Independence

Two variables are **independent** if: $\forall x, y \; P(x, y) = P(x)P(y)$

We denote this as $X \perp\!\!\!\perp Y$

Conditional Independence

X is **conditionally independent** of Y given Z

if and only if: $\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$

or, equivalently, if and only if $\forall x, y, z : P(x|z, y) = P(x|z)$

$$X \perp\!\!\!\perp Y | Z$$

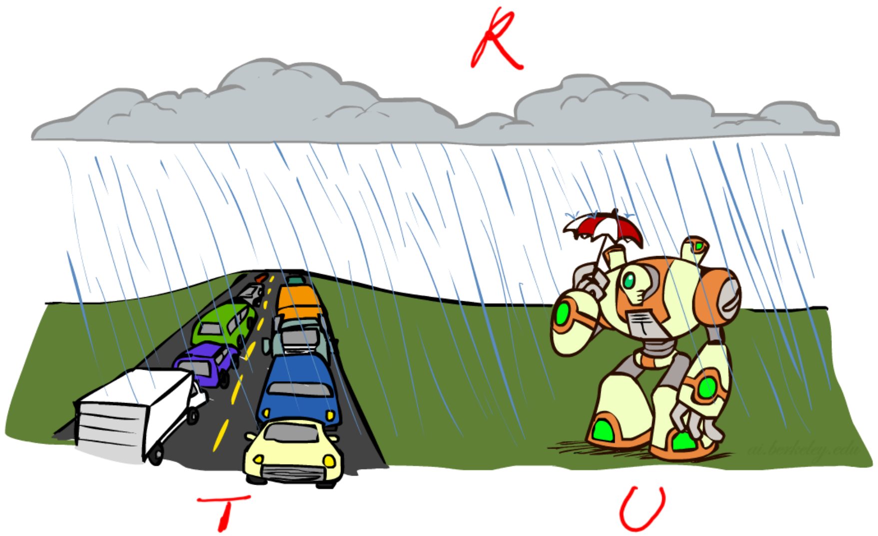
Conditional Independence

Traffic, Umbrella, Raining

$X \perp\!\!\!\perp Y \mid Z$

↑

Raining



$T \perp\!\!\!\perp U$?

$T \perp\!\!\!\perp U \mid R$

⇕

$P(T \mid R, U) = P(T \mid R)$

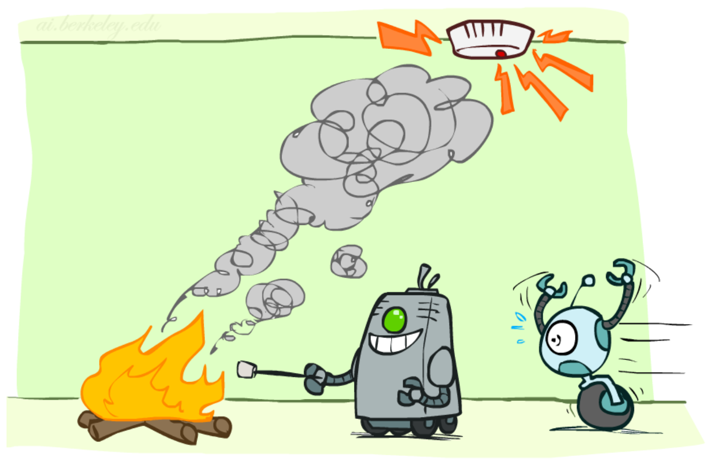
Conditional Independence

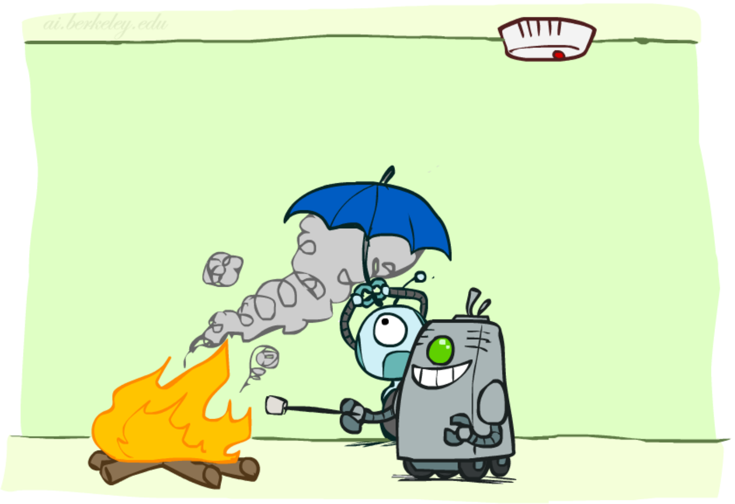
(Smoke detector)

Fire, Smoke, Alarm

$X \perp\!\!\!\perp Y \mid Z$

$P(\text{Alarm} \mid \text{smoke}) \stackrel{?}{=} P(\text{Alarm} \mid \text{smoke}, \text{Fire})$

