Probability and Structure in Natural Language Processing

Noah Smith, Carnegie Mellon University

2012 International Summer School in Language and Speech Technologies

Introduction

Motivation

- Statistical methods in NLP arrived ~20 years ago and now dominate.
- Mercer was right: "There's no data like more data."
  - And there's more and more data.

- Lots of new applications and new statistical techniques.
- My goal is to synthesize ideas you may have seen before ...

Thesis

- Most of the main ideas are related and similar to each other.
  - Different approaches to decoding.
  - Different learning criteria.
  - Supervised and unsupervised learning.
- Umbrella: probabilistic reasoning about discrete linguistic structures.

- This is good news!

Introduction

- Noah – professor at CMU since 2006
  - Language Technologies Institute
  - Machine Learning Department
  - *Linguistic Structure Prediction* (2011)
  - Courses: "Language and Statistics II," "Probabilistic Graphical Models," "Structured Prediction," "Algorithms for Natural Language Processing" at CMU

- This course was codesigned with **Shay Cohen**, now at Columbia University.

Plan

1.	Graphical models	M 8:00-9:30
2.	Probabilistic inference	M 13:30-15:00
3.	Decoding and structures	T 8:00-9:30
4.	Supervised learning	T 14:30-16:00
5.	Hidden variables	W 8:00-9:30
6.	The Bayesian approach	W 13:30-15:00

Exhortations

- The content is formal, but the style doesn't need to be.
- Ask questions!
  - Help me find the right pace.
  - Lecture 6 can be dropped if we need to slow down.
- The course starts in machine learning and moves toward NLP.
  - Be patient.

Lecture 1: Graphical Models

Random Variables

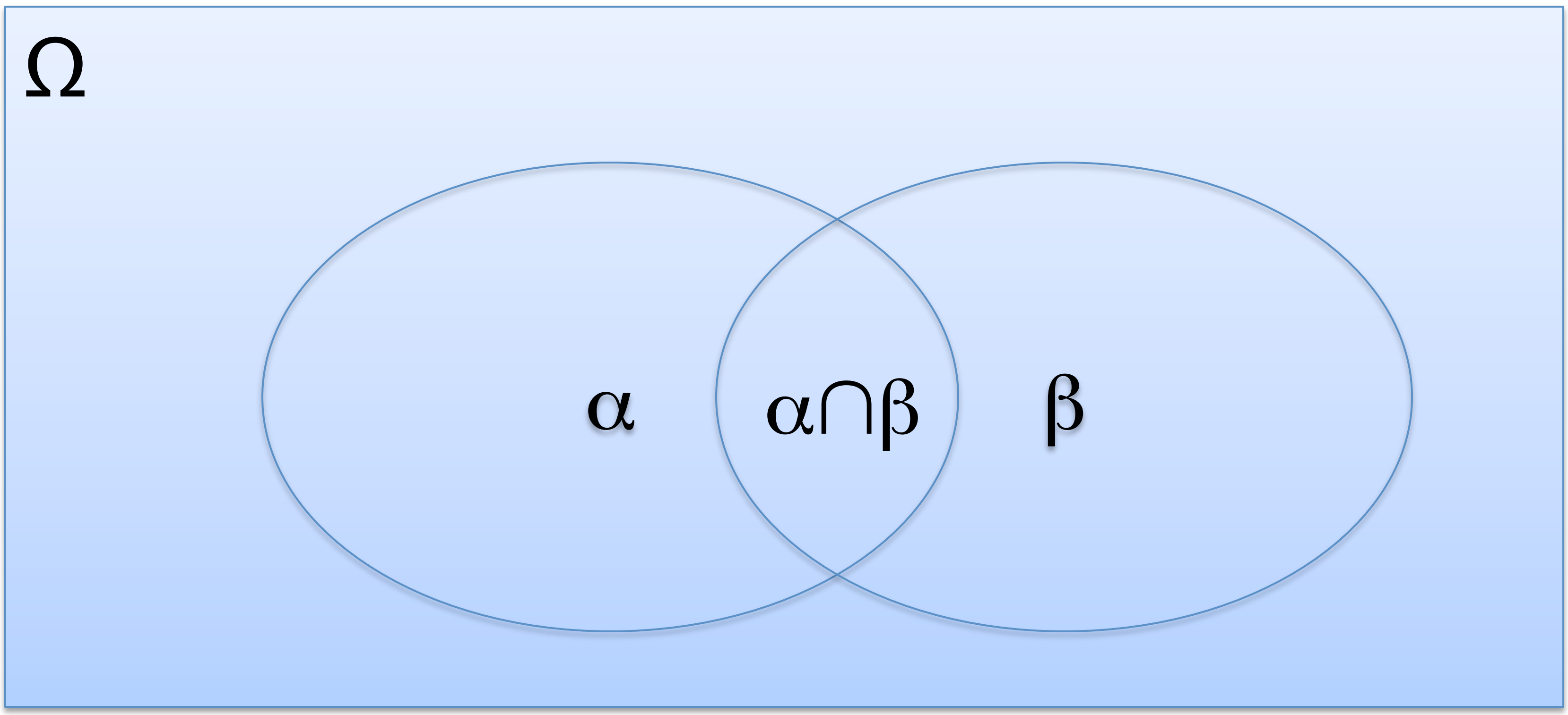
- Probability distributions usually defined by **events**
- Events are complicated!
  - We tend to *group* events by **attributes**
  - Person → Age, Grade, HairColor
- **Random variables** formalize attributes:
  - "Grade = A" is shorthand for event

$$\{\omega \in \Omega : f_{\text{Grade}}(\omega) = A\}$$

- Properties of random variable X:
  - Val(X) = possible values of X
  - For discrete (categorical): $\sum_{x \in \text{Val}(X)} P(X = x) = 1$
  - For continuous: $\int P(X = x)dx = 1$
  - Nonnegativity: $\forall x \in \text{Val}(X), P(X = x) \geq 0$

Conditional Probabilities

- After learning that α is true, how do we feel about β? $P(\beta \mid \alpha)$



Chain Rule

$$P(\alpha \cap \beta) = P(\alpha)P(\beta \mid \alpha)$$

$$P(\alpha_1 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 \mid \alpha_1) \cdots P(\alpha_k \mid \alpha_1 \cap \ldots \cap \alpha_{k-1})$$

Bayes Rule

likelihood

prior

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}$$

posterior

normalization constant

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$$

γ is an “external event”

Independence

- α and β are **independent** if $P(\beta|\alpha) = P(\beta)$

$$P \rightarrow (\alpha \perp \beta)$$

- **Proposition:** α and β are **independent** if and only if $P(\alpha \cap \beta) = P(\alpha)\, P(\beta)$

Conditional Independence

- Independence is rarely true.

- α and β are **conditionally independent** given γ if $P(\beta \mid \alpha \cap \gamma) = P(\beta \mid \gamma)$

$$P \rightarrow (\alpha \perp \beta \mid \gamma)$$

Proposition: $P \rightarrow (\alpha \perp \beta \mid \gamma)$ if and only if

$$P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma)\, P(\beta \mid \gamma)$$

Joint Distribution and Marginalization

$$P(\text{Grade}, \text{Intelligence}) =$$

	Intelligence = very high	Intelligence = high
Grade = A	0.70	0.10
Grade = B	0.15	0.05

- Compute the marginal over each individual random variable?

Marginalization: General Case

$$P(X_i = x) = \sum_{x_1 \in \mathrm{Val}(X_1), x_2 \in \mathrm{Val}(X_2), \ldots, x_{i-1} \in \mathrm{Val}(X_{i-1}), x_{i+1} \in \mathrm{Val}(X_{i+1}), \ldots, x_n \in \mathrm{Val}(X_n)} P(X_1 = x_1, X_2 = x_2, \ldots, X_i = x, \ldots, X_n = x_n)$$

$$P(X_i = x) = \sum_{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} P(x_1, x_2, \ldots, x_i, \ldots, x_n)$$

How many terms?

Basic Concepts So Far

- Atomic outcomes: assignment of $x_1,\ldots,x_n$ to $X_1,\ldots, X_n$

- Conditional probability: $P(X, Y) = P(X)\ P(Y|X)$

- Bayes rule: $P(X|Y) = P(Y|X)\ P(X) / P(Y)$

- Chain rule: $P(X_1,\ldots,X_n) = P(X_1)\ P(X_2|X_1)$
  $\ldots\ P(X_k|X_1,\ldots,X_{k-1})$

Sets of Variables

- **Sets** of variables **X**, **Y**, **Z**
- **X** is independent of **Y** given **Z** if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} | \mathbf{Z}=\mathbf{z})$,
    $\forall\ \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$

- Shorthand:
  - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \models (\mathbf{X} \perp \mathbf{Y} \mid \varnothing)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** P satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$

Free Parameters

- Consider assigning a value to $P(X = x)$ for each x in $Val(X)$. How many free parameters, if $|Val(X)| = k$?

- Now consider $P(X_1, X_2, \ldots, X_n)$. How many?

- Can we do it with fewer parameters?

(Marginal) Independence

- Let's make a very strong independence assumption:

$$\forall \mathbf{Y} \subseteq \mathbf{X}, \mathbf{Z} \subseteq \mathbf{X}, \mathbf{Y} \perp \mathbf{Z}$$

- Joint distribution: $P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i)$

- How many free parameters now?

