Readings:
K&F: 3.1, 3.2, 3.3

BN Semantics 1

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

September 15th, 2006

Let's start on BNs…

- Consider $P(X_i)$ ← k-1 parms.
  - Assign probability to each $x_i \in Val(X_i)$
  - k- Independent parameters
- Consider $P(X_1,\ldots,X_n)$
  - How many independent parameters if $|Val(X_i)|=k$?

$k^n - 1$ parameters

What if variables are independent?

- What if variables are independent?
  - $(X_i \perp X_j), \forall\, i,j$ $\quad (\{X_1, X_3\} \perp \{X_7, X_9\})$
  - Not enough!!! (See homework 1 ☺) 1.4
  - Must assume that $(\mathbf{X} \perp \mathbf{Y}), \forall\, \mathbf{X},\mathbf{Y}$ subsets of $\{X_1,\ldots,X_n\}$
- Can write
  - $P(X_1,\ldots,X_n) = \prod_{i=1\ldots n} P(X_i)$
- How many independent parameters now?

$n \cdot (K-1)$ params

Conditional parameterization – two nodes

- Grade is determined by Intelligence

$P(I)$ $\quad I \quad$ $Val[I] = \{H, VH\}$

$\downarrow$

$P(G|I)$ $\quad G \quad$ $Val[G] = \{A+, A\}$

$P(I,G) = P(I) \cdot P(G|I)$

$P(I,G)$ =

I \ G	A	A+
H	0.01	.15
VH	0.04	0.8

$P(I)$ =

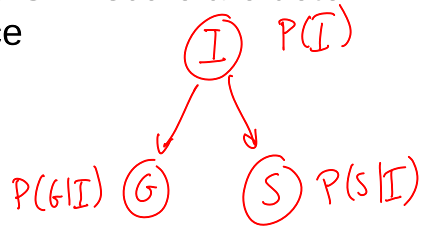
H	VH
.16	.84

$P(G|I)$ =

I \ G	A	A+
H	$\frac{0.01}{.016}$	$\frac{.15}{.16}$
V	$\frac{.04}{.84}$	$\frac{.8}{.84}$

Conditional parameterization – three nodes

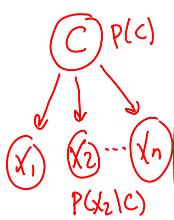
- Grade and SAT score are determined by Intelligence
- $(G \perp S \mid I)$



prob. of var given parents

The naïve Bayes model – Your first real Bayes Net

- Class variable: C
- Evidence variables: $X_1,\ldots,X_n$
- assume that $(\mathbf{X} \perp \mathbf{Y} \mid C)$, $\forall$ **X**,**Y** subsets of $\{X_1,\ldots,X_n\}$



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

Chain Rule:

$P(C, X_1, \ldots X_n) = P(C) \cdot P(X_1 \mid C) \cdot P(X_2 \mid C, X_1) \cdot P(X_3 \mid C, X_1, X_2) \cdots P(X_n \mid C, X_1 \ldots X_{n-1})$

$(X_1 \perp X_2 \mid C) \longrightarrow P(X_2 \mid C)$

$(X_3 \perp \{X_1, X_2\} \mid C) \longrightarrow P(X_3 \mid C)$

$(X_n \perp \{X_1, \ldots X_{n-1}\} \mid C) \longrightarrow P(X_n \mid C)$

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

What you need to know (From last class)

- Basic definitions of probabilities
- Independence
- Conditional independence
- The chain rule
- Bayes rule
- Naïve Bayes

Announcements

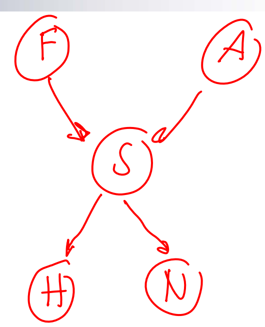
- Homework 1:
  - Out yesterday
  - Due September 27th – **beginning of class!**
  - It's hard – start early, ask questions
- Collaboration policy
  - OK to discuss in groups
  - Tell us on your paper who you talked with
  - Each person must write their **own unique paper**
  - No searching the web, papers, etc. for answers, we trust you want to learn
- Upcoming recitation
  - Monday 5:30-7pm in Wean 4615A – Matlab Tutorial
- Don't forget to register to the mailing list at:
  - https://mailman.srv.cs.cmu.edu/mailman/listinfo/10708-announce

