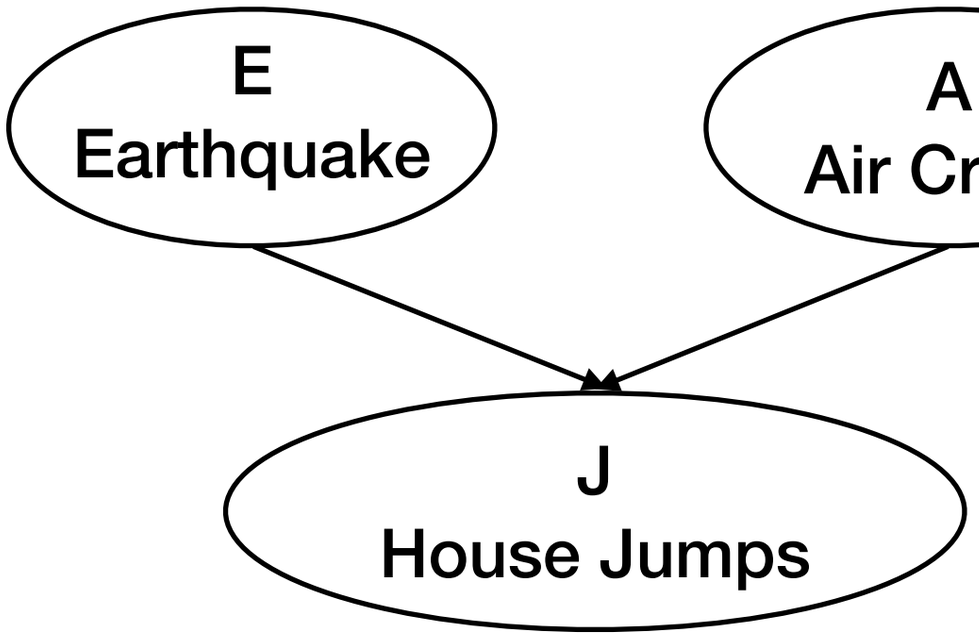


- Other than being acyclic, there are no other restrictions
  - A BN may be connected, disconnected, etc.
  - May be a chain, a tree, or a graph in general

Example

E	P(E)
0	0.8
1	0.2

A	P(A)
0	0.9
1	0.1



E	A	J	P(J|E, A)
0	0	0	0.99
0	0	1	0.01
0	1	0	0.03
0	1	1	0.97
1	0	0	0.05
1	0	1	0.95
1	1	0	0.01
1	1	1	0.99

Finite-value discrete variables may be modeled by conditional probability tables (CPTs)

Example

- P(J=1)=0.0072+0.0776+0.1710+0.0198=0.2756
- P(E=0, A=0|J=1) = 0.0072 / 0.2756 = 2.61%
- P(E=0, A=1|J=1) = 0.0776 / 0.2756 = 28.16%
- P(E=1, A=0|J=1) = 0.1710 / 0.2756 = 62.05%
- P(E=1, A=1|J=1) = 0.0198 / 0.2756 = 7%

E	P(E)
0	0.8
1	0.2

A	P(A)
0	0.9
1	0.1

E	A	J	P(J|E, A)	P(E, A, J)
0	0	0	0.99	0.7128
0	0	1	0.01	0.0072
0	1	0	0.03	0.0024
0	1	1	0.97	0.0776
1	0	0	0.05	0.0090
1	0	1	0.95	0.1710
1	1	0	0.01	0.0002
1	1	1	0.99	0.0198

Dependencies in BN

- Recall the purpose of BN: Model a distribution by
  - Capturing important dependencies
  - Ignoring unimportant dependencies
- **A natural question**
  - Given a BN, what variables are independent? What variables are dependent?

Dependencies in BN

- If $X \rightarrow Y$, then X, Y are dependent*

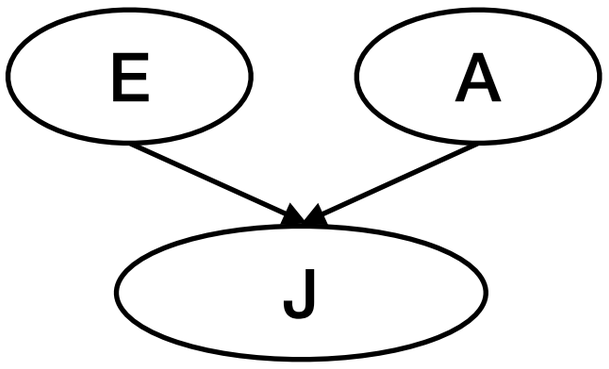
- If $X \rightarrow Z \rightarrow Y$, then X, Y are also dependent* when Z is unknown
- However, X, Y are independent if Z is known, $X \perp Y | Z$
- Example: Coursework → Mark → Grade

- $X \leftarrow Z \rightarrow Y$
- If Z is known, X, Y are independent, i.e., $X \perp Y | Z$
- if Z is unknown, X, Y are dependent*
- Example: Course 1 ← Instructor → Course 2

* Here, “dependency” means the variables may be dependent. They may be independent anyway with certain conditional probability tables