Independence Spectrum

various Bayesian networks

full independence assumptions

everything is dependent

$$\prod_{i=1}^{n} P(X_i)$$

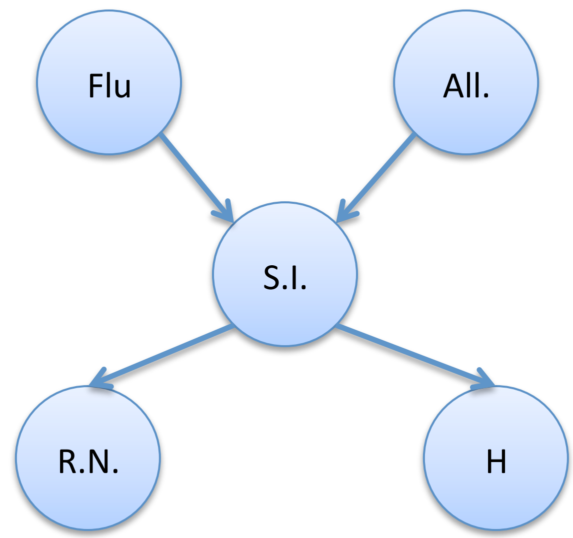
n parameters

$$P(X_1, \ldots, X_n)$$

$2^n - 1$ parameters

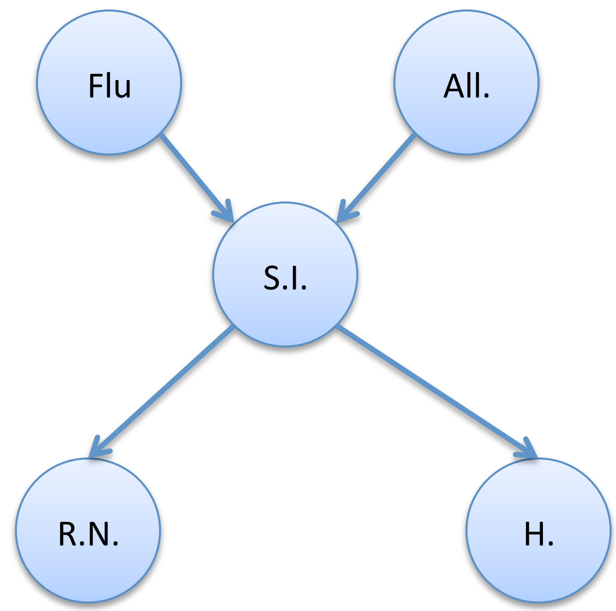
Causal Structure

- The flu causes sinus inflammation
- Allergies *also* cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches

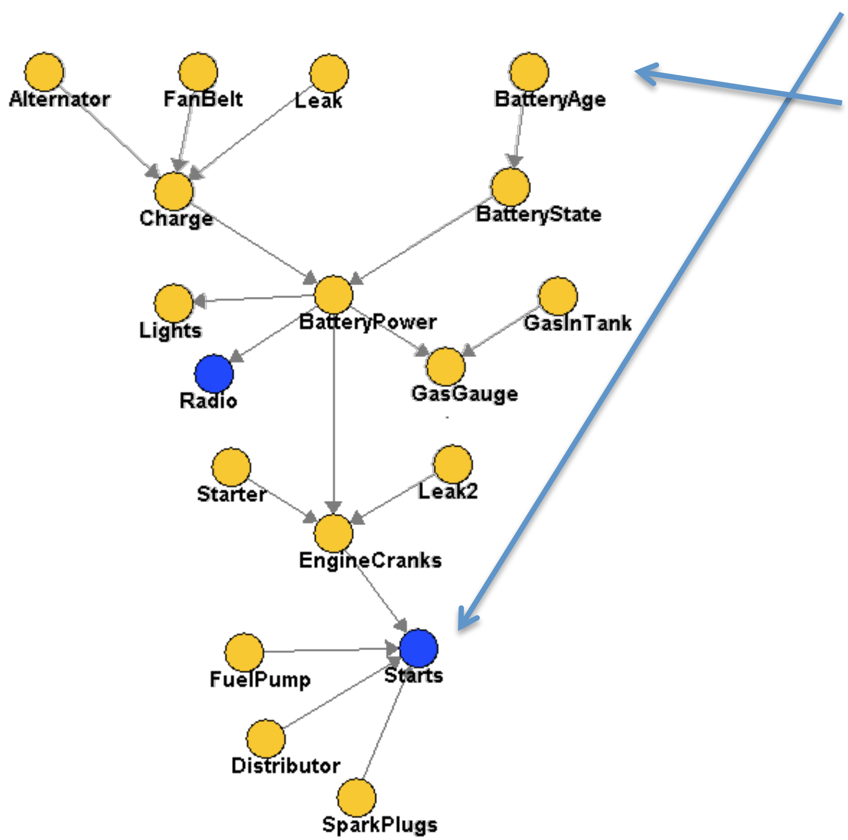


Querying the Model

- Inference (e.g., do you have allergies?)
- What's the best explanation?
- Active data collection (what is the next best r.v. to observe?)



A Bigger Example: Your Car

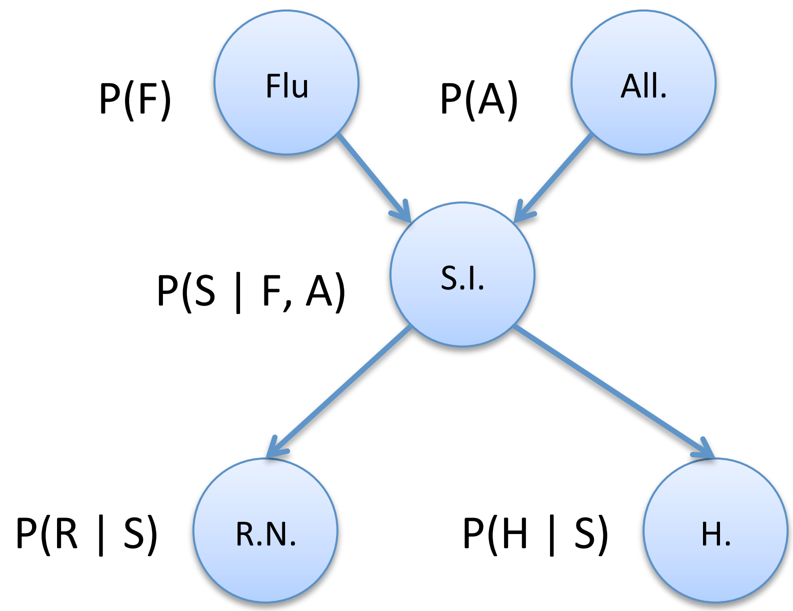


- The car doesn't start.
- What do we conclude about the battery age?

- 18 random variables
- Marginalization will have 2^{18} terms!

Factored Joint Distribution

- Want:
  P(F, A, S, R, H)
  = P(F)
  P(A)
  P(S | F, A)
  P(R | S)
  P(H | S)



- How many parameters?

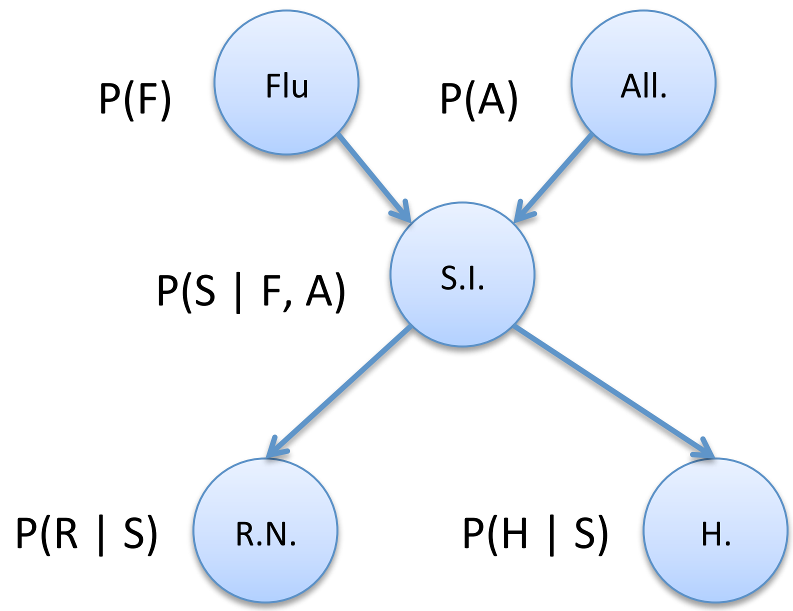
The BN Independence Assumption

- **Local Markov Assumption:** A variable X is independent of its non-descendants given its parents (and *only* its parents).

$$X \perp \mathbf{NonDescendants}(X) \mid \mathbf{Parents}(X)$$

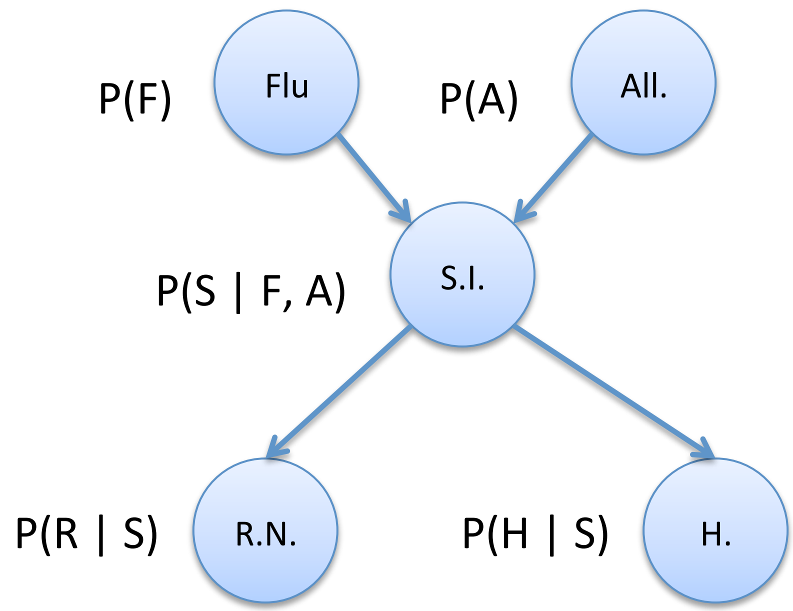
What's Independent?