This class

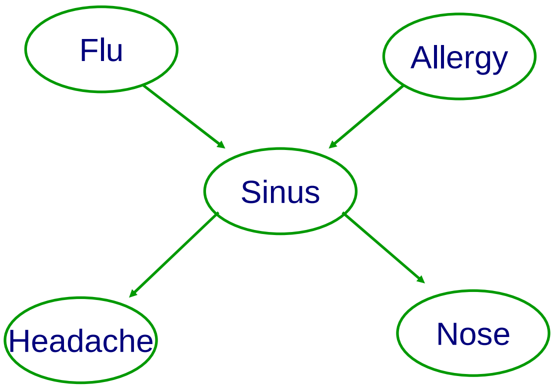
- We've heard of Bayes nets, we've played with Bayes nets, we've even used them in your research
- This class, we'll learn the semantics of BNs, relate them to independence assumptions encoded by the graph

Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

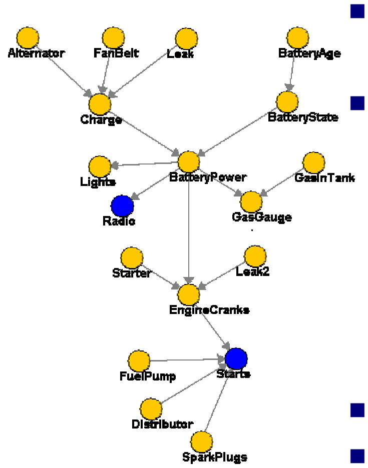


Possible queries



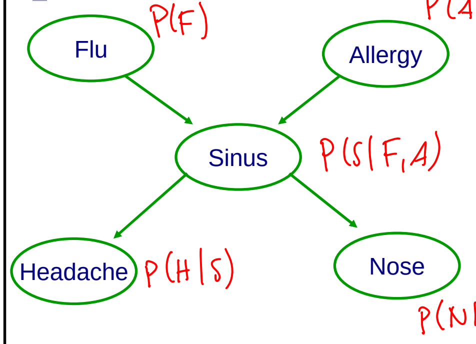
- Inference
  $P(F \mid H=t)$
- Most probable explanation
  $\max_{f,a,s} P(F=f, A=a, S=s \mid H=t, N=f)$
- Active data collection
  I say H=t
  tell me the value of N ?

Car starts BN



- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f)
- 2^{18} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution - Preview



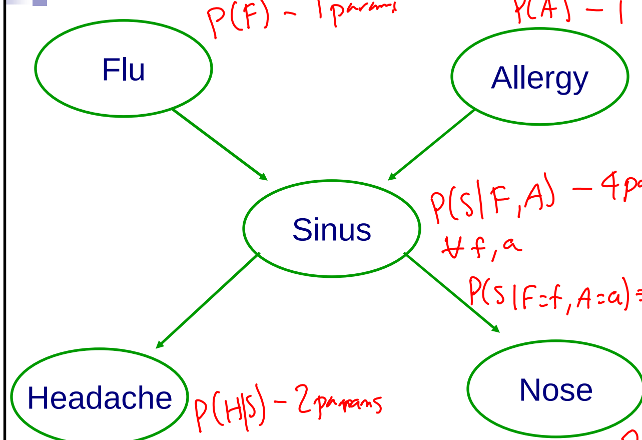
$P(F)$ $P(A)$ $P(S|F,A)$ $P(H|S)$ $P(N|S)$

$P(F,A,S,H,N)$

$= P(F) \cdot P(A) \cdot P(S|F,A)$

$P(H|S) \cdot P(N|S)$

Number of parameters (binary vars)



$P(F)$ – 1 param

$P(A)$ – 1

$P(S|F,A)$ – 4 params

$\forall f,a$

$P(S|F=f,A=a) \Rightarrow 1$ param

$P(H|S)$ – 2 params

$P(N|S)$ – 2 params

$\rightarrow$ 10 params.