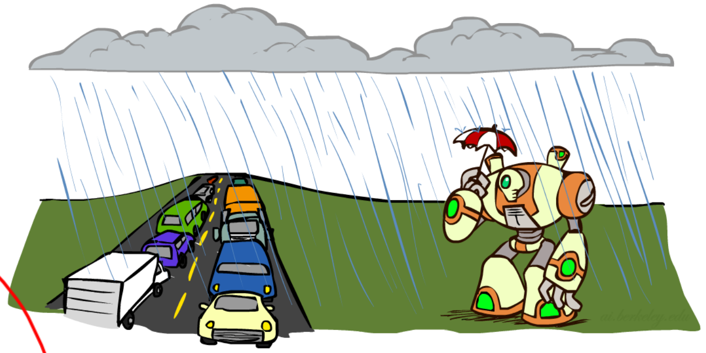


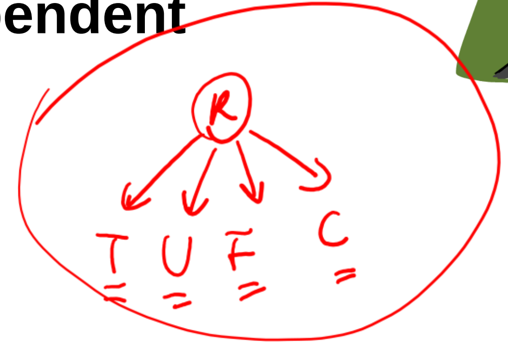


Independence vs. Conditional Independence

Rain
Traffic
Pedestrian holding umbrella
Flood in the house
Trip cancelled
…

Dependent





P(Traffic | Rain, Umbrella) = P(Traffic | Rain)   **Conditional Independent**

Conditional distribution / independence allows us to model the probability of a certain event only using relevant factors.

Bayesian Networks

Bayes Net

Bayesian Network Example

Traffic, Umbrella, Raining

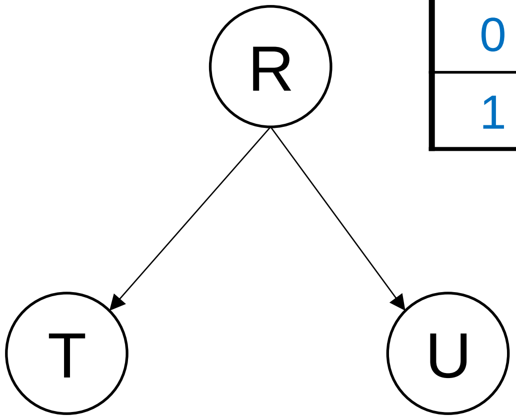
P(t, u, r)

= P(r) P(t | r) P(u | r, t) (always hold by chain rule)

= P(r) P(t | r) P(u | r)

$T \perp\!\!\!\perp U \mid R$

R	P(R)
0	0.7
1	0.3



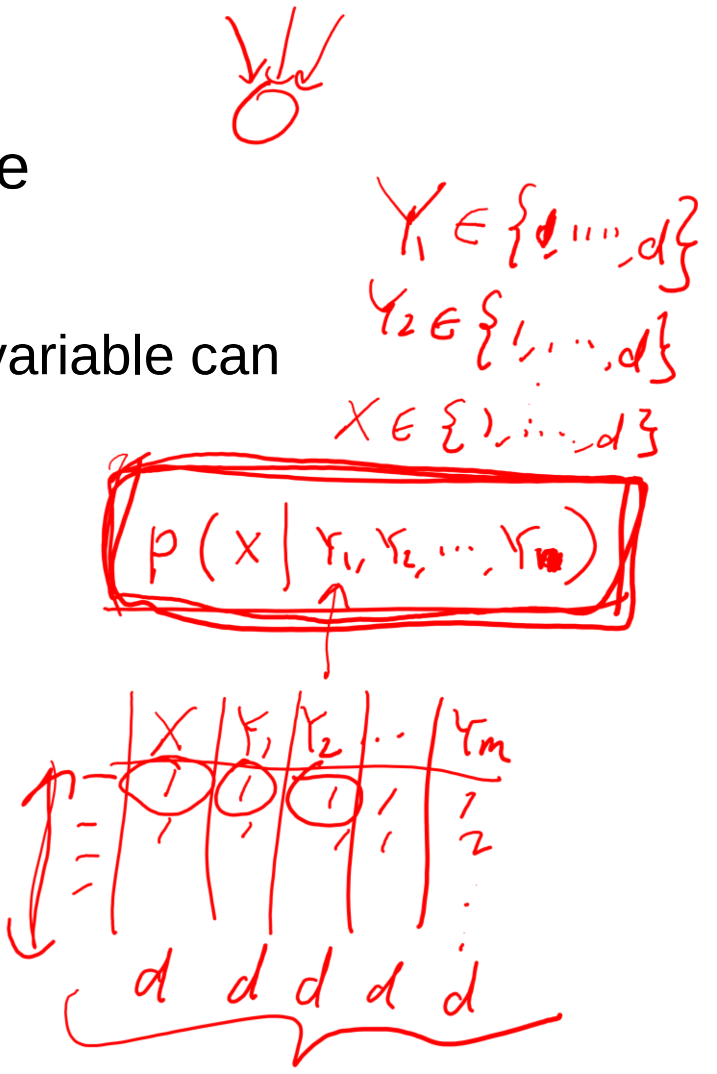
R	T	P(T|R)
0	0	0.5
0	1	0.5
1	0	0.2
1	1	0.8

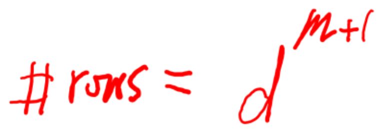
R	U	P(U|R)
0	0	0.8
0	1	0.2
1	0	0.1
1	1	0.9

Bayesian Network (BN)

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - Suppose a node as m parents, and suppose each random variable can take d different values
  - What is the size of the table?
- The BN models the joint probability as

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$





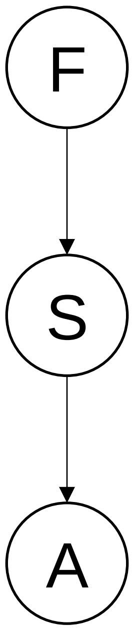
Bayesian Network Example

Fire, Smoke, Alarm

$P(f, s, a) = P(f)\, P(s \mid f)\, P(a \mid s)$ (by BN semantics)

Prove $F \perp\!\!\!\perp A \mid S$?

