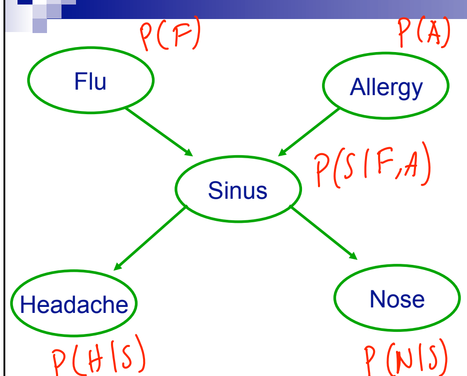
Readings:
K&F: 3.1, 3.2, 3.3.1, 3.3.2

BN Semantics 2 –
Representation Theorem
The revenge of d-separation

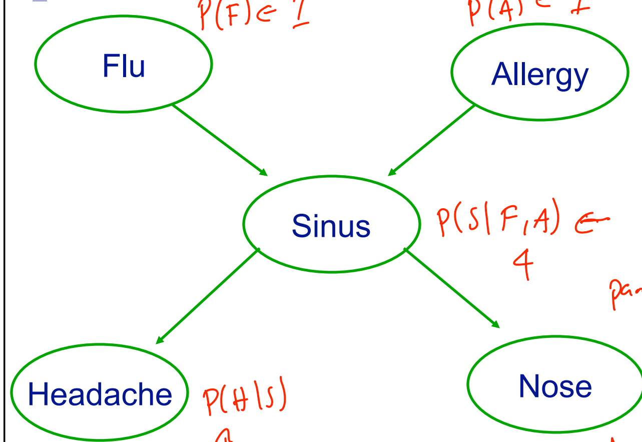
Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
September 17th, 2008

Factored joint distribution - Preview

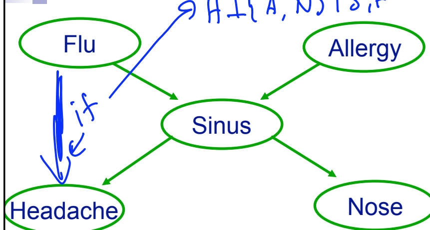


$P(F,A,S,H,N) =$
$P(F) \cdot P(A) \cdot P(S|F,A)$
$P(H|S) \cdot P(N|S)$

Number of parameters

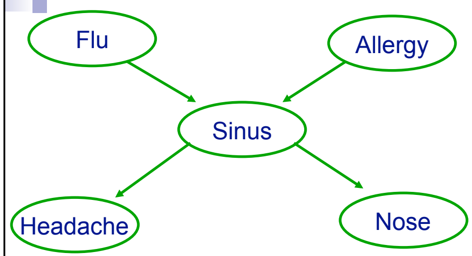


The independence assumption



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

Joint distribution



Why can we decompose? Local Markov Assumption!

A general Bayes net

- Set of random variables
- Directed acyclic graph
- CPTs
- Joint distribution:

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

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

Questions????

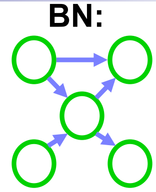
- What distributions can be represented by a BN?
- What BNs can represent a distribution?
- What are the independence assumptions encoded in a BN?
  - in addition to the local Markov assumption

Independencies in Problem

World, Data, reality:

True distribution *P* contains independence assertions

BN:

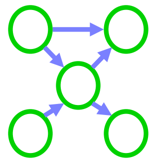


Graph *G* encodes local independence assumptions

Key Representational Assumption:

Today: The Representation Theorem – True Independencies to BN Factorization

BN:



Encodes independence assumptions

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

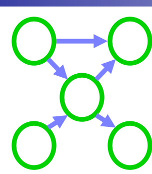
Obtain

Joint probability distribution:

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

Today: The Representation Theorem – BN Factorization to True Independencies

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*

Let's start proving it for naïve Bayes – From True Independencies to BN Factorization

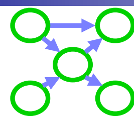
- Independence assumptions:
  - X_i independent given C
- Let's assume that *P* satisfies independencies must prove that *P* factorizes according to BN:
  - $P(C,X_1,\ldots,X_n) = P(C) \prod_i P(X_i|C)$
- Use chain rule!