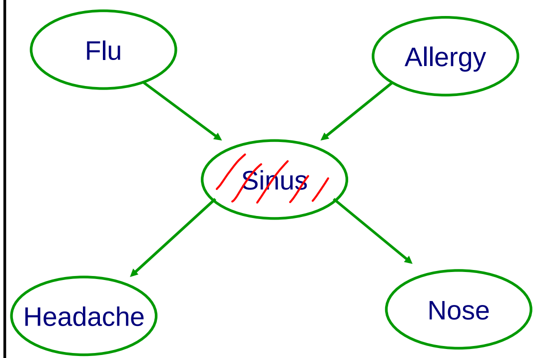
explicit representation

$P(F,A,S,H,N)$

$2^5 - 1 = 31$ params

!! :)

Key: Independence assumptions



$(F \perp A)$

not $(F \perp N)$

$(F \perp N \mid S)$

not $(H \perp N)$

$(H \perp N \mid S)$

Knowing sinus separates the variables from each other

Symptom

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A \Rightarrow P(F,A) = P(F) \cdot P(A)$

Flu = t	0.1
Flu = f	0.9

Allergy = t	0.2
Allergy = f	0.8

- More Generally:

	Flu = t	Flu = f
Allergy = t	$.1 \times .2$	$.2 \times .9$
Allergy = f		

Conditional independence

- Flu and Headache are not (marginally) independent

not $(F \perp H)$

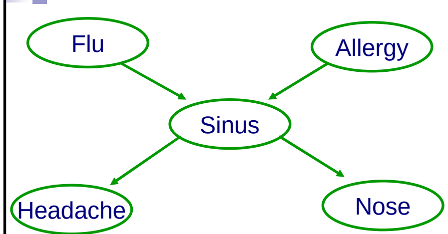
- Flu and Headache are independent given Sinus infection

$(F \perp H \mid S) \rightarrow P(F \mid S, H) = P(F \mid S)$

- More Generally:

$(X \perp Y \mid Z) \Rightarrow P(X \mid Y, Z) = P(X \mid Z)$

The independence assumption



Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parents
$(X_i \perp NonDescendants_{X_i} \mid Pa_{X_i})$

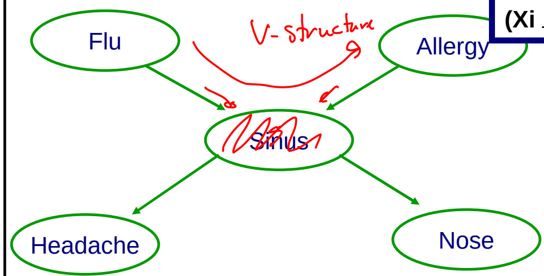
$(F \perp A \mid \phi) \equiv (F \perp A)$

$(N \perp \{F, A, H\} \mid S)$

Explaining away

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parents
($X_i \perp NonDescendants_{X_i} \mid Pa_{X_i}$)

V-structure



$(F \perp A)$

$P(F=t) \leq P(F=t \mid S=t)$

$P(A=t) \leq P(A=t \mid S=t)$

$P(A=t) \leq P(A=t \mid S=t, F=t) \leq P(A=t \mid S=t)$

$(F \perp A)$

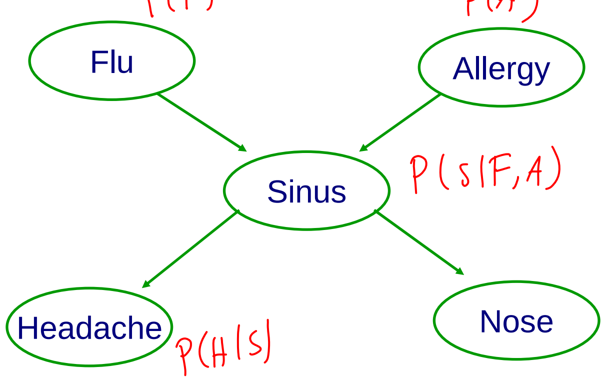
not $(F \perp A \mid S)$ S was the center of a V-structure

not $(F \perp A \mid H)$ H was the descendent of " "

What about probabilities? discrete settin
Conditional probability tables (CPTs)