$p(f)\, p(s \mid f)\, p(a \mid s, f)$



Bayesian Network Example

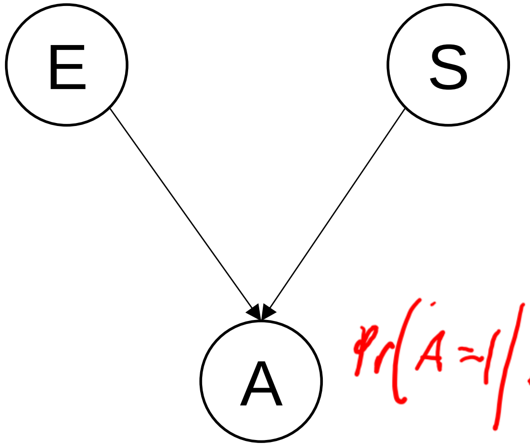
Earthquake, Smoke, Alarm

$P(e, s, a) = P(e)\,P(s)\,P(a \mid e, s)$

$E \perp\!\!\!\perp S$? Yes $\qquad E \perp\!\!\!\perp S \mid A$? No

E	P(E)
0	0.999
1	0.001

S	P(S)
0	0.999
1	0.001



$Pr(A=1 \mid E, S) = \begin{cases} 1, & \text{if } E=1 \text{ or } S=1 \\ 0, & \text{otherwise} \end{cases}$

Pr(Earthquake | Alarm) ? Pr(Earthquake | Alarm, Smoke)

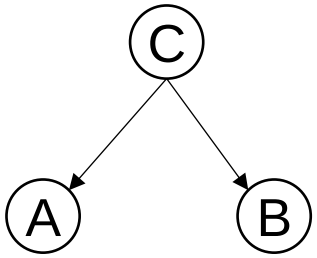
$Pr(\text{Earthquake} \mid \text{Alarm}) \approx 1/2 \qquad Pr(\text{Earthquake} \mid \text{Alarm, Smoke}) \approx 0.001$

"Explain away"

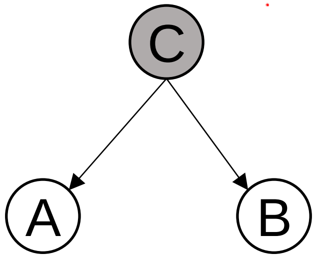
10^{-6}

Recap

- Common cause



$A \not\perp\!\!\!\perp B$

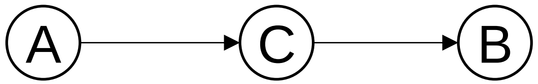


$A \perp\!\!\!\perp B \mid C$

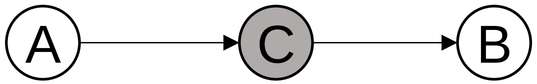
A and B are not independent *in general*