Dependencies in BN

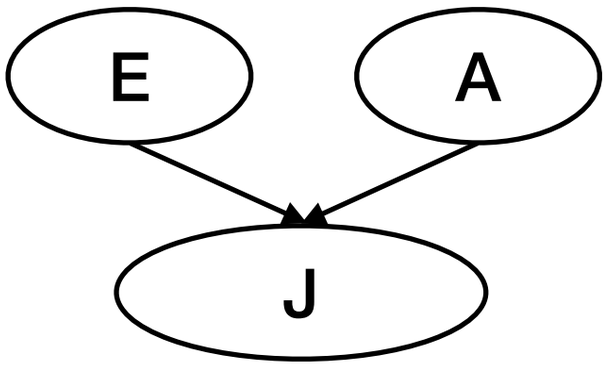
- "v-structure": $X \rightarrow Y \leftarrow Z$
- If Z is unknown, X, Y are independent, $X \perp Y$
- if Z is known, X, Y are dependent*



* Here, "dependency" means the variables may be dependent. They may be independent anyway with certain conditional probability tables

Example

- P(E=1 | A=1, J=1) = 0.0198 / (0.0198 + 0.0776) = 20.3%
- P(E=1 | A=0, J=1) = 0.1710 / (0.1710 + 0.0072) = 95.96%
- P(E=1 | J=1) = (0.1710+0.0198) /
  (0.0072+0.0776+0.1710+0.0198) = 69.23%



E	P(E)
0	0.8
1	0.2

A	P(A)
0	0.9
1	0.1

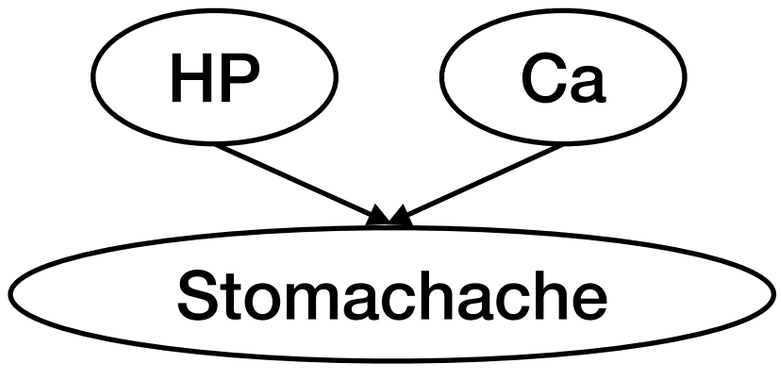
E	A	J	P(J|E, A)	P(E, A, J)
0	0	0	0.99	0.7128
0	0	1	0.01	0.0072
0	1	0	0.03	0.0024
0	**1**	**1**	**0.97**	**0.0776**
1	0	0	0.05	0.0090
1	0	1	0.95	0.1710
1	1	0	0.01	0.0002
1	**1**	**1**	**0.99**	**0.0198**

Example

- P(E=1 | A=1, J=1) = 0.0198 / (0.0198 + 0.0776) = 20.3%
- P(E=1 | A=0, J=1) = 0.1710 / (0.1710 + 0.0072) = 95.96%
- P(E=1 | J=1) = (0.1710+0.0198) /
  (0.0072+0.0776+0.1710+0.0198) = 69.23%
- **Intuition:** House jumping is an unlikely event. If it happens, we need to find a cause.
  - Either cause is also an unlikely event, but given house jumps, there should be some cause
  - If there is no air crash, then it's very likely that there's an earthquake.
  - If there is an air crash, then the chance of earthquake is lowered, because it's extremely unlikely that both air crash and earthquake happen
- This is known as **explaining away**

Another Example of Explaining away

- I had a stomachache
  - It might be either HP (infection of some bacteria) or cancer, but I was uncertain, so I felt concerned
  - I did a lab test, and it was HP. It was a relief, because HP is benign and can be treated by antibiotics
- PGM perspective
  - HP and Ca are (roughly speaking) independent events
  - HP test does not rule out the possibility of Ca
  - Given stomachache, HP explains Ca away. I.e., given stomachache and HP, the chance of cancer is much lowered.