Also called "Factors": in a BN a CPT is a factor

$P(F)$ $P(A)$ CPTs



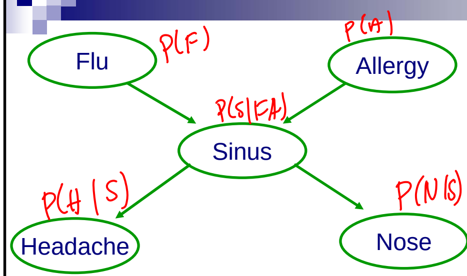
$P(S \mid F, A)$

$P(H \mid S)$

$P(N \mid S)$

Joint distribution

→ Chain rule
→ local Markov Assumption



$P(F,A,S,H,N) =$

$P(A) \cdot P(F) \cdot P(S|FA) \cdot P(H|S) \cdot P(N|S)$

chain rule

$= P(F) \cdot P(A|F) \cdot P(S|FA) \cdot P(H|SFA) \cdot P(N|SFA,H)$

local Markov

$= P(F) \cdot P(A) \, P(S|F,A) \cdot P(H|S) \cdot P(N|S)$

$(F \perp A) \Rightarrow P(A|F) = P(A)$

$(H \perp \{F,A\} | S) \Rightarrow P(H|SFA) = P(H|S)$

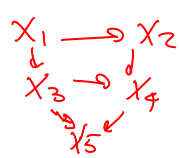
$(N \perp \{F,A,H\} | S) \Rightarrow P(N|F,A,S) = P(N|S)$

local

Why can we decompose? Markov Assumption!

A general Bayes net

- Set of random variables $\{X_1, \ldots, X_n\}$
- Directed acyclic graph DAG



- CPTs $P(X_i | Pa_{X_i})$; $Pa_{X_i} \subseteq \{X_1, \ldots, X_n\}$
- Joint distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

(almost)

- **Local Markov Assumption:**
  - A variable X is independent of its non-descendants given its parents – **(Xi ⊥ NonDescendantsXi | PaXi)**

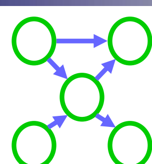
  (and only its parents)

Questions????

- What distributions can be represented by a BN?
- What BNs can represent a distribution?
- What are the independence assumptions encoded in a BN?
  - in addition to the local Markov assumption

Today: The Representation Theorem – Joint Distribution to BN

BN:



G

Encodes independence assumptions

$I_\ell(G)$

If conditional independencies in BN are subset of conditional independencies in *P*

Obtain

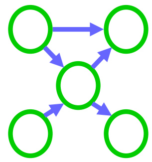
Joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

$I_\ell(G) \subseteq I(P)$ ✓ indep. in P

Today: The Representation Theorem – BN to Joint Distribution

BN:



Encodes independence assumptions

If joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Obtain

Then conditional independencies in BN are subset of conditional independencies in *P*

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

Let's start proving it for naïve Bayes – From joint distribution to BN

- Independence assumptions:
  - X_i independent given C
- Let's assume that *P* satisfies independencies must prove that *P* factorizes according to BN:
  - $P(C,X_1,\ldots,X_n) = P(C) \prod_i P(X_i|C)$
- Use chain rule!

$P(C, X_1 \ldots X_n) = P(C) \cdot P(X_1|C) \; P(X_2|X_1,C) \cdots$

$= P(C) \cdot P(X_1|C) \; P(X_2|C) \cdots$

Let's start proving it for naïve Bayes – From BN to joint distribution 1

- Let's assume that *P* factorizes according to the BN:
  - $P(C,X_1,\ldots,X_n) = P(C) \prod_i P(X_i|C)$
- Prove the independence assumptions:
  - X_i independent given C
  - Actually, $(\mathbf{X} \perp \mathbf{Y} \mid C)$, $\forall$ **X**,**Y** subsets of $\{X_1,\ldots,X_n\}$