They could still be independent *in special cases*

- Causal chain

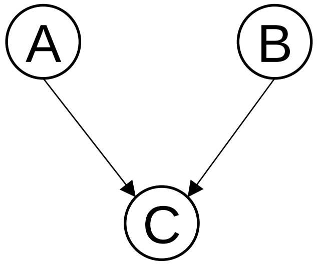


$A \not\perp\!\!\!\perp B$

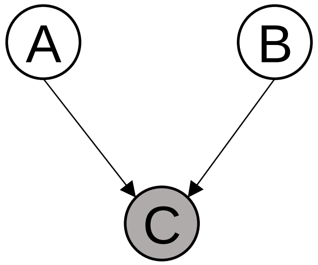


$A \perp\!\!\!\perp B \mid C$

- Common effect

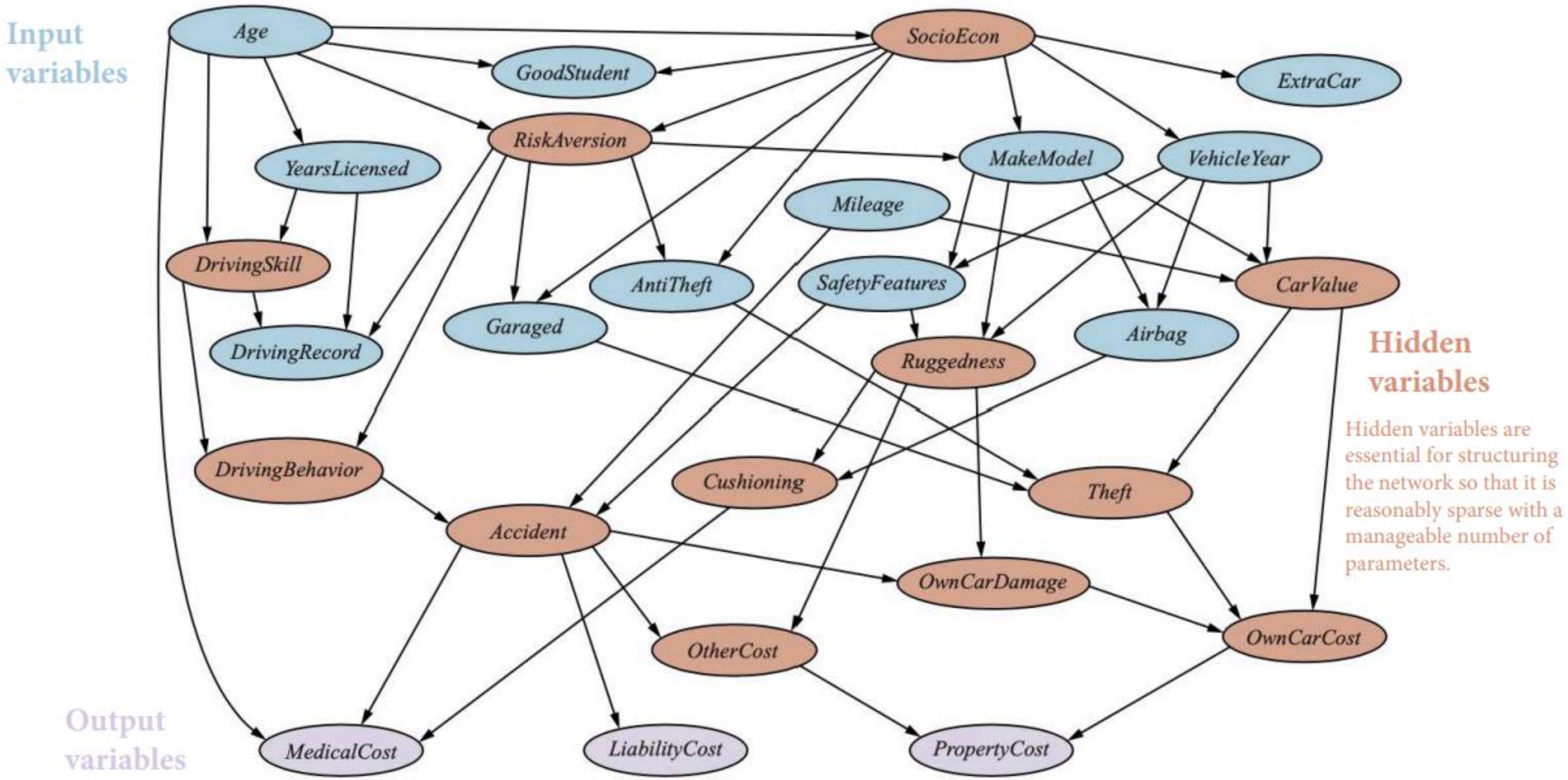


$A \perp\!\!\!\perp B$

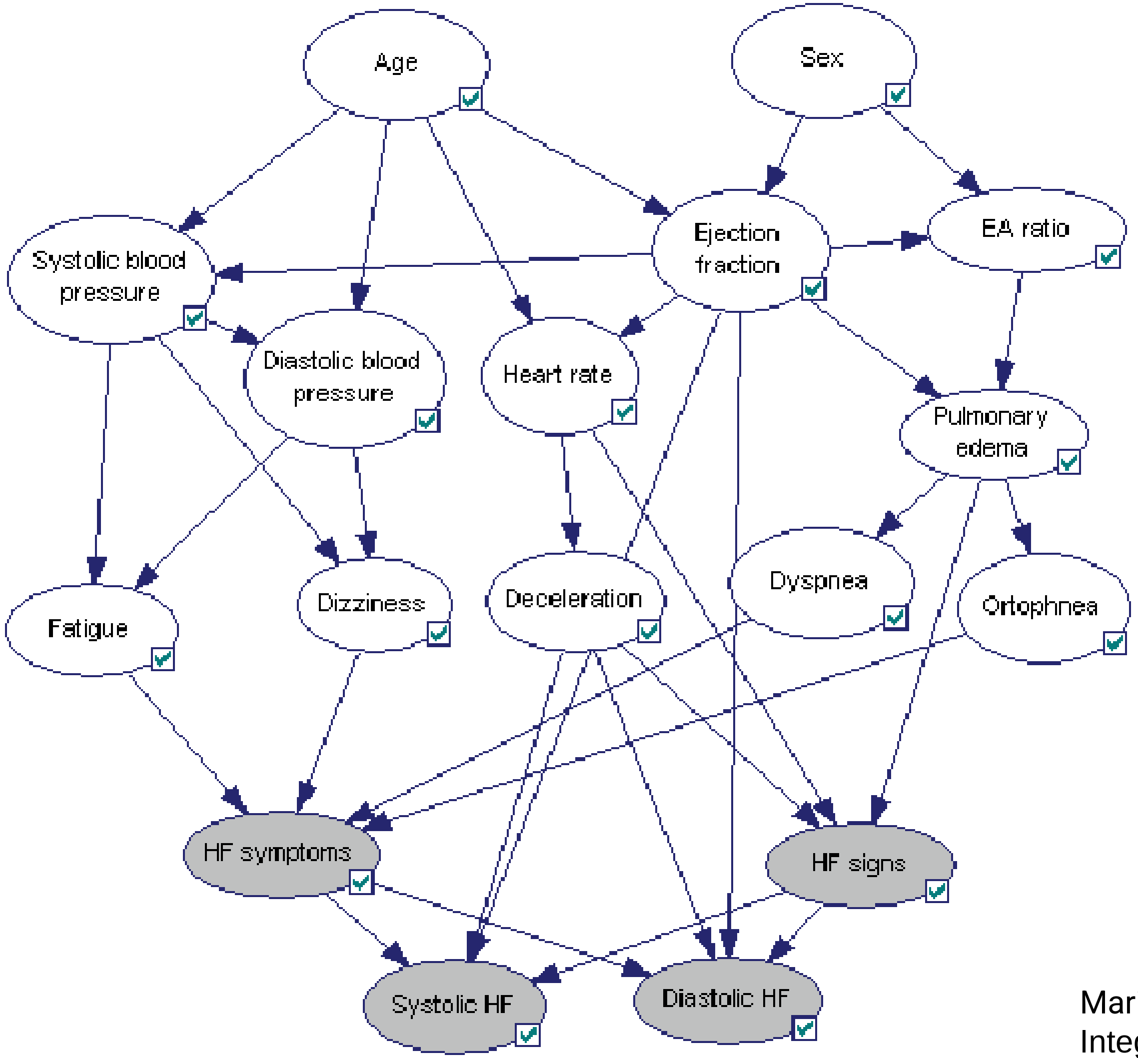


$A \not\perp\!\!\!\perp B \mid C$

Example: Car Insurance



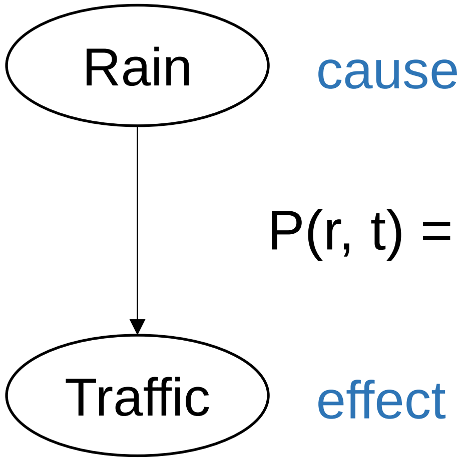
Example: Medical Diagnosis



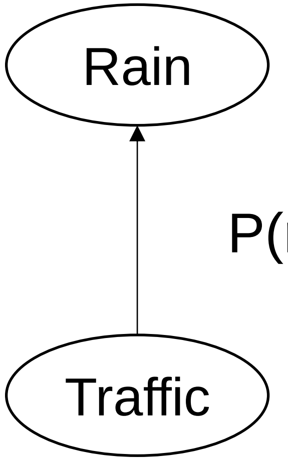