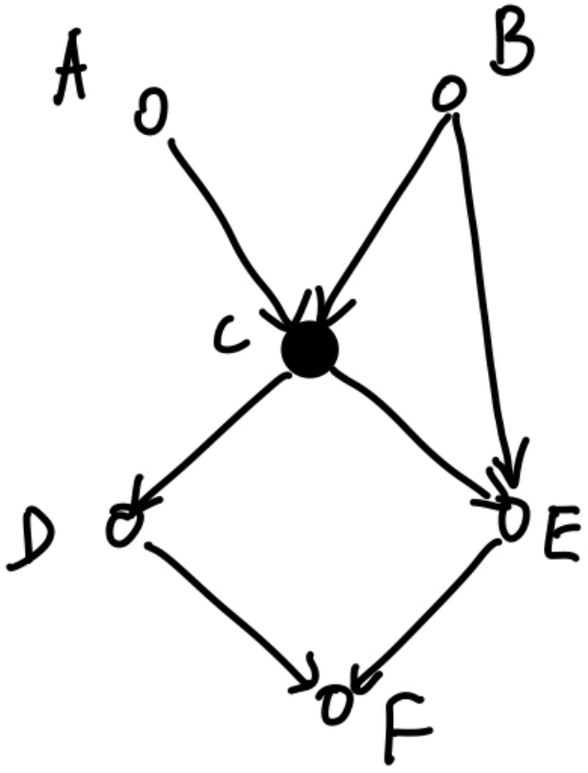


Jargons

- **Def (active trail)**: $X_1 - X_2 - \cdots - X_n$ is an active trail if, for every node X_i with $i = 2, \cdots, n-1$,
  - X_i **or its descendent** is given if X_i has a v-structure $X_{i-1} \to X_i \leftarrow X_{i+1}$.
  - X_i is not given otherwise.

- **Def (d-separation).** X and Y are d-separated given observations $\boldsymbol{Z}$ in graph G if there is no active trail between X and Y, denoted by d-sep$_G(X, Y \mid \boldsymbol{Z})$

Example



- Blank node: unobserved (not given)
- Black node: observed (given)

d-sep(A, E|C) ?

A — C — E: inactive

A — C — B — E: active

No, d-sep(A, E|C) does not hold

d-sep(D, E|C) ?

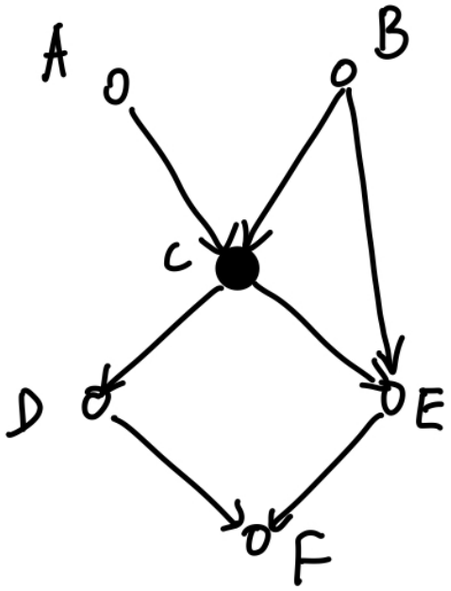
D-C-E, D-C-B-E, D-F-E: none is active

d-sep(D, E|C) holds

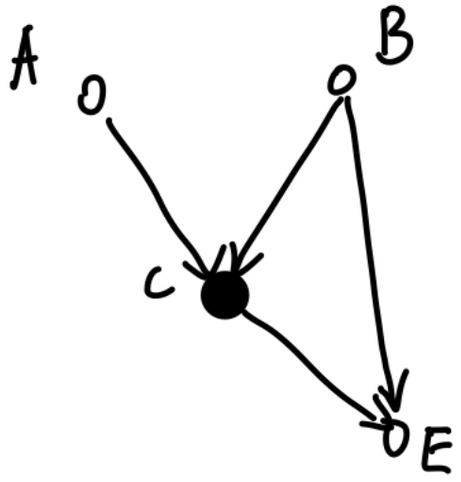
Getting rid of v-structures

- V-structures are sometimes confusing. We may determine d-sep without v-structures
- For a query about d-sep$(X, Y \mid \boldsymbol{Z})$
  - Draw variables mentioned and their ancestors
  - Moralize of the observed nodes, i.e., marry parents (if multiple parents, draw an edge pairwise)
  - Delete given variables and their edges
  - Check if the variables in question are connected

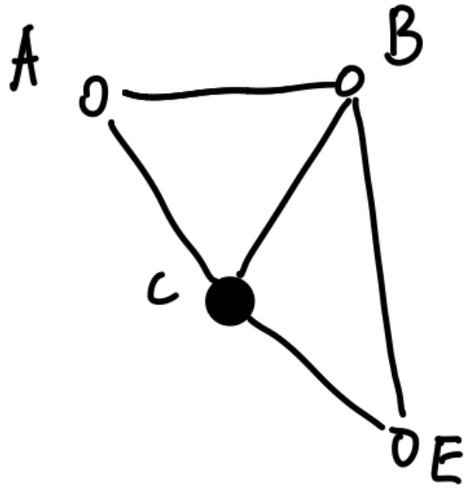
Question: d-sep(A, E | C)?



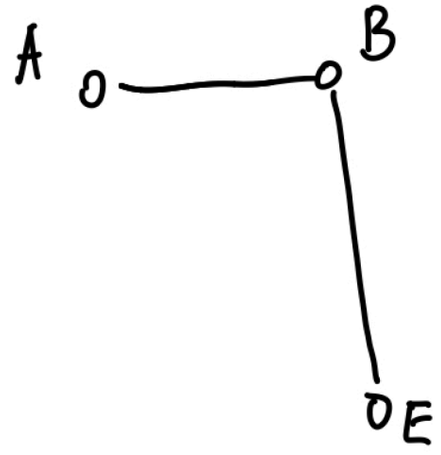
Original graph



Ancestral graph



Moralizing



Soundness

- D-separations are said in terms of graphs
- Independencies are said in terms of probability
- They are very related

Theorem (soundness). Suppose a probability distribution P factorizes according to a Bayesian network G. If d-sep$_G(X, Y | \boldsymbol{Z})$, then $X \perp Y | \boldsymbol{Z}$.