- $F \perp A \mid \varnothing$



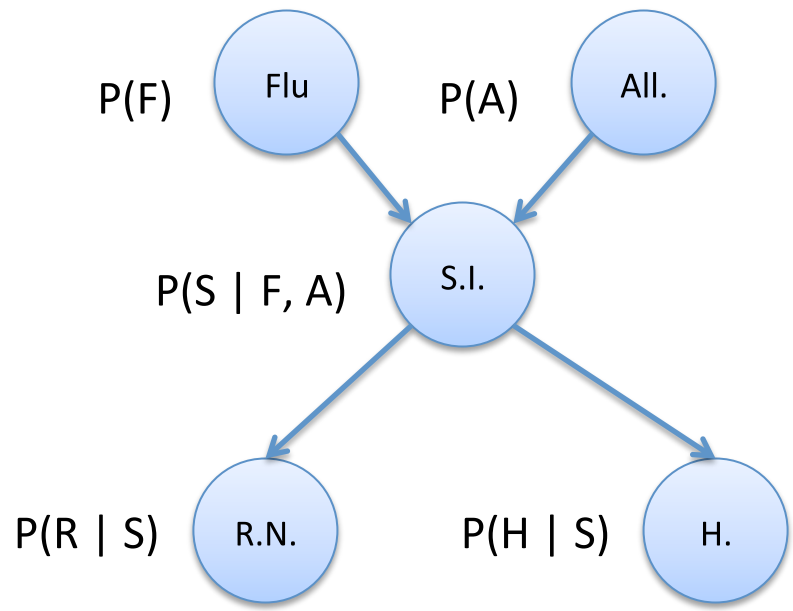
What's Independent?

- F ⊥ A | ∅
- A ⊥ F | ∅



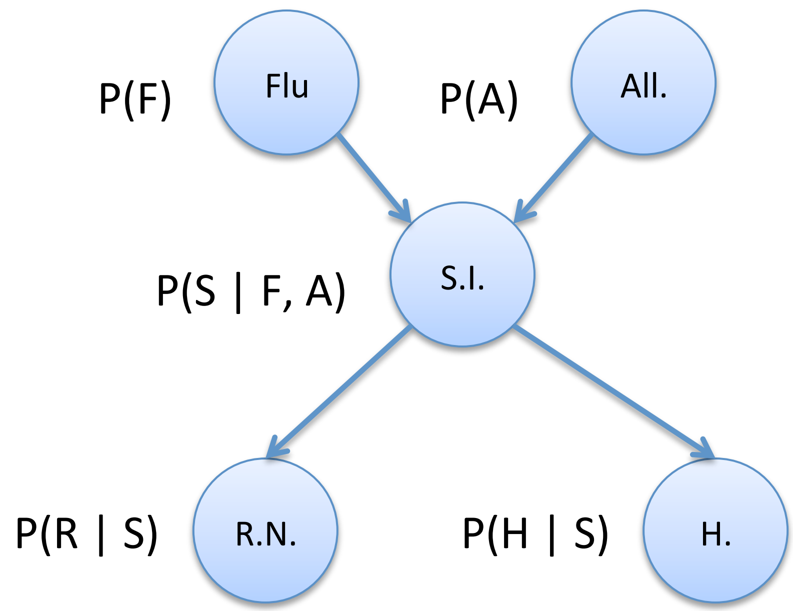
What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$
- S?



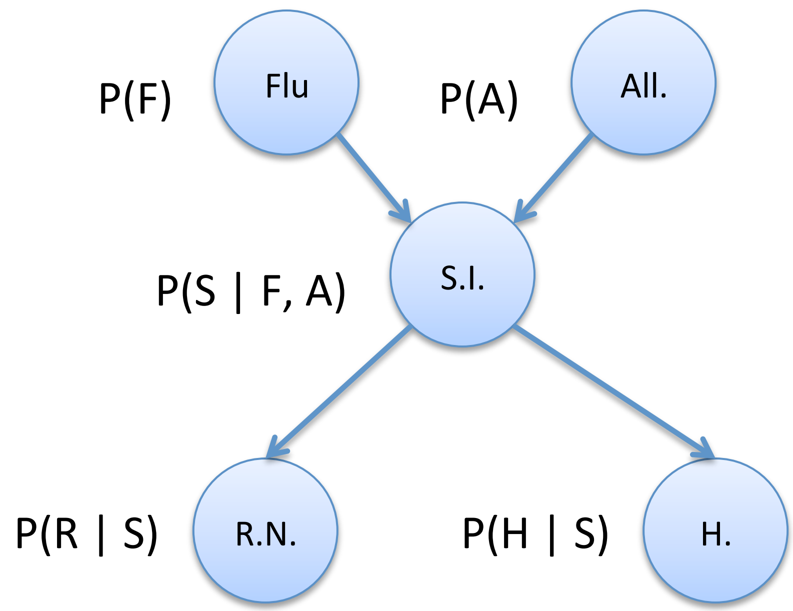
What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$
- S?
- $R \perp \{F, A, H\} \mid S$



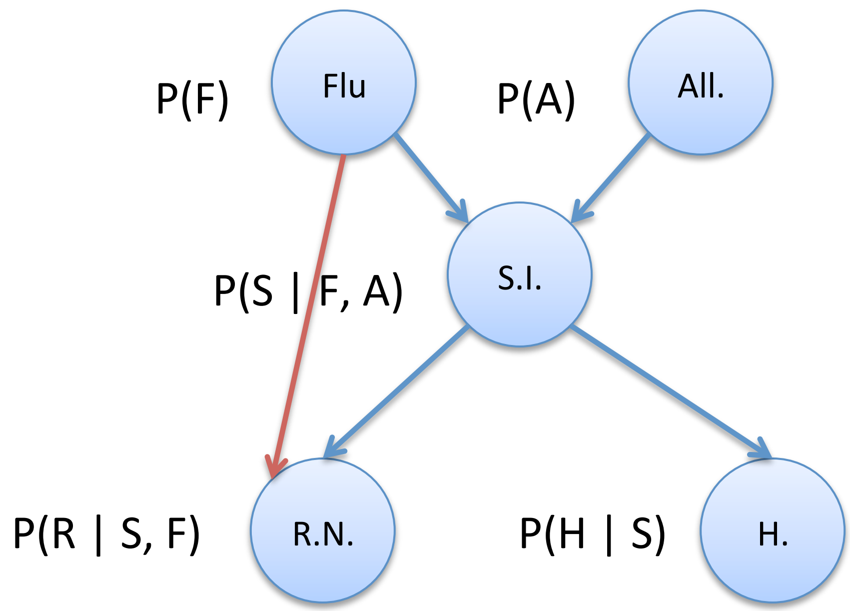
What's Independent?

- $F \perp A \mid \varnothing$
- $A \perp F \mid \varnothing$
- S?
- $R \perp \{F, A, H\} \mid S$
- $H \perp \{F, A, R\} \mid S$



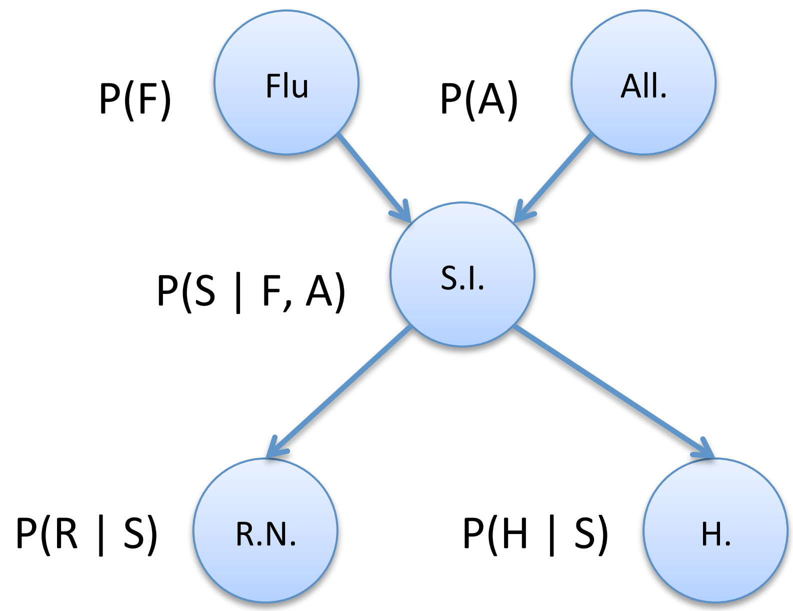
New Edge: What's Independent?