$P(X|Y,C) = P(X|C)$ equiv. $P(X,Y|C) = P(X|C) \cdot P(Y|C)$

$\mathbf{X} = \{X_1, X_2\}$ $\quad \mathbf{Y} = \{X_3, X_4\}$ $\quad P(X_1 \ldots X_4|C) = P(X_1X_2|C) \cdot P(X_3X_4|C)$

$$P(X_1 \cdots X_4|C) = \frac{\cancel{P(C)} \cdot P(X_1|C) \cdots P(X_4|C)}{\cancel{P(C)}} = \underbrace{P(X_1|C) \cdot P(X_2|C)}_{P(X_1,X_2|C)} \cdot \underbrace{P(X_3|C) \cdot P(X_4|C)}_{P(X_3,X_4|C)}$$

$$P(X_1X_2|C) = \sum_{X_3X_4} P(X_1 \cdots X_4|C)$$

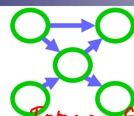
$\Rightarrow (\{X_1X_2\} \perp \{X_3X_4\} \mid C)$ q.e.d.

Let's start proving it for naïve Bayes – From BN to joint distribution 2

- Let's consider a simpler case
  - Grade and SAT score are determined by Intelligence
  - $P(I,G,S) = P(I)P(G|I)P(S|I)$
  - Prove that $P(G,S|I) = P(G|I)\, P(S|I)$

Today: The Representation Theorem

BN:



Encodes independence assumptions

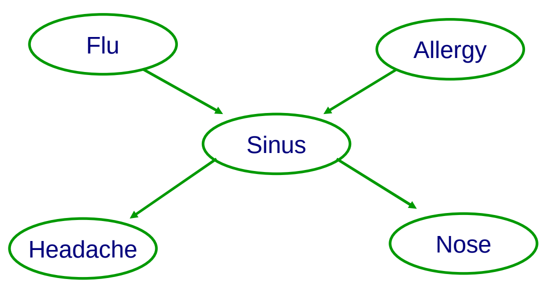
I-map: G is an I-map for P

If conditional independencies in BN are subset of conditional independencies in *P*	Obtain	Joint probability distribution:
$I_\ell(G) \subseteq I(P)$		$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$

If joint probability distribution:	Obtain	Then conditional independencies in BN are subset of conditional independencies in *P*
$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$		

Local Markov assumption & I-maps

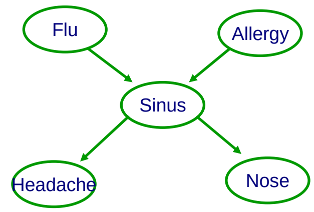
- Assumptions
  Local independence assumptions in BN structure G: $I_\ell(G)$
- Truth:
  Independence assertions of *P*: $I(P)$
- BN structure G is an ***I-map*** (independence map) for P if: $I_\ell(G) \subseteq I(P)$



Local Markov Assumption:
A variable X is independent of its non-descendants given its parents
($X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}$)

Factorized distributions

- Given
  - Random vars $X_1,\ldots,X_n$
  - *P* distribution over vars
  - BN structure *G* over same vars
- *P* factorizes according to *G* if



$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

BN Representation Theorem – I-map to factorization

If conditional independencies in BN are subset of conditional independencies in *P*

Obtain

Joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

G* is an I-map of *P

P* factorizes according to *G

BN Representation Theorem – I-map to factorization: **Proof**

G* is an I-map of *P

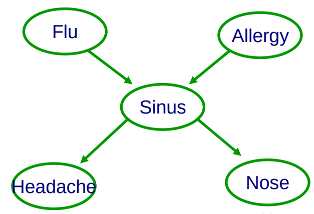
Obtain

P* factorizes according to *G

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

ALL YOU NEED:

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents
($X_i \perp$ NonDescendants$_{X_i}$ | Pa$_{X_i}$)



Defining a BN

- Given a set of variables and conditional independence assertions of P
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, $\mathbf{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov assumption holds – X_i independent of rest of $\{X_1,\ldots,X_{i-1}\}$, given parents $\mathbf{Pa}_{X_i}$
  - Define/learn CPT – $P(X_i \mid \mathbf{Pa}_{X_i})$

BN Representation Theorem – Factorization to I-map

If joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Obtain

Then conditional independencies in BN are subset of conditional independencies in *P*