(Soundness means that d-sep captures “correct” independencies)

But converse $X \perp Y | \boldsymbol{Z}$ => d-sep$_G(X, Y | \boldsymbol{Z})$ is **false**.

Completeness

Theorem (completeness). If not d-sep$_G(X, Y \mid \boldsymbol{Z})$, then there exists some (but almost all) distribution P that factorizes according to G where $X \perp Y \mid \boldsymbol{Z}$ does not hold.

Completeness means d-sep also captures all independencies.

However, this does not assert for every distribution, but **some** (although almost all) distribution, that factorizes according to G.

Proof sketch: Check a local structure + Induction

I-map

$I(G)$: independencies captured by G

$$I(G) = \{X \perp Y | \boldsymbol{Z} : \text{d-sep}(\text{X}, \text{Y} | \boldsymbol{Z})\}$$

$I(P)$: independencies in the distribution P

Def (I-map). G is an I-map of P if

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

G is appropriate to model P

G cannot assert more independencies than P, but can drop independencies in P

I-map

$I(G)$: independencies captured by G

$$I(G) = \{X \perp Y | \boldsymbol{Z} : \text{d-sep}(\text{X}, \text{Y} | \boldsymbol{Z})\}$$

$I(P)$: independencies in the distribution P

Def (I-map). G is an I-map of P if

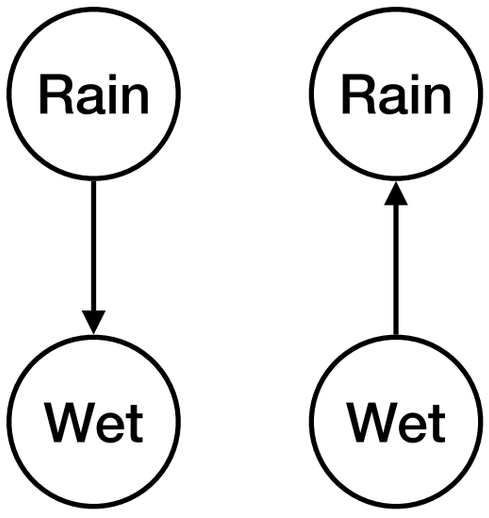
$$I(G) \quad \subseteq \quad I(P) \quad ?$$

G is appropriate to model P

Cause and Effect in BN

- BN models cause and effect in a loose sense.
- The physical cause and effect is not modeled by conditional probabilities

- Example:
  - Since it rained yesterday, the road is wet
  - Since the road is wet, it must have rained yesterday