- F ⊥ A | ∅
- A ⊥ F | ∅
- S?
- R ⊥ {~~F,~~ A, H} | S, F
- H ⊥ {F, A, R} | S



A Puzzle

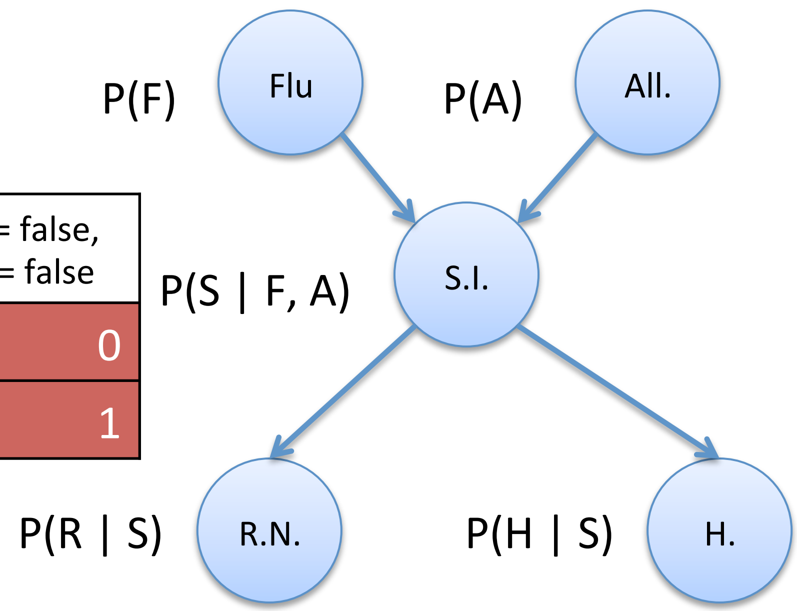
- $F \perp A \mid S$?



A Puzzle

- F ⊥ A | S ?

P(S|F, A)	F = true, A = true	F = true, A = false	F = false, A = true	F = false, A = false
true	0	1	1	0
false	1	0	0	1



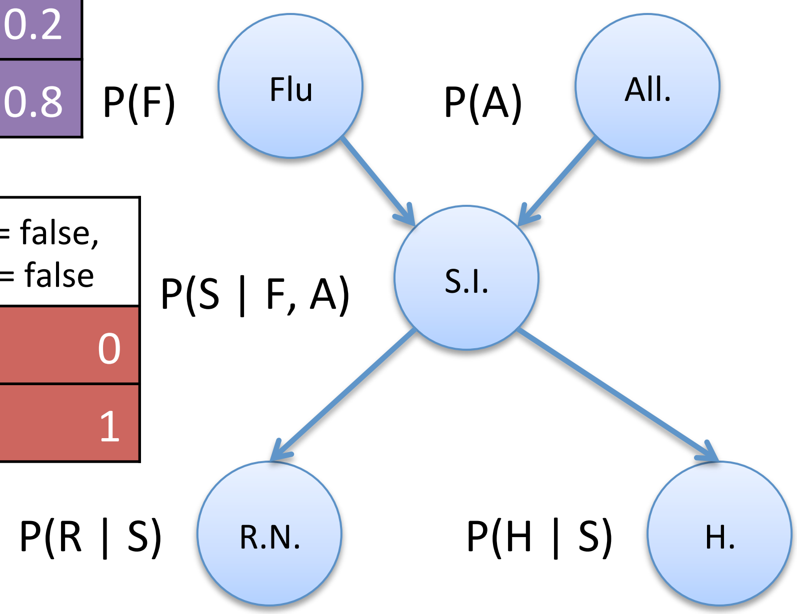
A Puzzle

true	0.2
false	0.8

true	0.2
false	0.8

- F ⊥ A | S ?

P(S|F, A)	F = true, A = true	F = true, A = false	F = false, A = true	F = false, A = false
true	0	1	1	0
false	1	0	0	1



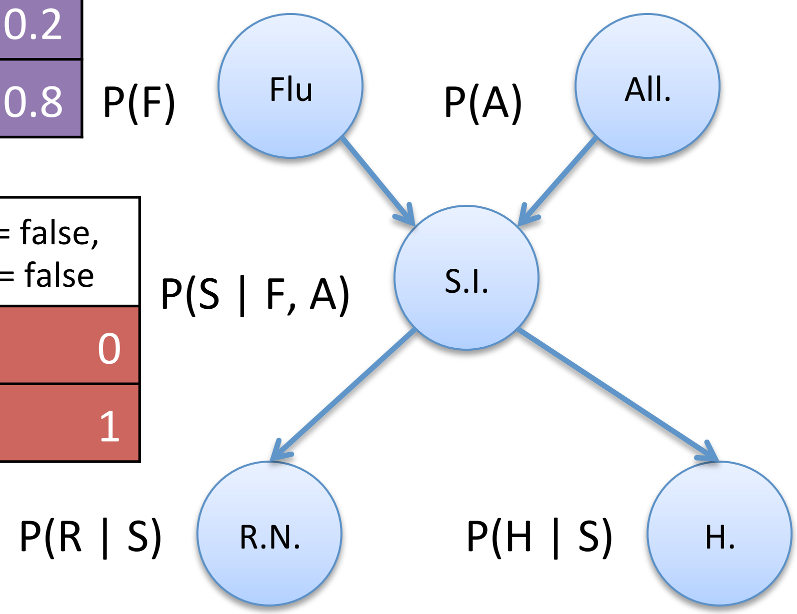
A Puzzle

true	0.2
false	0.8

true	0.2
false	0.8

- $F \perp A \mid S$?

P(S|F, A)	F = true, A = true	F = true, A = false	F = false, A = true	F = false, A = false
true	0	1	1	0
false	1	0	0	1



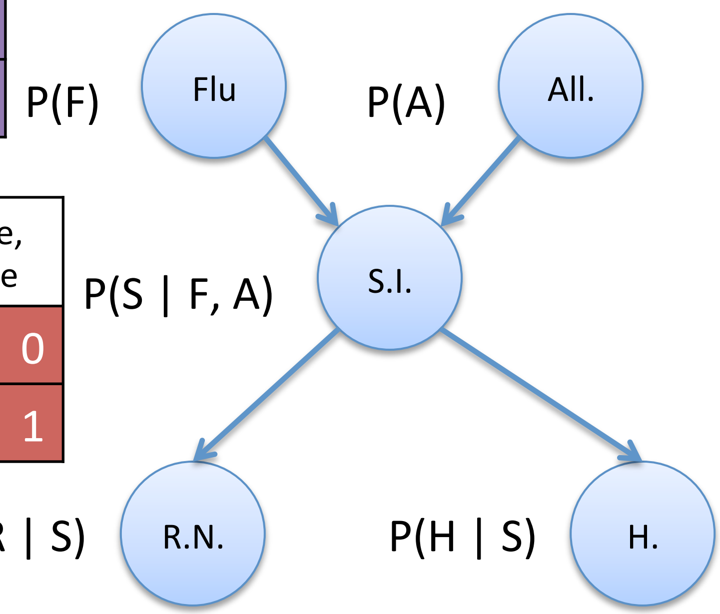
- $P(F = \text{true}) = 0.2$
- $P(F = \text{true} \mid S = \text{true}) = 0.5$
- $P(F = \text{true} \mid S = \text{true}, A = \text{true}) = 0$

A Puzzle

true	0.2
false	0.8

- F ⊥ A | S ?

true	0.2
false	0.8



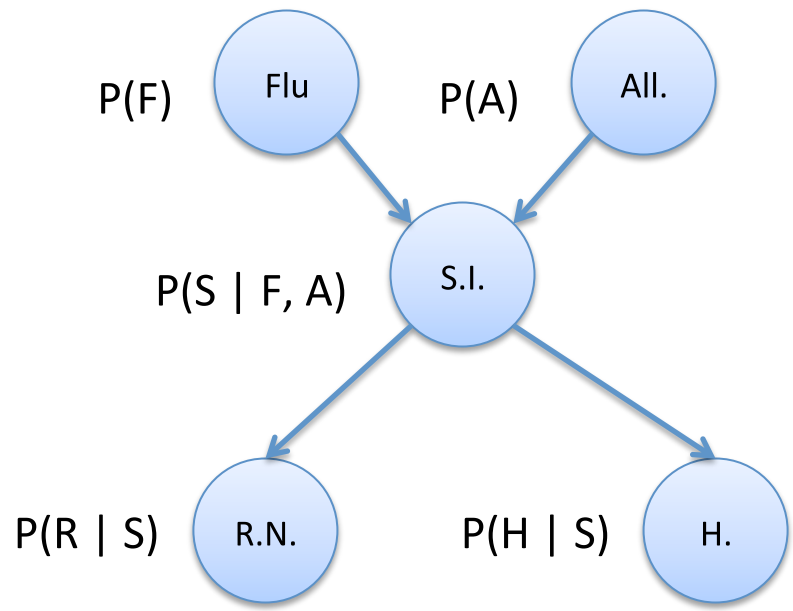
P(S|F, A)	F = true, A = true	F = true, A = false	F = false, A = true	F = false, A = false
true	ε	1	1	0
false	1 - ε	0	0	1

- P(F = true) = 0.2
- P(F = true | S = true) = (ε + 4)/(ε + 8)
- P(F = true | S = true, A = true) = ε

A Puzzle

- F ⊥ A | S ?