Let's start proving it for naïve Bayes – From BN Factorization to True Independencies

- 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 $\{X_1,\ldots,X_n\}$

Today: The Representation Theorem

BN:



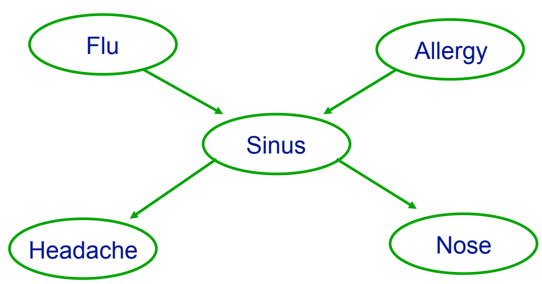
Encodes independence assumptions

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

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***

Local Markov assumption & I-maps

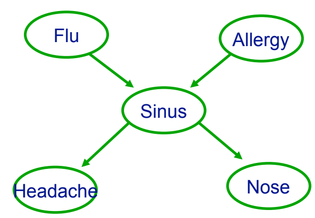
- Local independence assumptions in BN structure G:
- Independence assertions of *P*:
- BN structure G is an ***I-map*** (independence map) if:



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

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, part 1**

G* is an I-map of *P

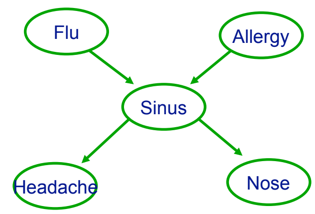
Obtain

P* factorizes according to *G

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

Topological Ordering:

- Number variables such that:
  - parent has lower number than child
  - i.e., $X_i \rightarrow X_j \Rightarrow i<j$
  - ***Key: variable has lower number than all of its***
- DAGs always have (many) topological orderings
  - find by a modification of breadth first search



BN Representation Theorem – I-map to factorization: **Proof, part 2**

G* is an I-map of *P

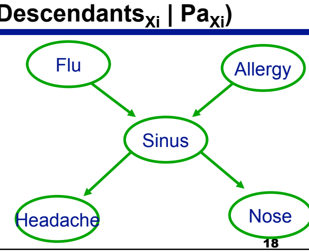
Obtain

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 and only its parents
($X_i \perp NonDescendants_{X_i} \mid Pa_{X_i}$)

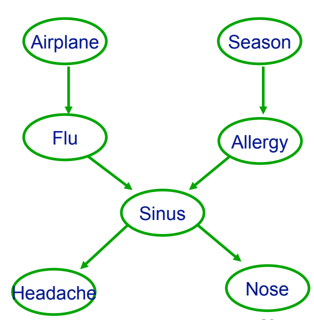


Defining a BN

- 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 $\mathbf{Pa}_{Xi}$
  - Define/learn CPT – $P(X_i| \mathbf{Pa}_{Xi})$

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***.
  - **Corollary 1**: __ is strictly more expressive than ___
  - **Corollary 2**: If G is an I-map for P, then adding edges still an I-map
- **Proof**:



Announcements

- Homework 1:
  - Out today
  - Due in 2 weeks – **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
- Audit policy
  - No sitting in, official auditors only, see course website
- Recitation tomorrow
  - Wean 5409, 5pm

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*

What you need to know thus far

- Independence & conditional independence
- Definition of a BN
- Local Markov assumption
- The representation theorems
  - Statement: *G* is an I-map for *P* if and only if *P* factorizes according to *G*
  - Interpretation

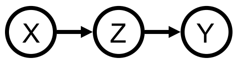
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp NonDescendants_{Xi} \mid \mathbf{Pa}_{Xi})$
- 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!!!

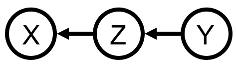
Understanding independencies in BNs – BNs with 3 nodes

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

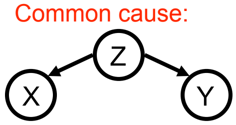
Indirect causal effect:



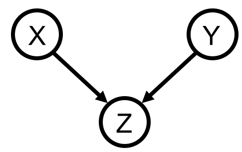
Indirect evidential effect:



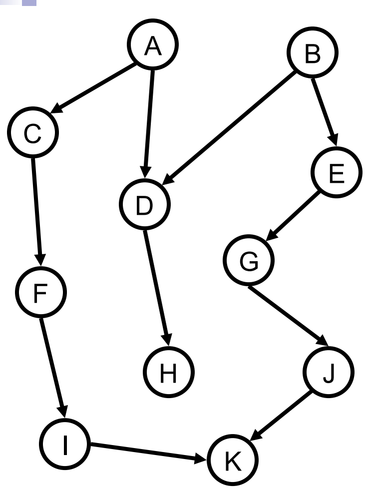
Common cause:



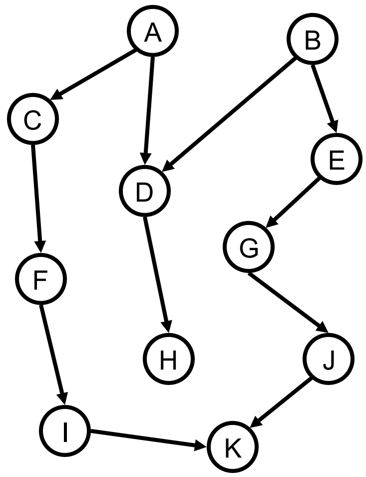
Common effect:



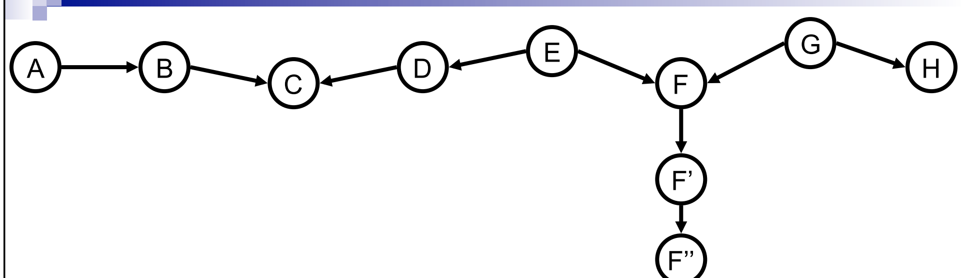
Understanding independencies in BNs – Some examples



Understanding independencies in BNs – Some more examples



An active trail – Example



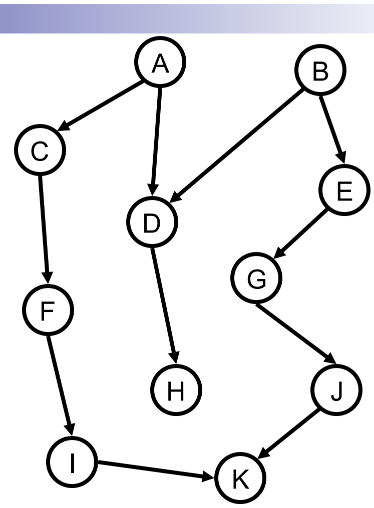
When are A and H independent?

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

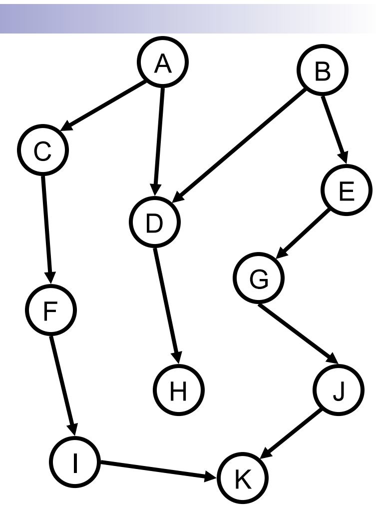


More generally: Soundness of d-separation

- 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

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

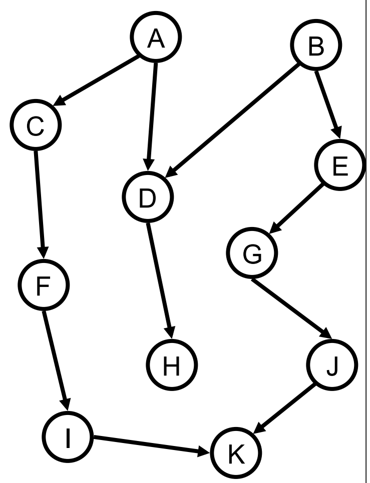
- **Theorem: Completeness of d-separation**
  - For "almost all" distributions where *P* factorizes over to *G*, we have that I(*G*) = I(*P*)
    - *"almost all" distributions*: except for a set of measure zero of parameterizations of the CPTs (assuming no finite set of parameterizations has positive measure)
    - Means that if all sets **X** & **Y** that are not d-separated given **Z**, then ¬(**X**⊥**Y**|**Z**)
- **Proof sketch for very simple case:**

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:
  - $P \Rightarrow (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}), \forall\, \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$
- Often we have context-specific independence (CSI)
  - $\exists\, \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z}): P \Rightarrow (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z})$
  - Many factors may affect your grade
  - But if you are a frequentist, all other factors are irrelevant ☺

Algorithm for d-separation

- How do I check if X and Y are d-separated given **Z**
  - There can be exponentially-many trails between X and Y
- Two-pass linear time algorithm finds all d-separations for X
- 1. Upward pass
  - Mark descendants of **Z**
- 2. Breadth-first traversal from X
  - Stop traversal at a node if trail is "blocked"
  - (Some tricky details apply – see reading)



What you need to know

- d-separation and independence
  - sound procedure for finding independencies
  - existence of distributions with these independencies
  - (almost) all independencies can be read directly from graph without looking at CPTs