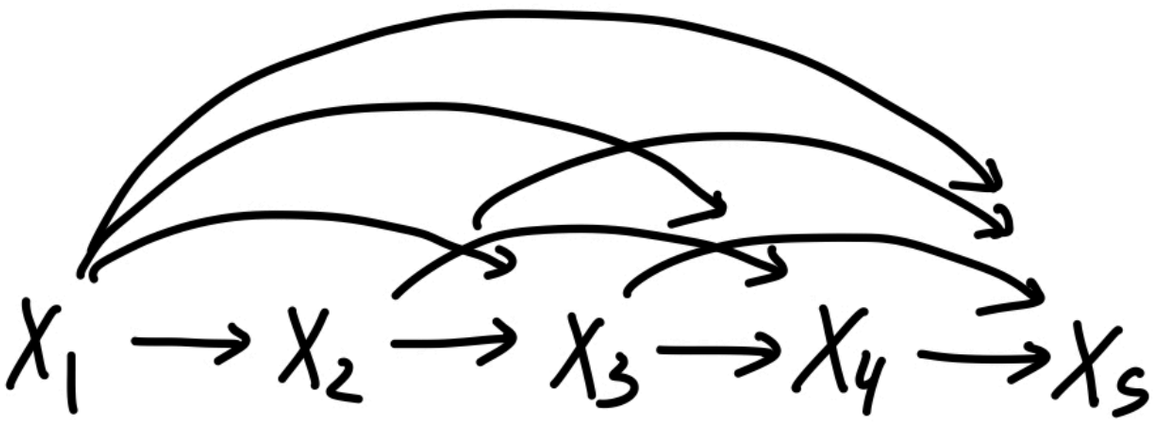


$$P(R, W) = P(R)P(W \mid R) = P(W)P(R \mid W)$$

Both are valid BNs and can model any joint distribution on R and W

BN can model any discrete finite-value distributions

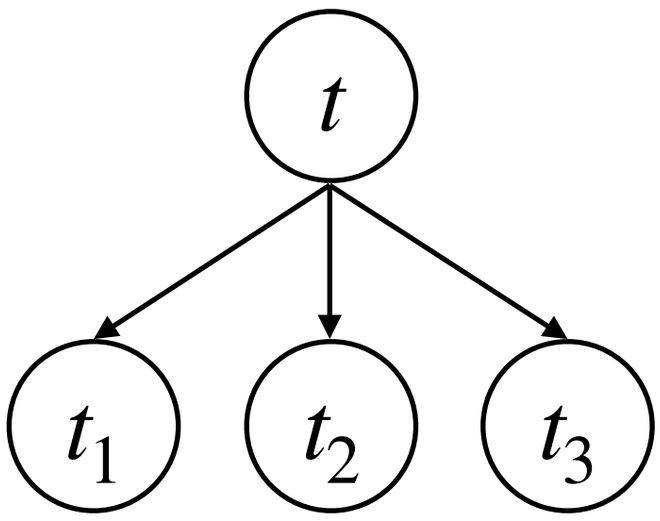
- $P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2 \mid X_1)\cdots P(X_n \mid X_1, \cdots, X_{n-1})$
- Tabular parametrization of CDTs
  - Inefficient (how many parameters)?
- Function parametrization (e.g., by neural networks)
  - Actually outperforms models with dependency assumptions
  - Application: text generation



Cause and Effect in BN

- Designing BN by cause and effect simplifies the structure

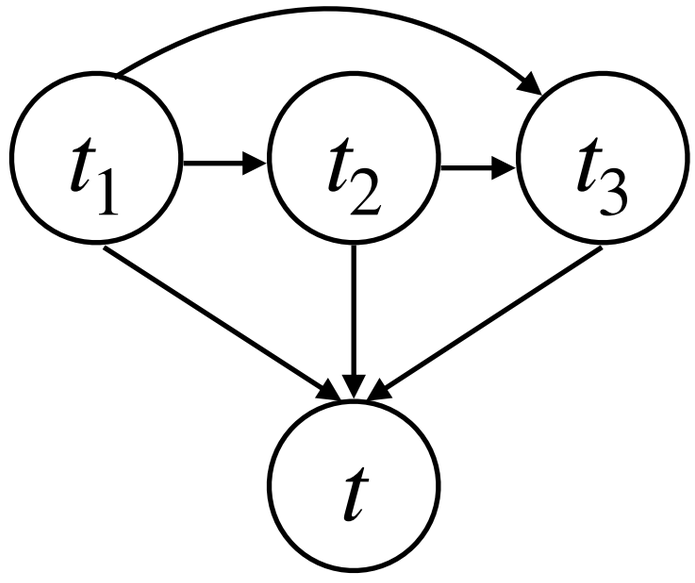
- t: standard time, t_i: time by clock i
  - An intuitive BN: every clock is adjusted by the standard time
  - But what if we reverse $t \rightarrow t_i$
    - Requirement: Retain all dependencies



Cause and Effect in BN

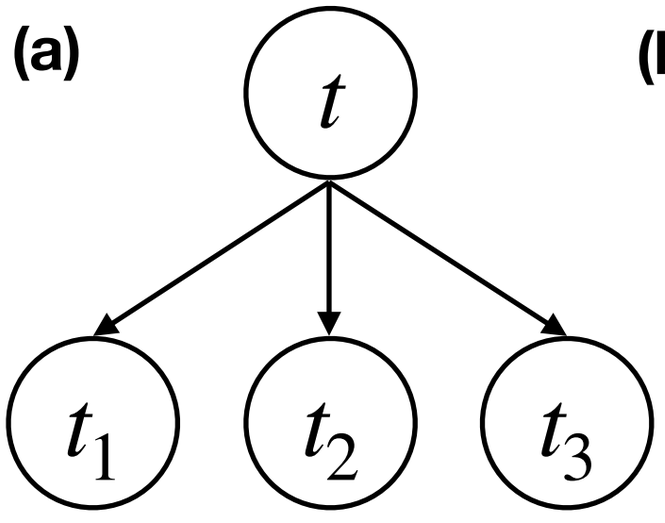
- Designing BN by cause and effect simplifies the structure
- t: standard time, t_i: time by clock i
  - In (a), we know $t_1 \perp t_2$ does not hold in general, but (b) loses such dependency. To recover that, we need to connect t_i, t_j for every i, j

Retaining all dependencies, but in efficient parameterization

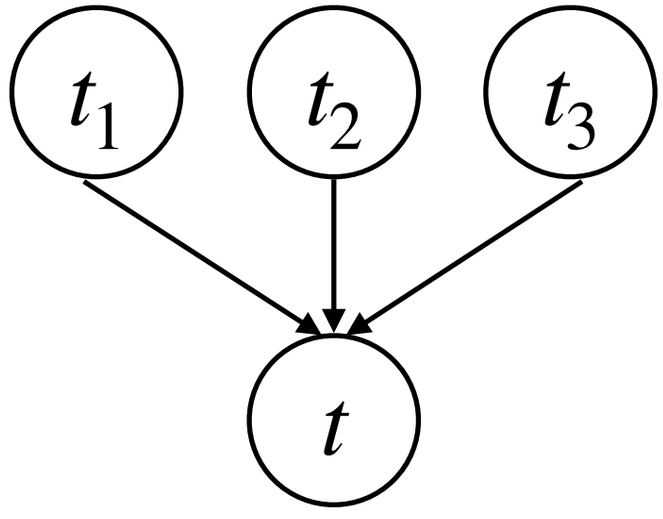


Dependencies lost

(a)