Marin Prcela et al. Information Gain of Structured Medical Diagnostic Tests - Integration of Bayesian Networks and Ontologies

Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents) and easier to think about
- BNs need not be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - Arrows that reflect correlation, but not necessary causality

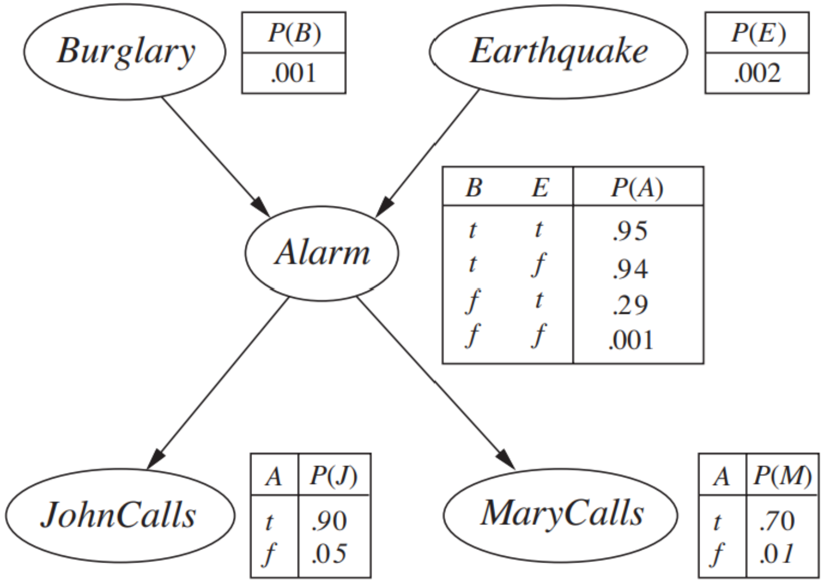


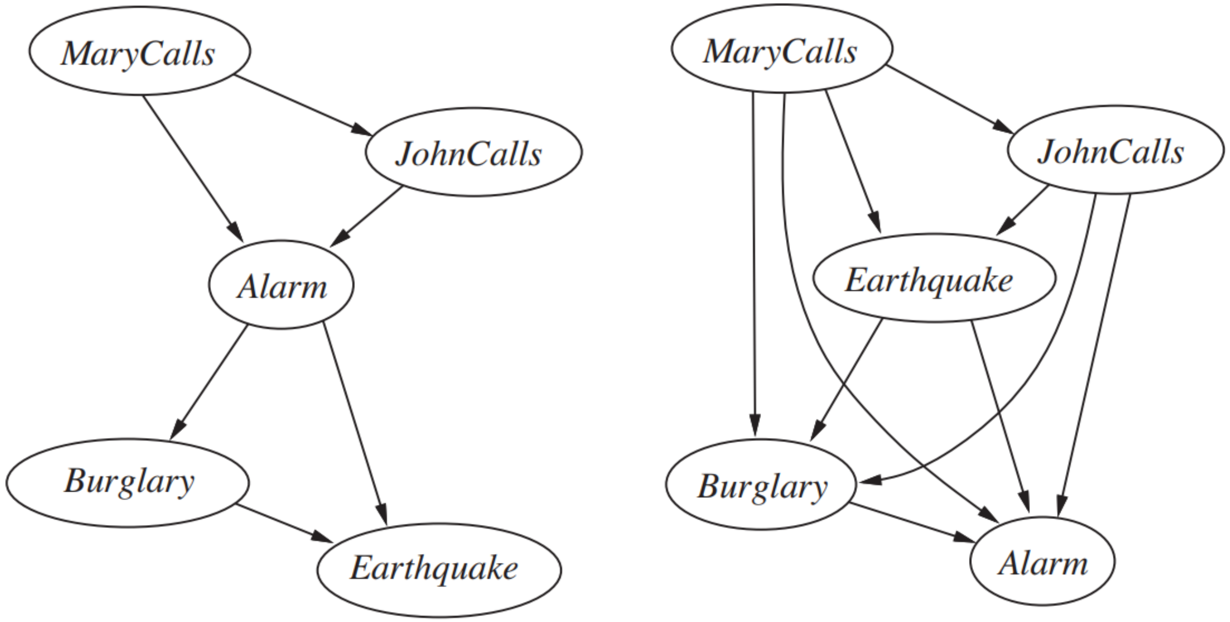
$P(r, t) = P(r)\,P(t \mid r)$



$P(r, t) = P(t)\,P(r \mid t)$

Causality?





