Graphical Models

Aarti Singh
Slides Courtesy: Carlos Guestrin

Machine Learning 10-701/15-781
Nov 10, 2010

Recitation

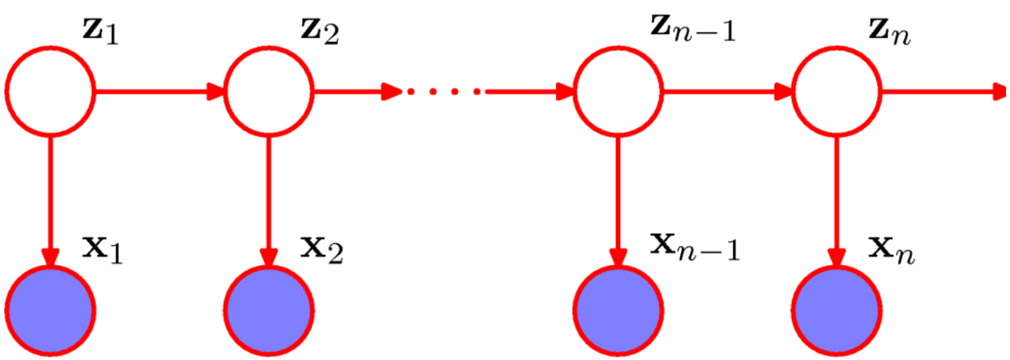
- HMMs & Graphical Models
- Strongly recommended!!
- Place: NSH 1507 (Note)
- Time: 5-6 pm