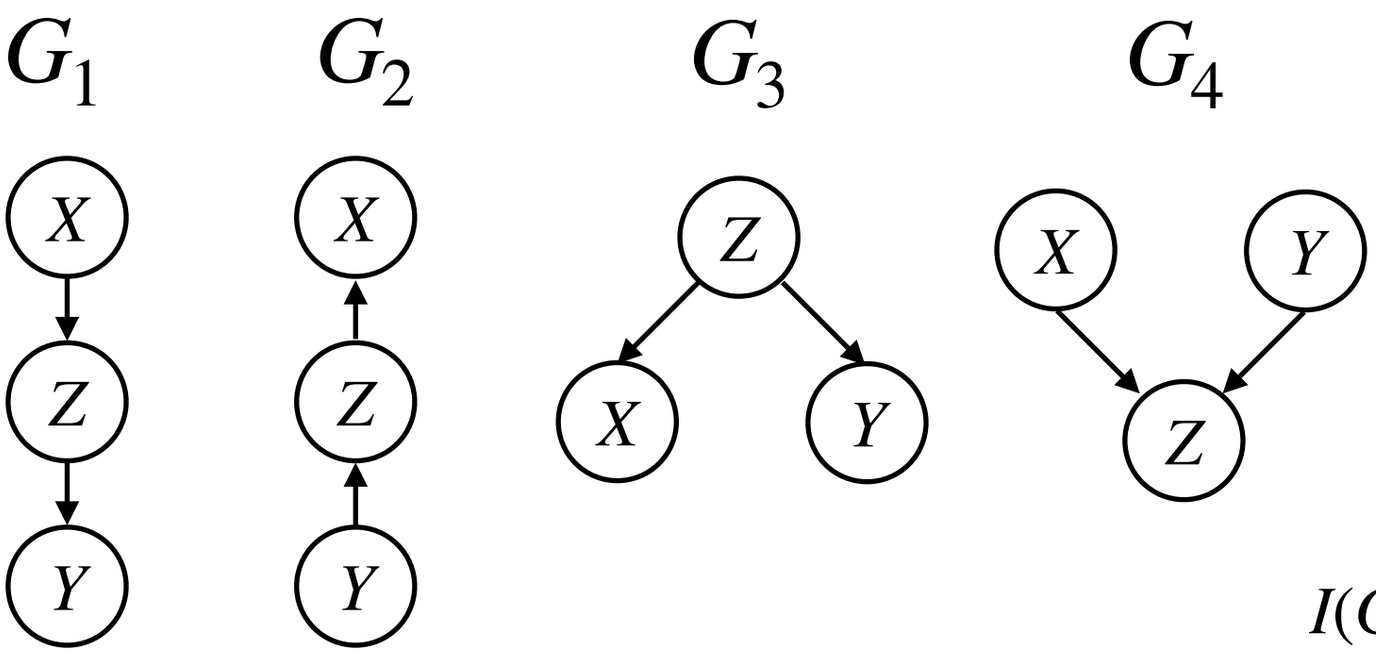


(b)



I-equivalence

- Different BNs may capture exactly the same independencies



$$I(G_4) = \{X \perp Y\}$$

$$\not\perp$$

$$I(G_1) = I(G_2) = I(G_3) = \{X \perp Y | Z\}$$

V-structure has different independencies than other structures
If not v-structure, then it doesn't matter

I-equivalence

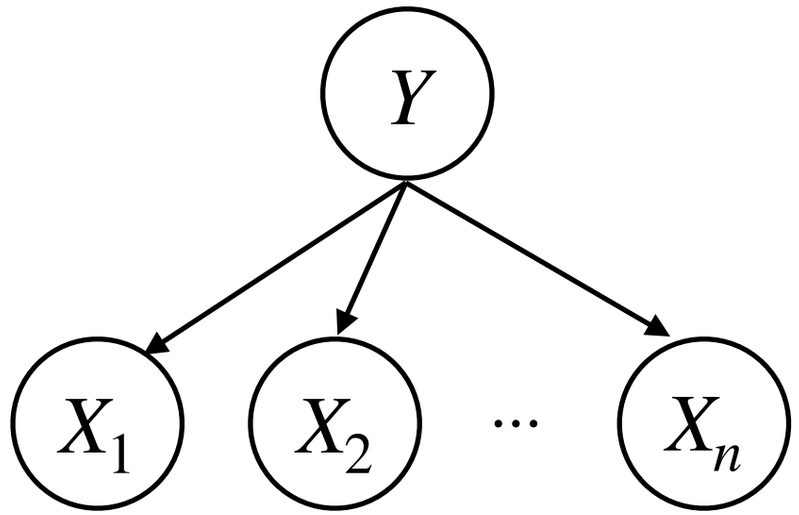
- Def (I-equivalence): If $I(G_1) = I(G_2)$, then G_1 and G_2 are I-equivalent.