P* factorizes according to *G

G* is an I-map of *P

BN Representation Theorem – Factorization to I-map: **Proof**

If joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Obtain

Then conditional independencies in BN are subset of conditional independencies in *P*

P* factorizes according to *G

G* is an I-map of *P

Homework 1!!!! ☺

The BN Representation Theorem

If conditional independencies in BN are subset of conditional independencies in *P* → **Obtain** → Joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Important because:
Every *P* has at least one BN structure *G*

If joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

→ **Obtain** → Then conditional independencies in BN are subset of conditional independencies in *P*

Important because:
Read independencies of *P* from BN structure *G*

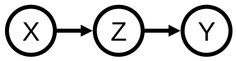
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp \text{NonDescendants}_{Xi} \mid \mathbf{Pa}_{Xi})$
- But then we talked about other (in)dependencies
  - e.g., explaining away
- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!

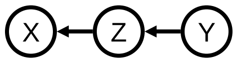
Understanding independencies in BNs – BNs with 3 nodes

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents

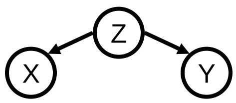
Indirect causal effect:



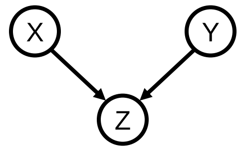
Indirect evidential effect:



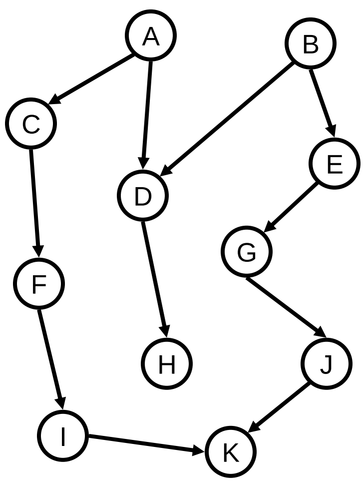
Common cause:



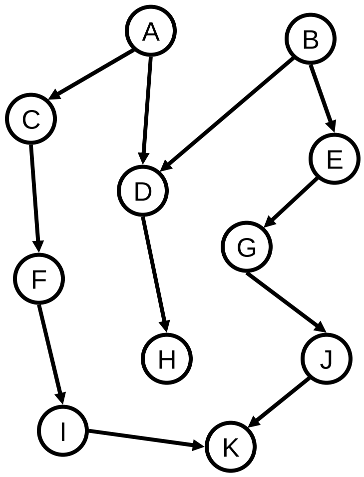
Common effect:



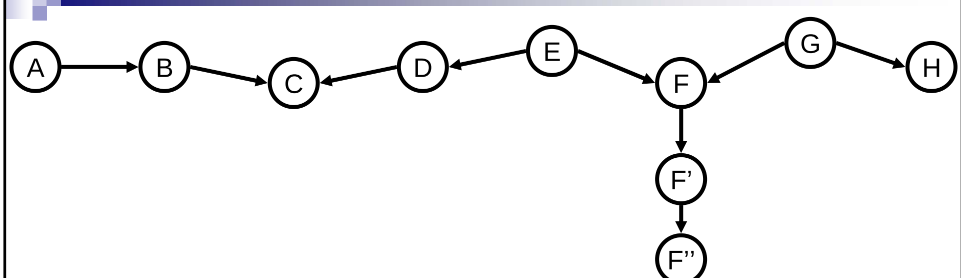
Understanding independencies in BNs – Some examples



Understanding independencies in BNs – Some more examples



An active trail – Example



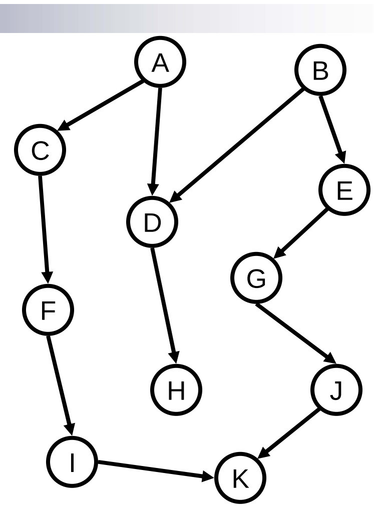
When are A and H independent?