(a) (b)

Independence Given Evidence

General question: Are two variables $X,\ Y$ independent of each other conditioned on $Z = \{Z_1, Z_2, \ldots\}$?

Or: Are X and Y "D-separated" by Z?

Algorithm

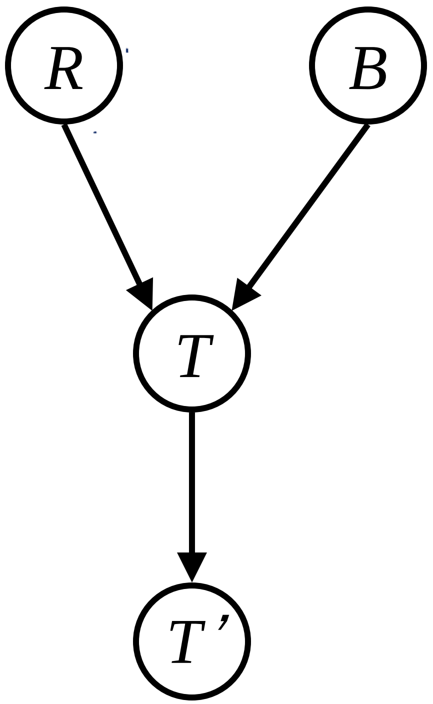
1. Consider just the **ancestral subgraph** consisting of X, Y, Z, and their ancestors.
2. Add links between any unlinked pair of nodes that share a common child; now we have the so-called **moral graph**.
3. Replace all directed links by undirected links.
4. If Z blocks all paths between X and Y in the resulting graph, then Z d-separates X and Y.

Example

$R \perp\!\!\!\perp B$ *Yes*

$R \perp\!\!\!\perp B | T$

$R \perp\!\!\!\perp B | T'$



Example

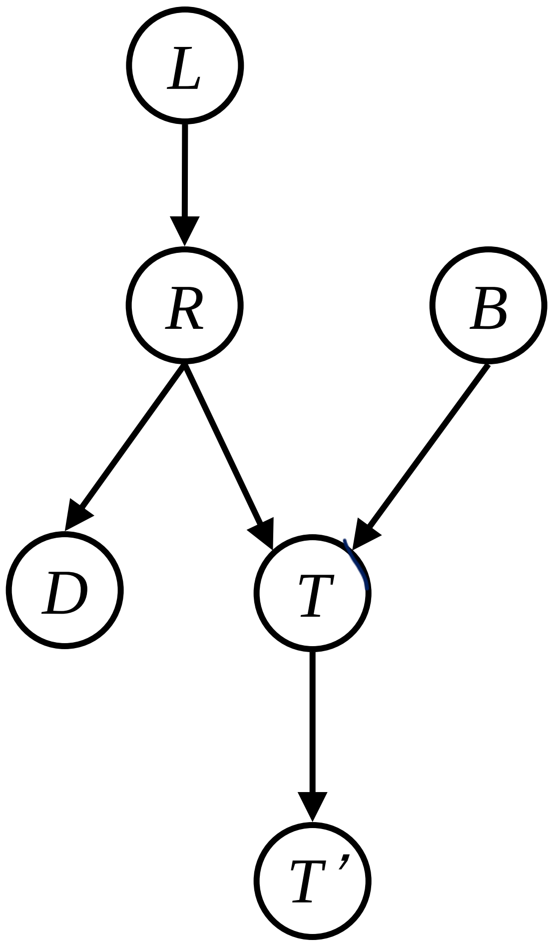
$L \perp\!\!\!\perp T' | T$ Yes

$L \perp\!\!\!\perp B$ Yes

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$ Yes



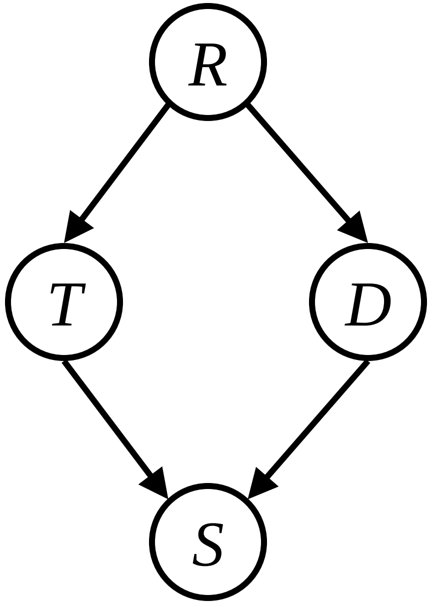
Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