- In general, **no**.
  - This independence statement does not follow from the Local Markov assumption.
- ¬ (F ⊥ A | S)



Recipe for a Bayesian Network

- Set of random variables **X**
- Directed acyclic graph (each X_i is a vertex)
- Conditional probability tables, P(X | **Parents**(X))
- Joint distribution:

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Parents}(X_i))$$

- Local Markov Assumption
  - A variable X is independent of its non-descendants given its parents (and *only* its parents).

$$X \perp \mathbf{NonDescendants}(X) \mid \mathbf{Parents}(X)$$

Questions

1. Given a BN, what distributions can be represented?
2. Given a distribution, what BNs can represent it?
3. In addition to the Local Markov Assumption, what other independence assumptions are encoded in a given BN?

Representation Theorem

The conditional independencies in our BN are a subset of the independencies in P.

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Parents}(X_i))$$

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

Questions

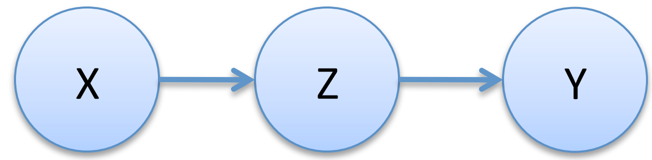
1. Given a BN, what distributions can be represented?
2. Given a distribution, what BNs can represent it?
3. In addition to the Local Markov Assumption, what other independence assumptions are encoded in a given BN?

Independencies

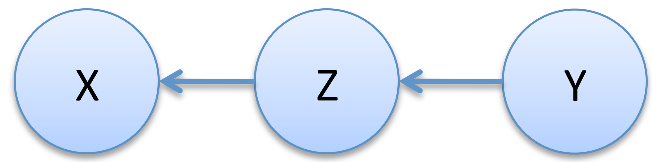
- Local Markov Assumption:
  $X_i \perp$ **NonDescendants**(X_i) | **Parents**(X_i)
- Are there other independencies that we can derive?
  - Yes.
  - Let's consider some three-node Bayesian networks.

Three-Node BNs

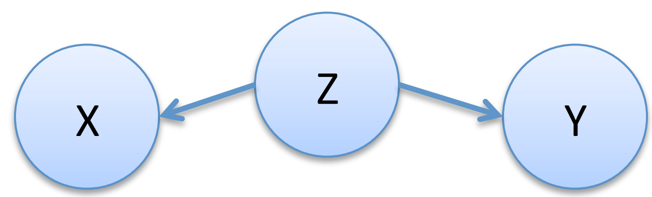
- Indirect causal effect



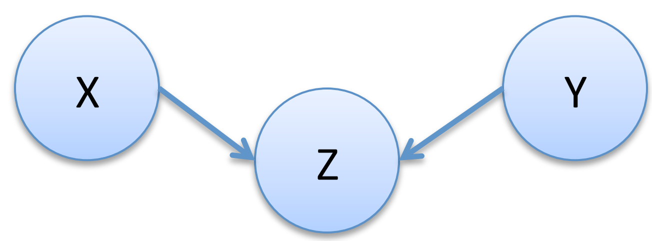
- Indirect evidential effect



- Common cause
  $(X \perp Y \mid Z), \neg(X \perp Y)$

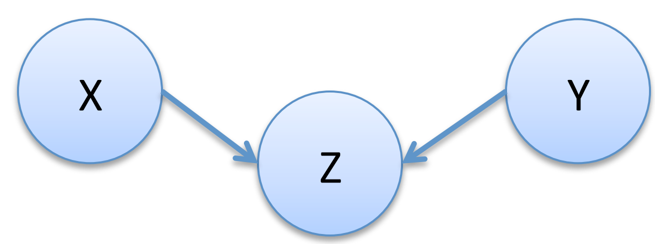


- Common effect
  (V-structure)
  $(X \perp Y), \neg(X \perp Y \mid Z)$



V-Structures, or Colliders

- Let $Z = X \oplus Y$.
  - Yes, random variables can be deterministic functions!
- In this case, if I know Z, then X and Y are dependent, because they cannot be equal!
- $\neg(X \perp Y \mid Z)$



What We Want

- A general test for conditional independence in a Bayesian network!

- Surprisingly enough, we can characterize all independence assumptions in a Bayesian network based on the simple constructs of three-node BNs

Observations and Conditional Independence

- Note: when we observe a certain outcome of a variable, we condition on its value
- "X and Y are independent when we observe Z": $X \perp Y \mid Z$

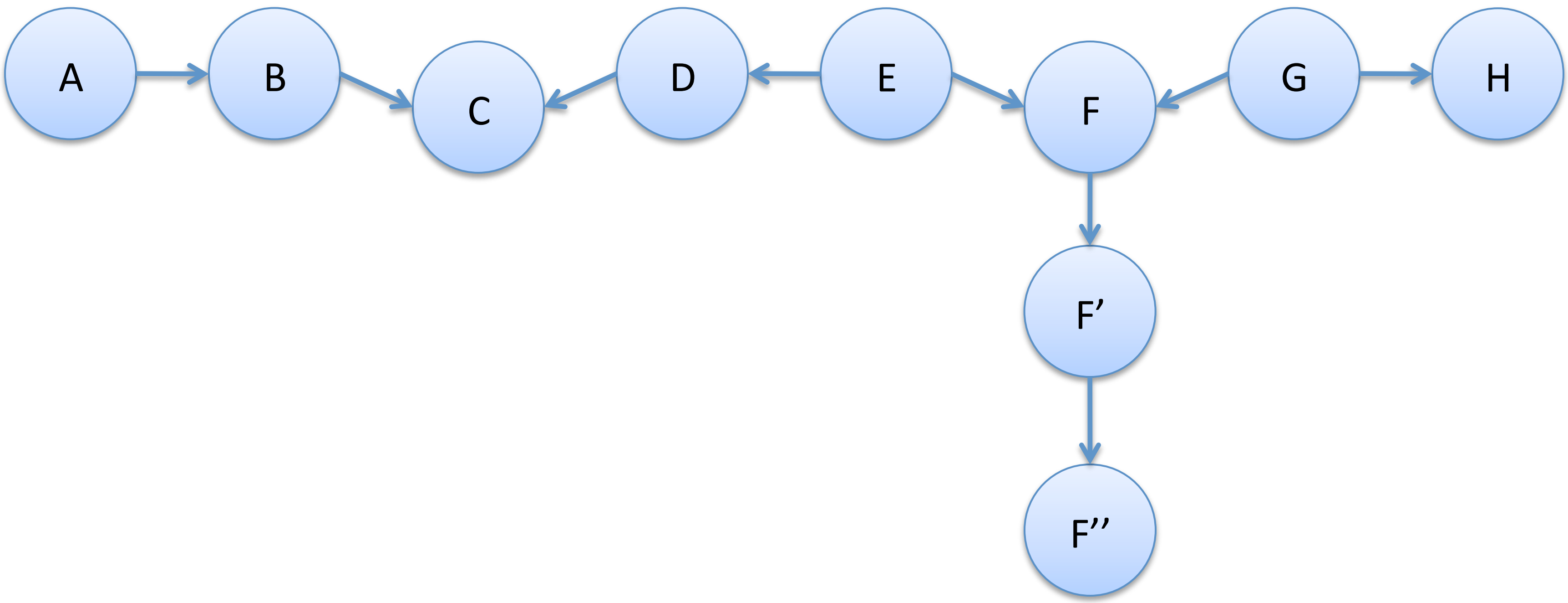
Active Trails, Formalized

- Trail: undirected path that doesn't visit any nodes more than once
- A trail $X_1 \rightleftarrows X_2 \rightleftarrows \ldots \rightleftarrows X_k$ is an **active trail** if, for each consecutive triplet in the trail:
  - $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$ and X_i is not observed.
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$ and X_i is not observed.
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$ and X_i is not observed.
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ and X_i (or one of its descendents) **is** observed.