Active trails formalized

- A trail $X_1 - X_2 - \cdot\cdot\cdot - X_k$ is an **active trail** when variables $\boldsymbol{O} \subseteq \{X_1,\ldots,X_n\}$ are observed if for each consecutive triplet in the trail:
  - $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, and X_i **is observed** ($X_i \in \boldsymbol{O}$), or **one of its descendents**

Active trails and independence?

- **Theorem**: Variables $\mathbf{X_i}$ **and** $\mathbf{X_j}$ **are independent given** $\boldsymbol{Z} \subseteq \{X_1,\ldots,X_n\}$ if the is **no active trail** between X_i and X_j when variables $\boldsymbol{Z} \subseteq \{X_1,\ldots,X_n\}$ are observed

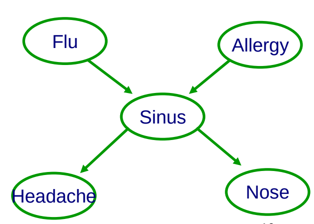


More generally: Soundness of d-separation

- Given BN structure *G*
- Set of independence assertions obtained by d-separation:
  - $I(G) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : \text{d-sep}_G(\mathbf{X};\mathbf{Y}|\mathbf{Z})\}$
- **Theorem: Soundness of d-separation**
  - If *P* factorizes over *G* then $I(G) \subseteq I(P)$
- **Interpretation:** d-separation only captures true independencies
- Proof discussed when we talk about undirected models

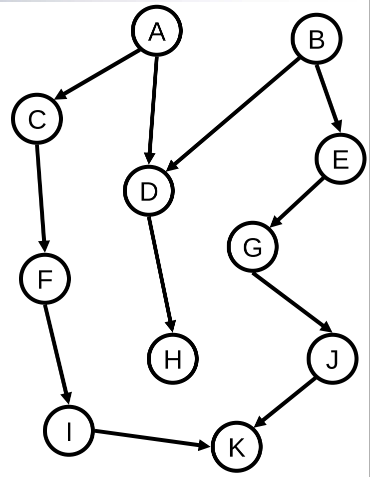
Adding edges doesn't hurt

- **Theorem**: Let **G** be an I-map for ***P***, any DAG **G'** that includes the same directed edges as **G** is also an I-map for ***P***.
- **Proof sketch**:



Existence of dependency when not d-separated

- **Theorem:** If X and Y are not d-separated given **Z**, then X and Y are dependent given **Z** under some *P* that factorizes over *G*
- **Proof sketch**:
  - Choose an active trail between X and Y given **Z**
  - Make this trail dependent
  - Make all else uniform (independent) to avoid "canceling" out influence



More generally: Completeness of d-separation

- **Theorem: Completeness of d-separation**
  - For "almost all" distributions that *P* factorize over to *G*, we have that I(*G*) = I(*P*)
  - *"almost all" distributions*: except for a set of measure zero of parameterizations of the CPTs (assuming no finite set of parameterizations has positive measure)
- **Proof sketch:**

Interpretation of completeness

- **Theorem: Completeness of d-separation**
  - For "almost all" distributions that P factorize over to G, we have that $I(G) = I(P)$
- BN graph is usually sufficient to capture all independence properties of the distribution!!!!
- But only for complete independence:
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}),\ \forall\ \mathbf{x} \in Val(\mathbf{X}),\ \mathbf{y} \in Val(\mathbf{Y}),\ \mathbf{z} \in Val(\mathbf{Z})$
- Often we have context-specific independence (CSI)
  - $\exists\ \mathbf{x} \in Val(\mathbf{X}),\ \mathbf{y} \in Val(\mathbf{Y}),\ \mathbf{z} \in Val(\mathbf{Z})$: $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z})$
  - Many factors may affect your grade
  - But if you are a frequentist, all other factors are irrelevant ☺

What you need to know

- Independence & conditional independence
- Definition of a BN
- The representation theorems
  - Statement
  - Interpretation
- d-separation and independence
  - soundness
  - existence
  - completeness

Acknowledgements

- JavaBayes applet
  - http://www.pmr.poli.usp.br/ltd/Software/javabayes/Home/index.html