- Theorem. G_1 and G_2 are I-equivalent if
  - They have the same skeleton (undirected graphs are the same)

AND

  - They have the same v-structures

Example: Naïve Bayes



- Features $X_1, \cdots, X_n$
- Label Y
- Generation story
  - First, a label is generated
  - Then, every feature is generated independently given the category Y
- Application: Text classification
  - Y: spam VS not spam
  - X_i: whether ith word occurs or not

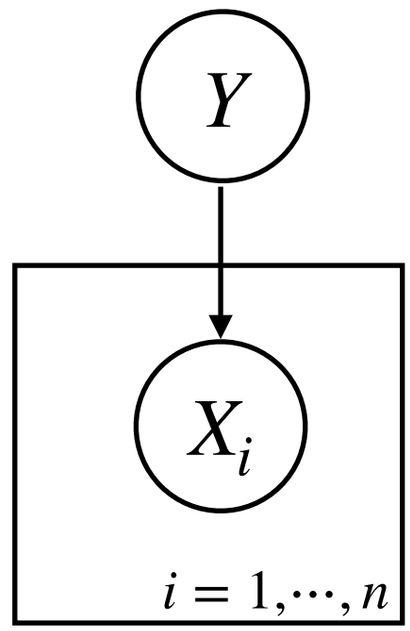


Plate diagram: a rectangle means repeatedly drawing variables

Example: Markov Model

- Temporal data $X_1, \cdots, X_n$ **Parametrization**
- Generation story
  - X_1 is generated **Initial probability π**
  - For every time step $t \geq 2$, X_t is generated based on X_{t-1} **Transition probability T**
- Applications:
  - X_i: weather on Day i
  - X_i: location of a moving object

T is shared among different time steps

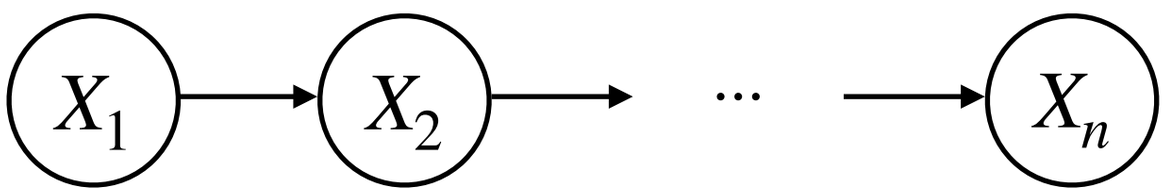


Plate diagrams are confusing for temporal graphs. Expand it.

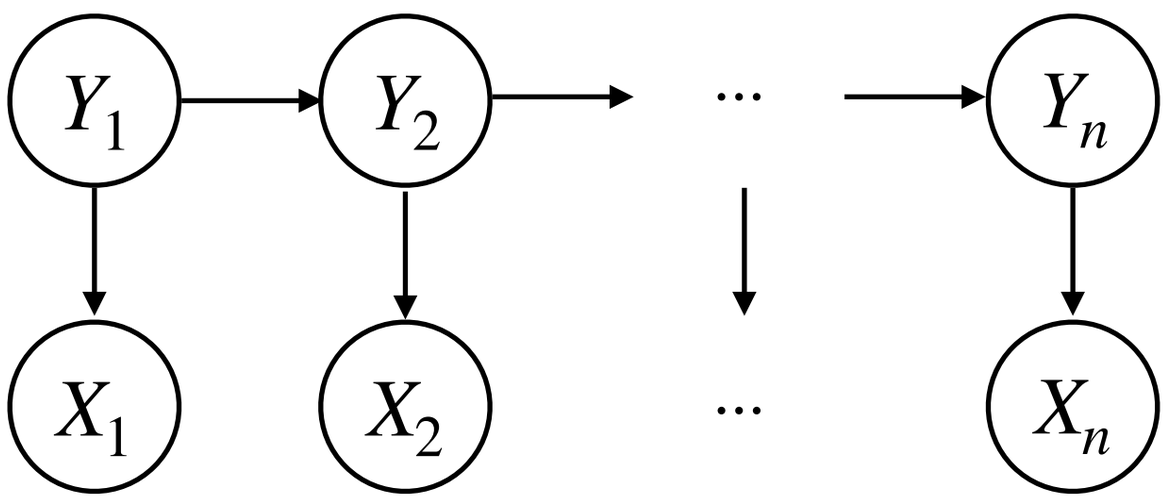
Example: Hidden Markov Model

- Still temporal data, where $Y_1, \cdots, Y_n$ follow a Markov model
- Generation story