D-Separation

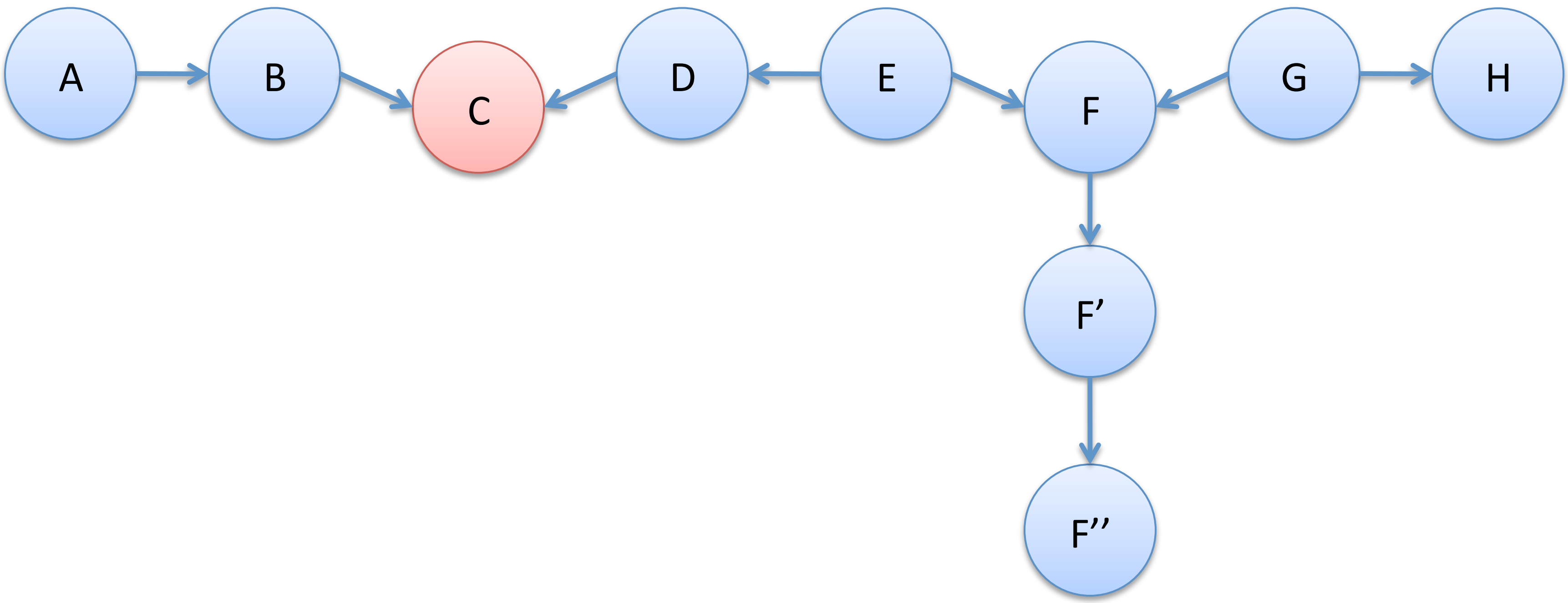
- Three sets of nodes: **X**, **Y**, and observed nodes **Z**
- **X** and **Y** are **d-separated** given **Z** if there is no active trail from any $X \in \mathbf{X}$ to any $Y \in \mathbf{Y}$ given **Z**.

Another Example



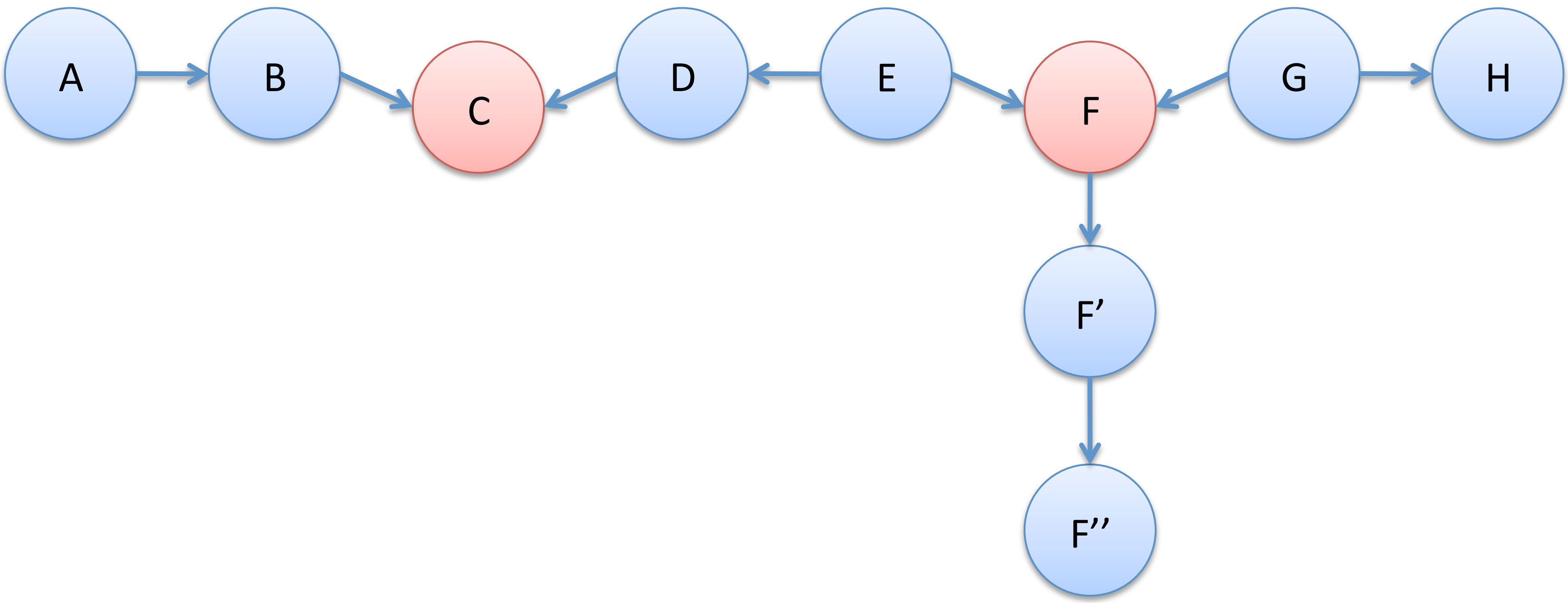
- If I observe nothing, then $A \perp H$.

Another Example



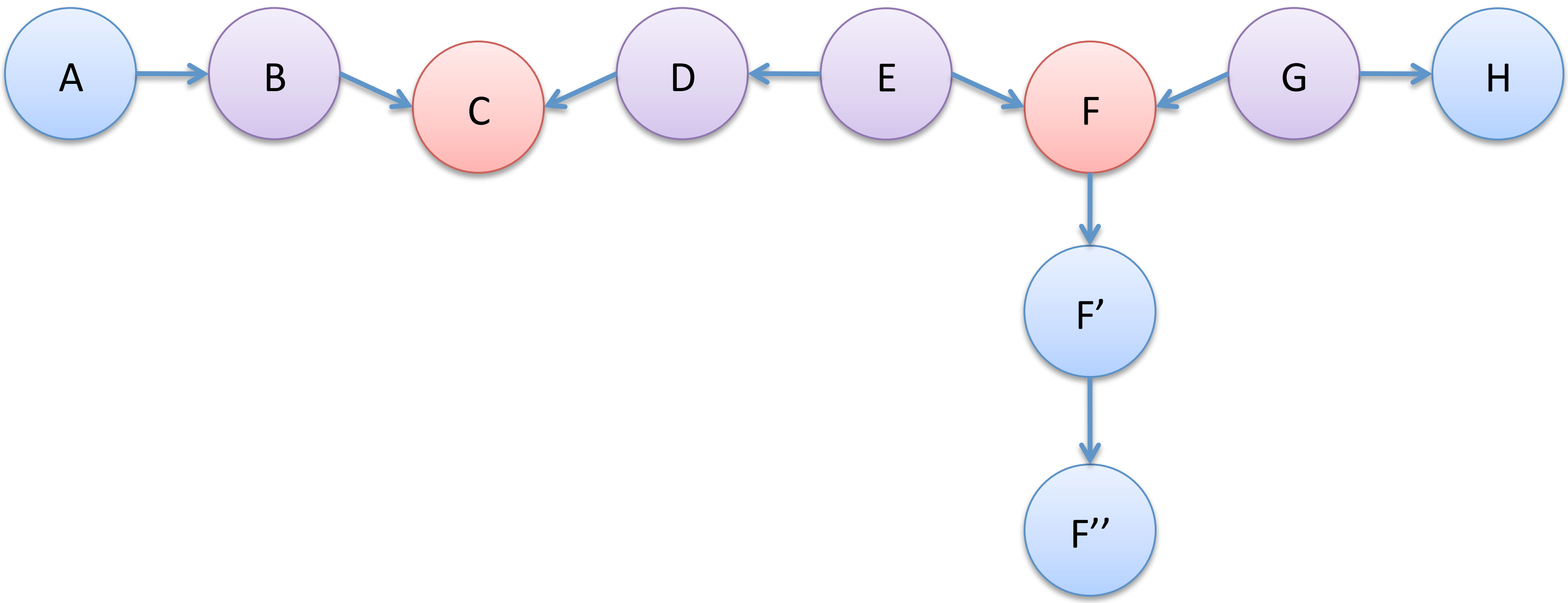
- If I observe C, then A ⊥ H.