$$T \perp\!\!\!\perp D$$

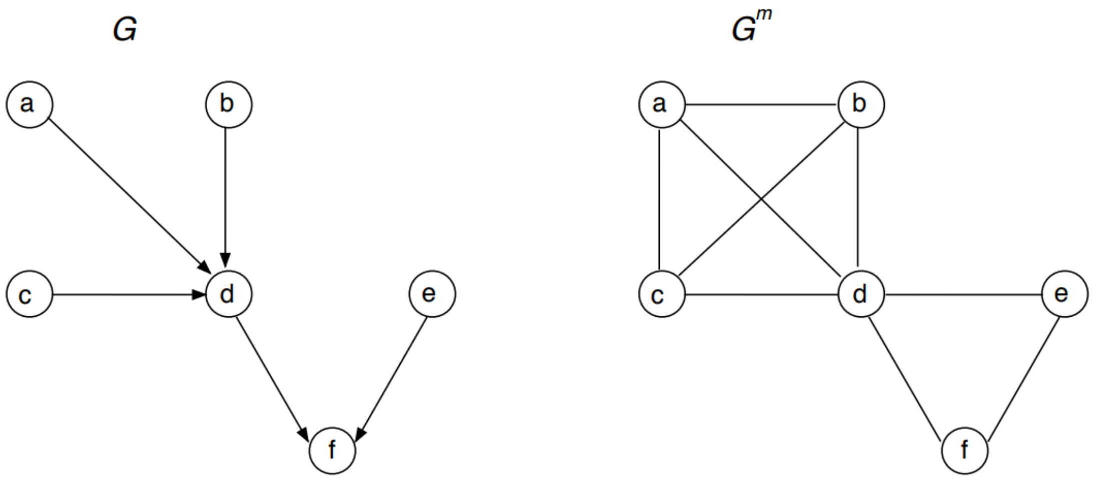
$$T \perp\!\!\!\perp D | R$$ Yes

$$T \perp\!\!\!\perp D | R, S$$



Proof Sketch

Statement: If X and Y and separated by Z in the moral graph, then X ⫫ Y | Z



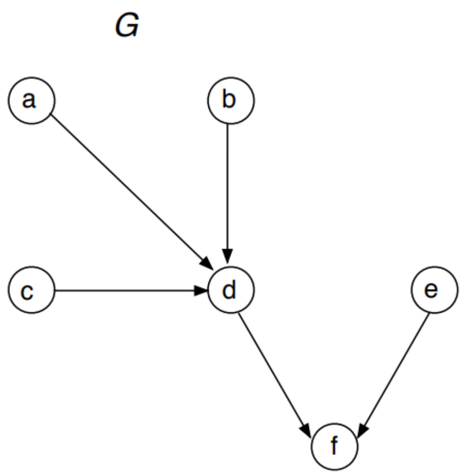
The moral graph gives a way to **"factorize"** the joint distribution of BN.

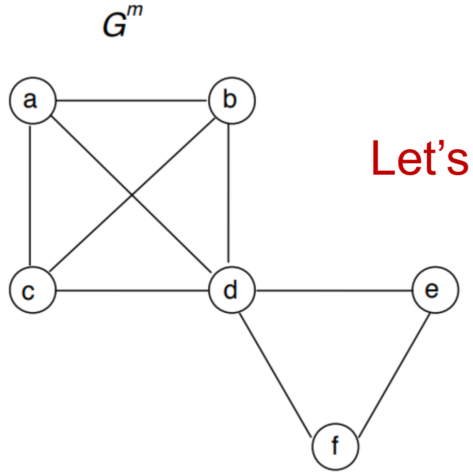
Each **clique** in the moral graph is a **factor**.

$$\underbrace{P(a)\,P(b)\,P(c)\,P(d \mid a, b, c)}_{\phi(a,\, b,\, c,\, d)}\ \underbrace{P(e)\,P(f \mid d, e)}_{\phi(d,\, e,\, f)} = \phi(a, b, c, d)\,\phi(d, e, f)$$

Proof Sketch

Statement: If X and Y and separated by Z in the moral graph, then X ⫫ Y | Z





Let's try to prove a ⫫ f | d

$$P(a|d) = \frac{P(a,d)}{P(d)} = \frac{\sum_f \phi(a,d)\phi(d,f)}{\sum_{a,f} \phi(a,d)\phi(d,f)} = \frac{\phi(a,d)\sum_f \phi(d,f)}{\sum_a \phi(a,d)\sum_f \phi(d,f)} = \frac{\phi(a,d)}{\sum_a \phi(a,d)}$$

$$P(a|d,f) = \frac{P(a,d,f)}{P(d,f)} = \frac{\phi(a,d)\phi(d,f)}{\sum_a \phi(a,d)\phi(d,f)} = \frac{\phi(a,d)}{\sum_a \phi(a,d)}$$