Parametrization

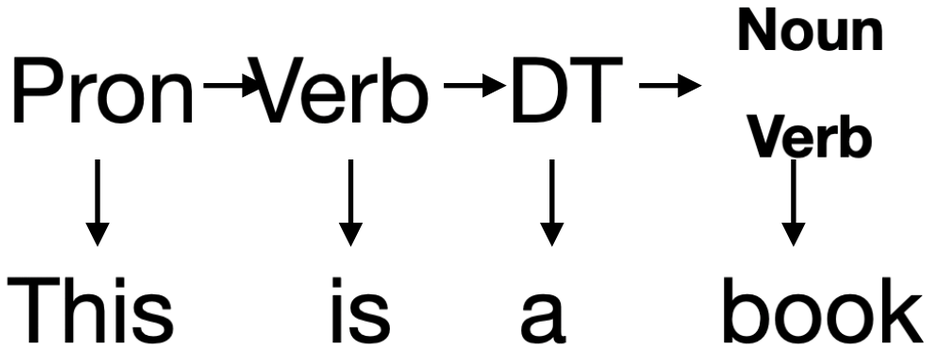
  - Y_1 is generated **Initial probability π**
  - At every step $t \geq 2$
    - Y_t is generated based on Y_{t-1} **Transition probability T**
    - X_t is generated based on Y_t **Emission probability E**

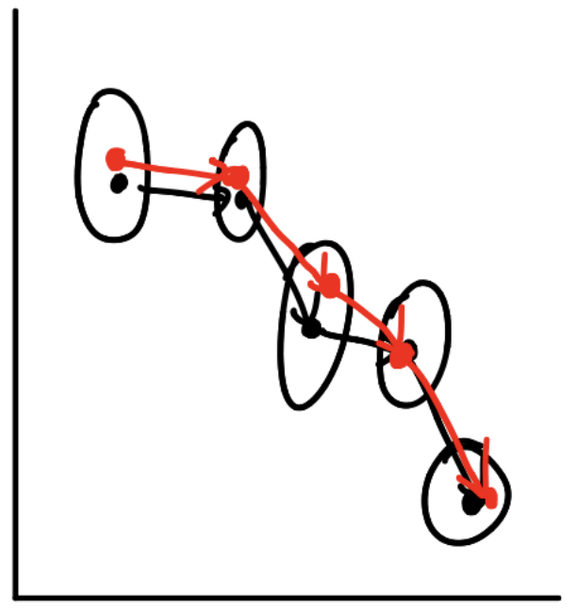
Shared among different time steps

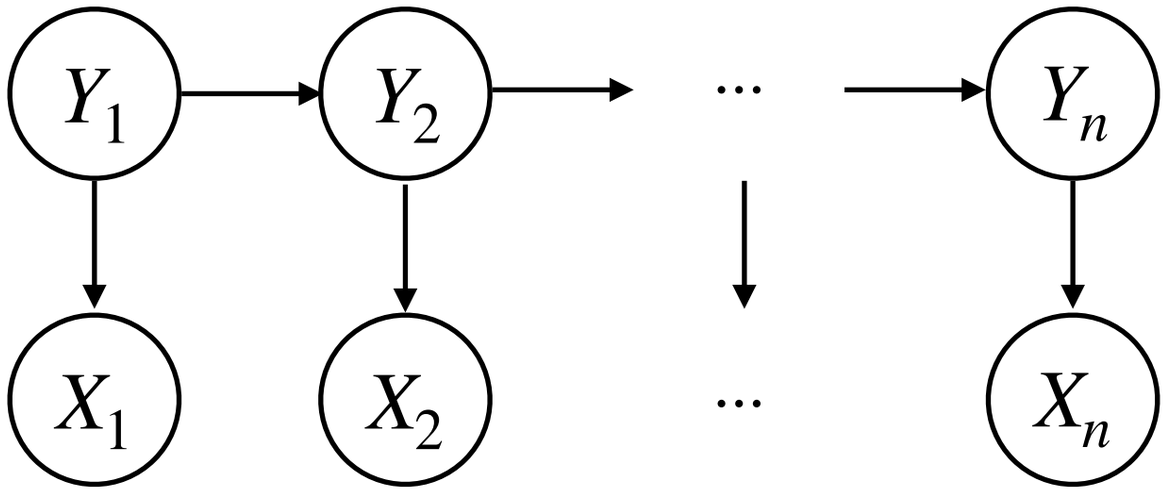


Example: Hidden Markov Model

- Application
  - POS tagging
    - Y_i: POS tag, X_i: word
  - State space representation
    - Y_i: actual position of an object
    - X_i: Position read from radar







Learning the Parameters

- Consider a BN with tabular CDTs
- Training data $\{X_1^{(m)}, \cdots X_n^{(m)}\}_{m=1}^{M}$
- Then, learning BN parameters is simply counting
- Example

$$\hat{P}(X = x \mid Y = y) = \frac{\sum_{m=1}^{M} \mathrm{I}(X^{(m)} = x, Y^{(m)} = y)}{\sum_{m=1}^{M} \mathrm{I}(Y^{(m)} = y)}$$

I(*) is an indicator function. I(*)=1 if * is true, I(*)=0 otherwise.

- Some estimated probability may be 0
  - Add-k smoothing: add k counts for every configuration. k may be a real number, e.g., 0.1.

Summary

- BN factorizes a joint distribution by conditional distributions
- Dependencies and independencies
  - v-structure VS other structures
  - Active trail ⇒ (may be) dependent
  - No active trail ⇒ independent
- D-sep, I-map, I-equivalence
- BN can model any distribution
- Designing BN in an intuitive way simplifies parametrization