Another Example



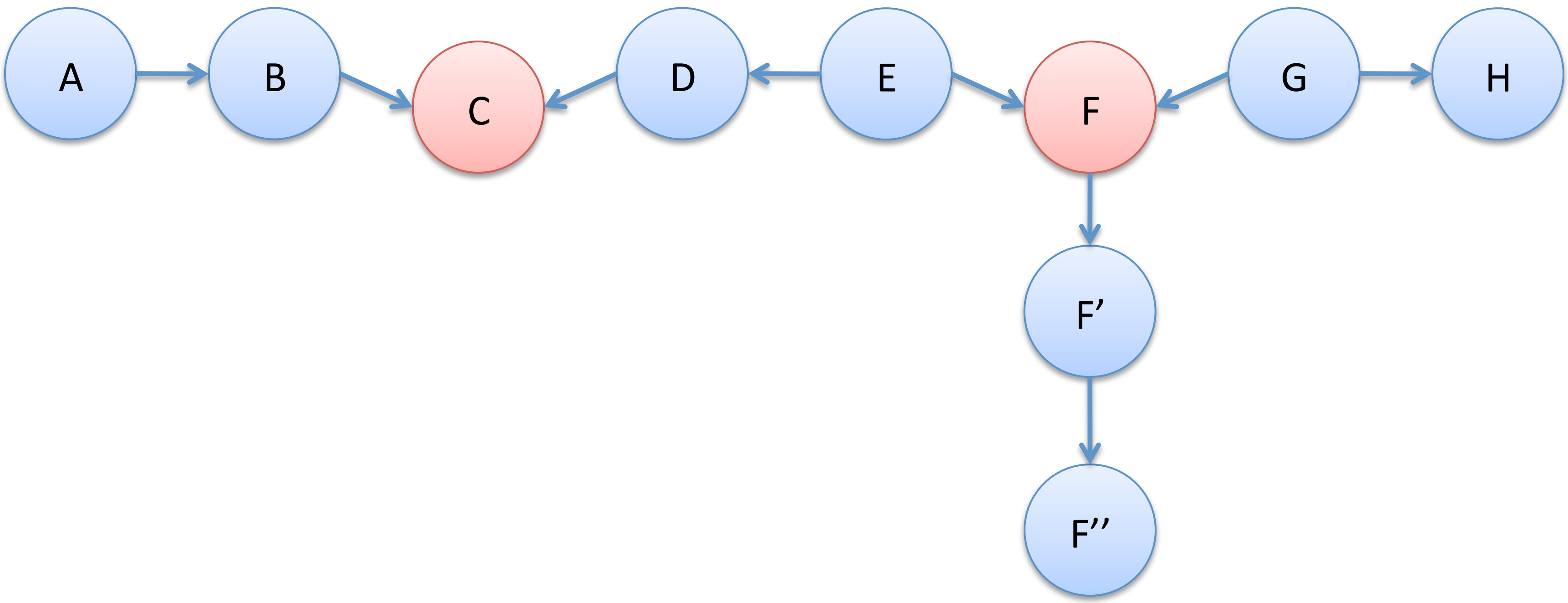
- If I observe C and F, then ¬(A ⊥ H).

Another Example



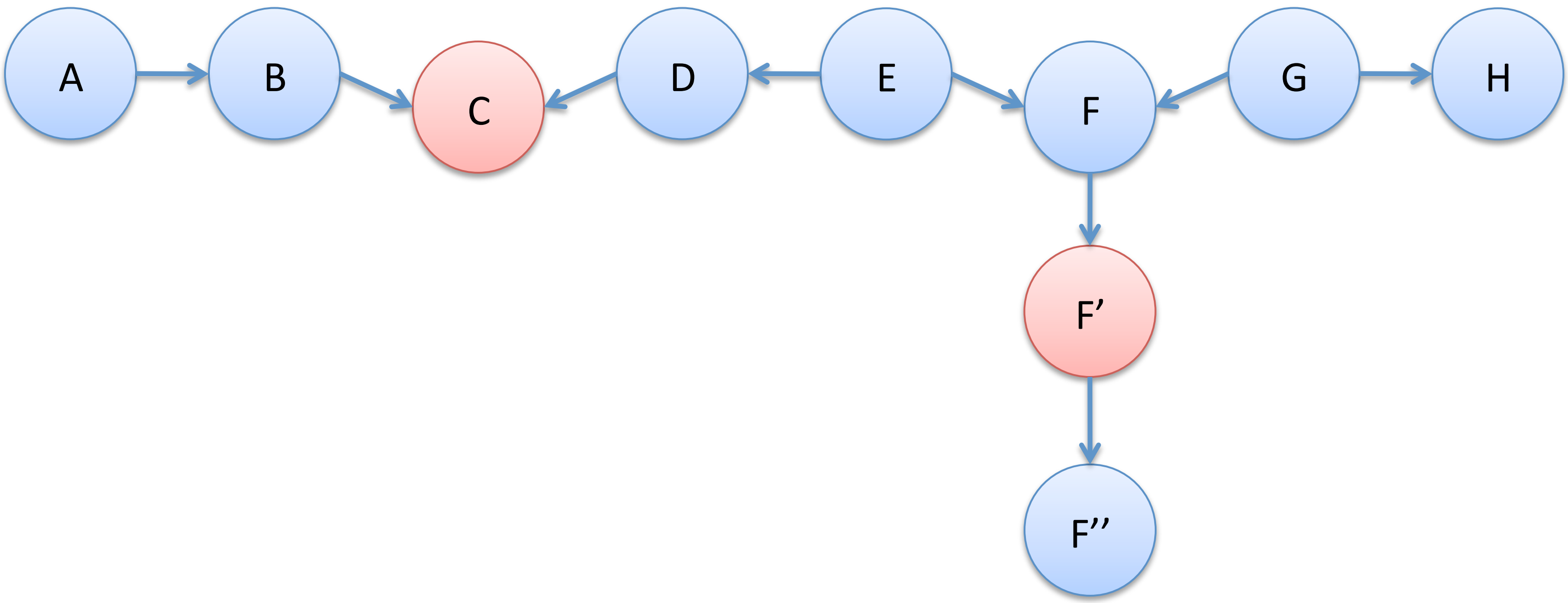
- If I observe C and F, then $\neg(A \perp H)$.
  - But if I observe B, D, E, and/or G, then $A \perp H$.

Another Example



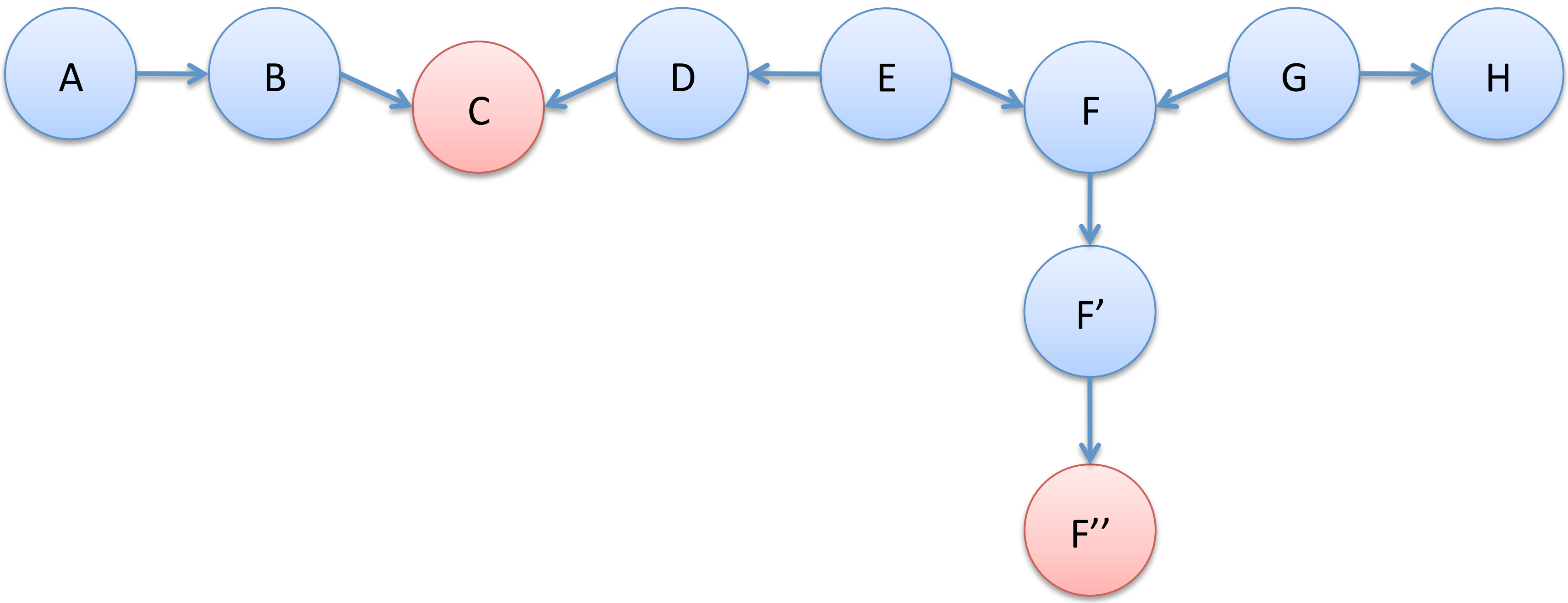
- If I observe C and F, then $\neg(A \perp H)$.

Another Example



- If I observe C and F', then ¬(A ⊥ H).

Another Example



- If I observe C and F’’, then $\neg(A \perp H)$.

Intuition

- Two variables can be dependent if there is a trail between them.
  - "Flow of influence" along active trails
- D-separation gives us a way to think about how that "flow of influence" could be blocked.
  - No active trail $\Rightarrow$ d-separation $\Rightarrow$ no dependence

Where We Are

- D-separation and independence
  - D-separation is a sound procedure for finding independencies: $I(G) \subseteq I(P)$
  - We can find a distribution respecting any such independency.
  - Almost all independencies can be read from the graph without recourse to the conditional probability tables. $I(G) \approx I(P)$.
    - Sometimes independencies can happen as an accident based on the probabilities!