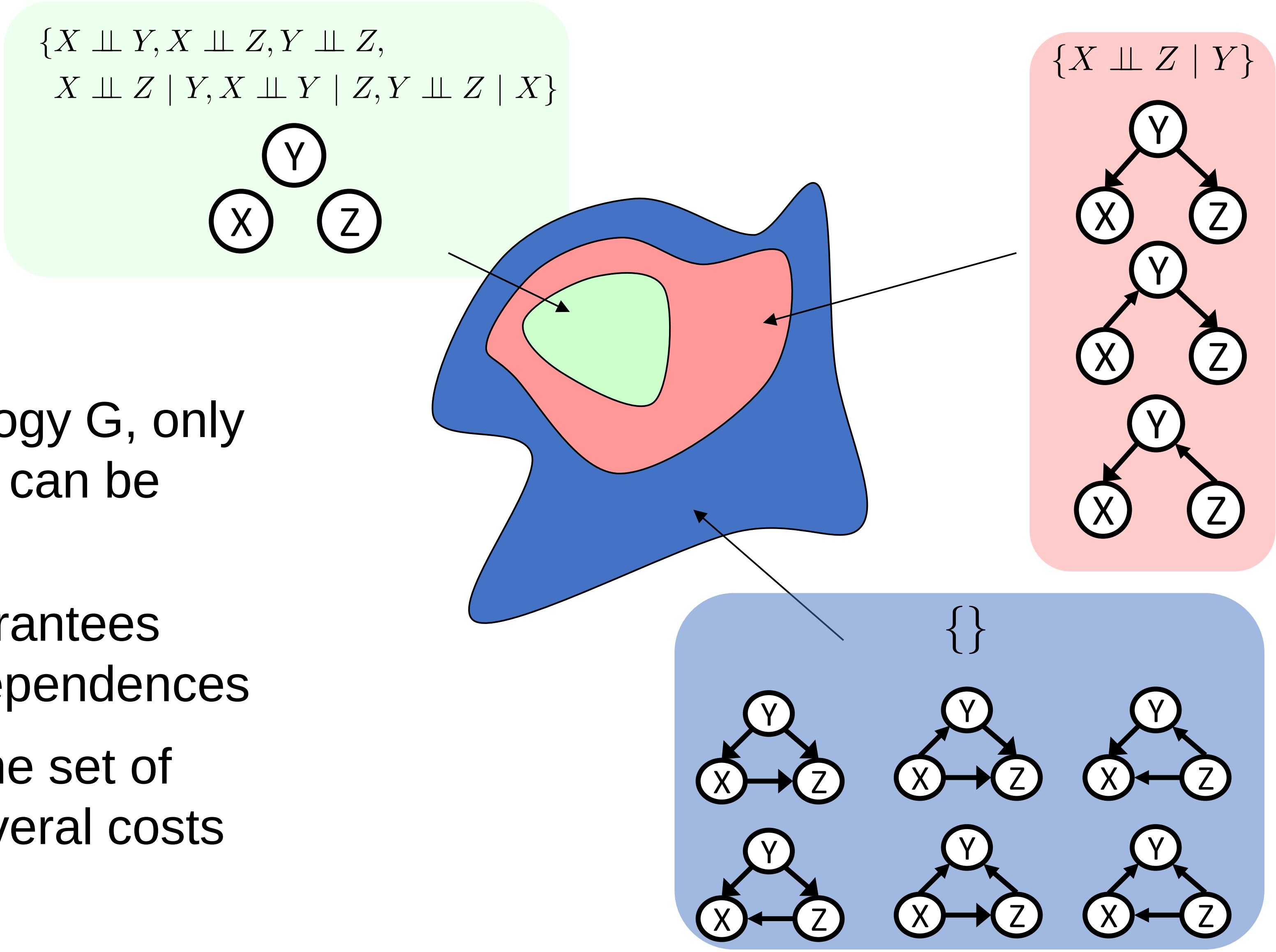
Structure Implications

- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

Topology Limits Distributions



- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- Adding arcs increases the set of distributions, but has several costs

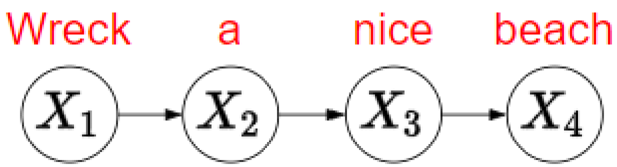
Application: Language Modeling

- Markov Model

Probabilistic program: Markov model

For each position $i = 1, 2, \ldots, n$:

Generate word $X_i \sim p(X_i \mid X_{i-1})$



Application: Object Tracking

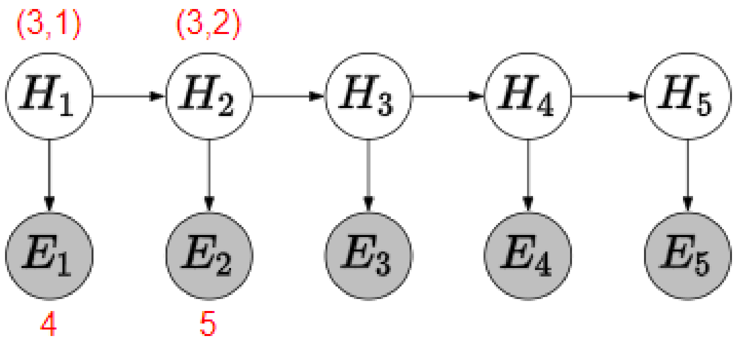
- Hidden Markov Model

Probabilistic program: hidden Markov model (HMM)

For each time step $t = 1, \ldots, T$:

Generate object location $H_t \sim p(H_t \mid H_{t-1})$

Generate sensor reading $E_t \sim p(E_t \mid H_t)$



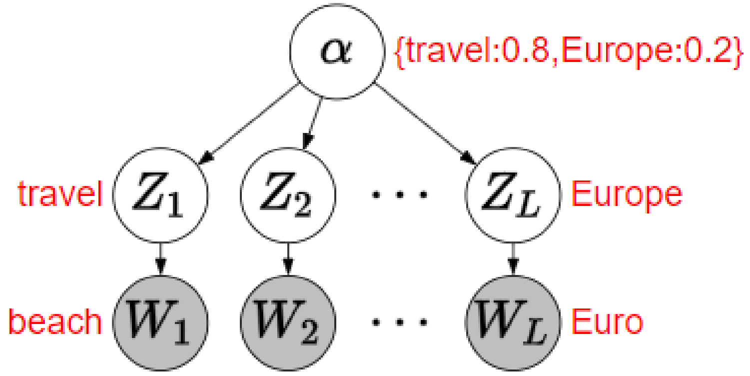
Inference: given sensor readings, where is the object?

Application: Topic Modeling

- Latent Dirichlet Allocation

Probabilistic program: latent Dirichlet allocation

Generate a distribution over topics $\alpha \in \mathbb{R}^K$
For each position $i = 1, \ldots, L$:
    Generate a topic $Z_i \sim p(Z_i \mid \alpha)$
    Generate a word $W_i \sim p(W_i \mid Z_i)$



Document classification, information retrieval, customer segmentation, …

Inference: given a text document, what topics is it about?

Next Time

- Inference