Markov Networks

Perfect Maps (P-Maps)

- A graph G is a **P-map** for a distribution P if I(G) = I(P).

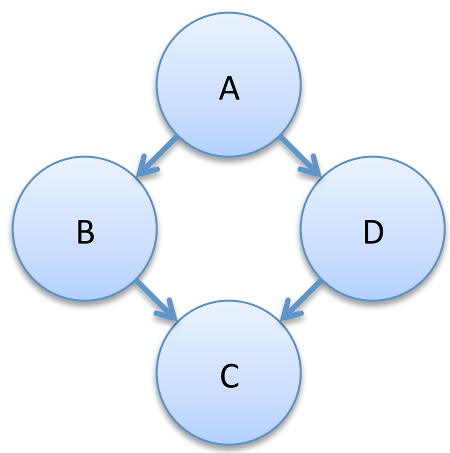
- Can we always construct one?

Motivating Example:
No Bayesian Network is a P-Map

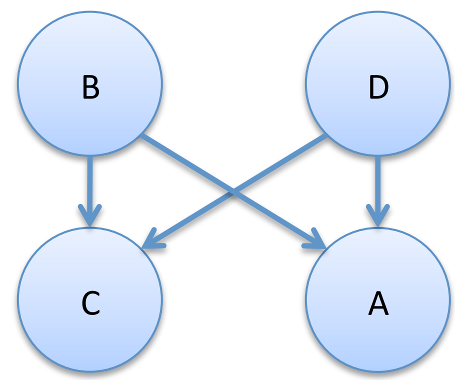
- Swinging couples or misunderstanding students

I(P):

- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg B \perp D$
- $\neg A \perp C$



Fails to capture:
$B \perp D \mid A, C$



Fails to capture:
$\neg B \perp D$

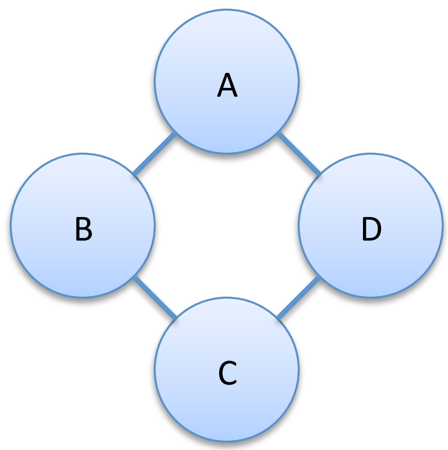
- Alice only talks to Bob and Debbie; Bob only talks to Charles and Alice; Charles only talks to Bob and Debbie; Debbie only talks to Alice and Charles

Motivating Example: This Markov Network is a P-Map!

- Swinging couples or misunderstanding students

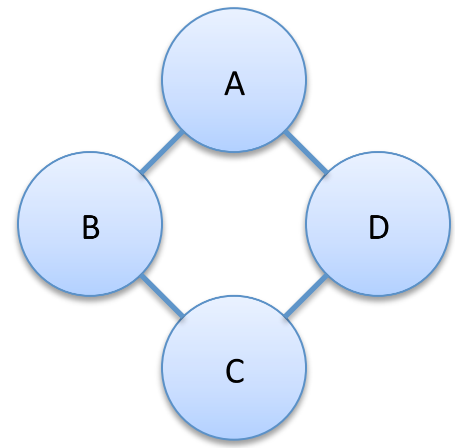
I(P):

- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg\, B \perp D$
- $\neg\, A \perp C$



Markov Networks

- Each random variable is a vertex.
- Undirected edges.
- **Factors** are associated with subsets of nodes that form cliques.
  - A factor maps assignments of its nodes to nonnegative values.

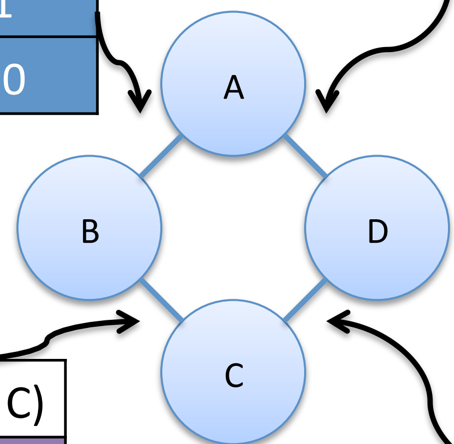


Markov Networks

- In this example, associate a factor with each edge.
  - Could also have factors for single nodes!

A	B	$\varphi_1(A, B)$
0	0	30
0	1	5
1	0	1
1	1	10

A	D	$\varphi_4(A, D)$
0	0	100
0	1	1
1	0	1
1	1	100



B	C	$\varphi_2(B, C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\varphi_3(C, D)$
0	0	1
0	1	100
1	0	100
1	1	1

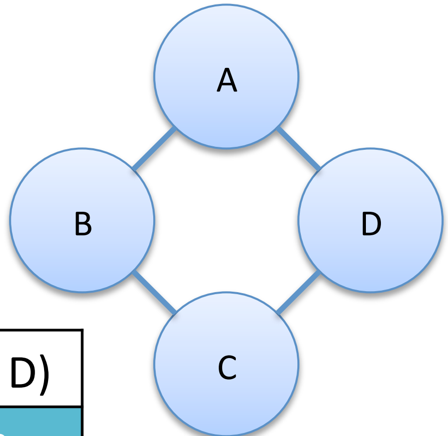
Markov Networks

- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$



A	B	φ_1(A, B)	B	C	φ_2(B, C)	C	D	φ_3(C, D)	A	D	φ_4(A, D)
0	0	30	0	0	100	0	0	1	0	0	100
0	1	5	0	1	1	0	1	100	0	1	1
1	0	1	1	0	1	1	0	100	1	0	1
1	1	10	1	1	100	1	1	1	1	1	100

Markov Networks

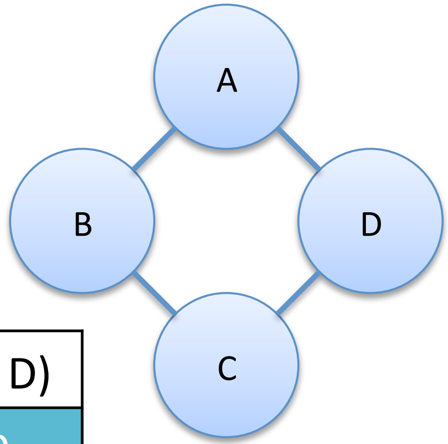
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	φ_1(A, B)
0	0	30
0	1	5
1	0	1
1	1	10

B	C	φ_2(B, C)
0	0	100
0	1	1
1	0	1
1	1	100

C	D	φ_3(C, D)
0	0	1
0	1	100
1	0	100
1	1	1

A	D	φ_4(A, D)
0	0	100
0	1	1
1	0	1
1	1	100

Markov Networks

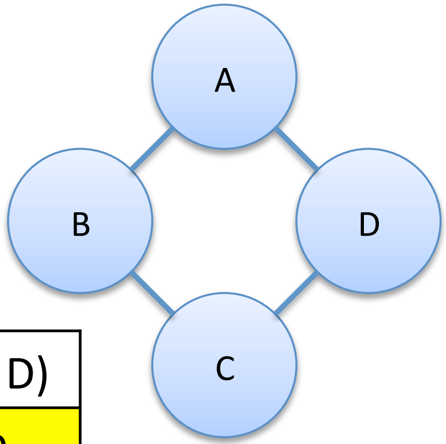
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	φ_1(A, B)
0	0	30
0	1	5
1	0	1
1	1	10

B	C	φ_2(B, C)
0	0	100
0	1	1
1	0	1
1	1	100

C	D	φ_3(C, D)
0	0	1
0	1	100
1	0	100
1	1	1

A	D	φ_4(A, D)
0	0	100
0	1	1
1	0	1
1	1	100

P(0, 1, 1, 0)
= 5,000,000 / Z
= 0.69

Markov Networks

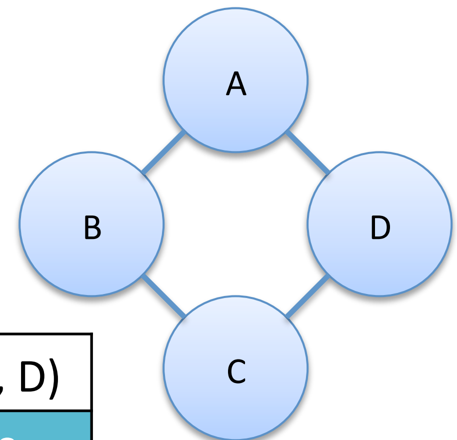
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	$\varphi_1(A, B)$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$\varphi_2(B, C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\varphi_3(C, D)$
0	0	1
0	1	100
1	0	100
1	1	1

A	D	$\varphi_4(A, D)$
0	0	100
0	1	1
1	0	1
1	1	100

P(1, 1, 0, 0)
= 10 / Z
= 0.0000014

Markov Networks (General Form)

- Let $\mathbf{D}_i$ denote the set of variables (subset of **X**) in the ith clique.
- Probability distribution is a **Gibbs** distribution:

$$
\begin{aligned}
P(\boldsymbol{X}) &= \frac{U(\boldsymbol{X})}{Z} \\
U(\boldsymbol{X}) &= \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i) \\
Z &= \sum_{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X})} U(\boldsymbol{x})
\end{aligned}
$$