Min

iid to dependent data

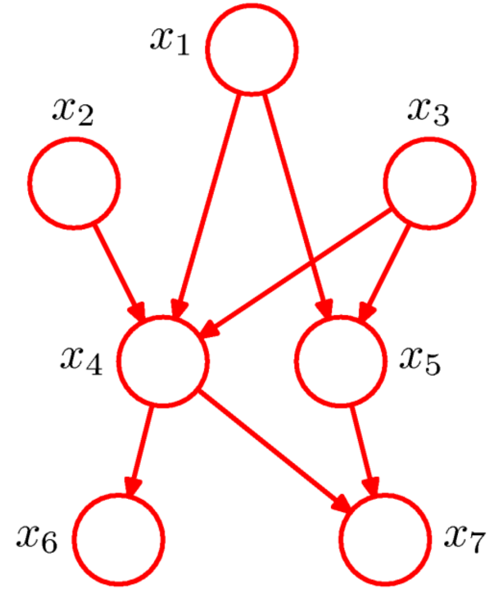
HMM

- sequential dependence



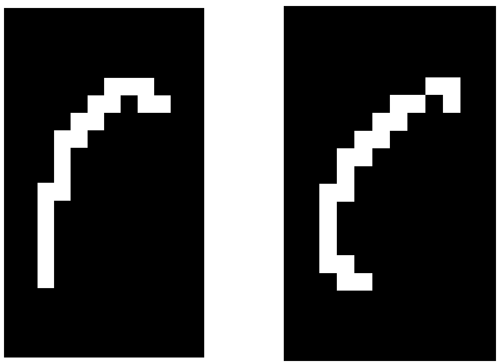
Graphical Models

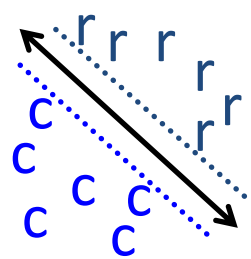
- general dependence

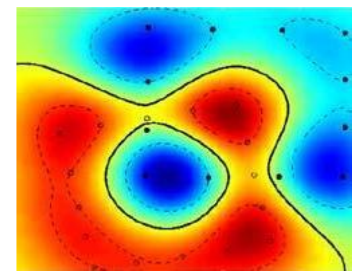


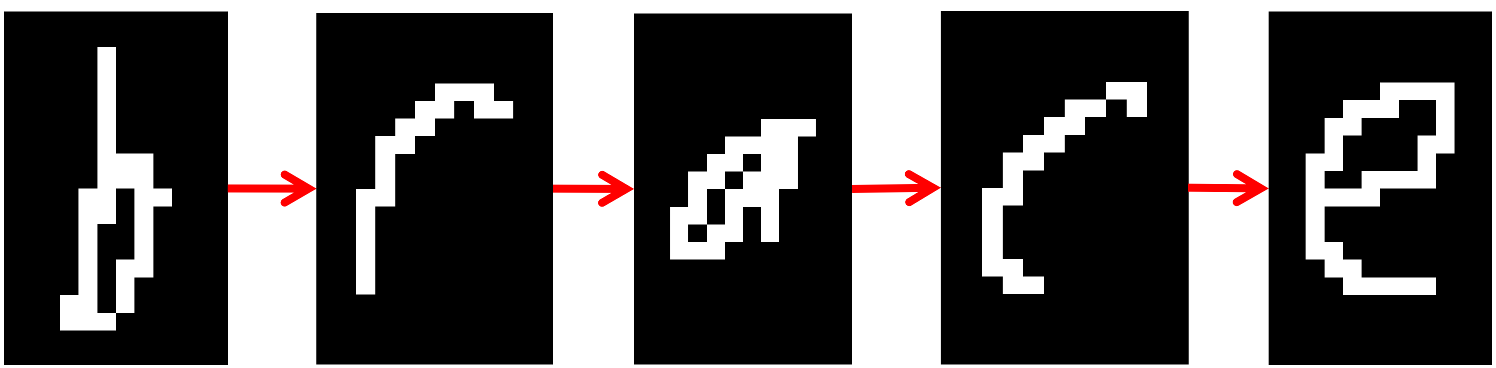
Applications

- Character recognition, e.g., kernel SVMs





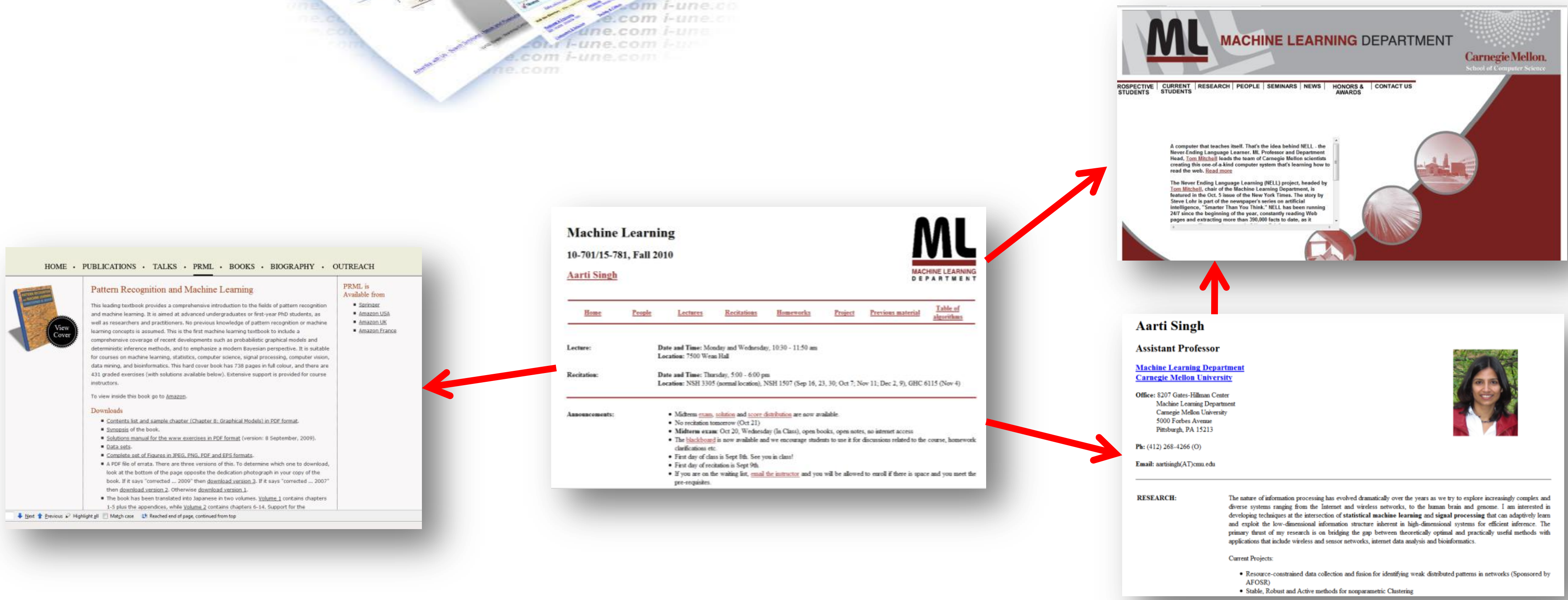




Applications

- Webpage Classification

Sports
Science
News



Applications

- Speech recognition
- Diagnosis of diseases
- Study Human genome
- Robot mapping
- Modeling fMRI data
- Fault diagnosis
- Modeling sensor network data
- Modeling protein-protein interactions
- Weather prediction
- Computer vision
- Statistical physics
- Many, many more …

Graphical Models

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
  - Graphical models express sets of conditional independence assumptions via graph structure
  - Graph structure plus associated parameters define *joint probability distribution over set of variables/nodes*

- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

Topics in Graphical Models

- Representation
  - Which joint probability distributions does a graphical model represent?
- Inference
  - How to answer questions about the joint probability distribution?
    - Marginal distribution of a node variable
    - Most likely assignment of node variables
- Learning
  - How to learn the parameters and structure of a graphical model?

Conditional Independence

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

probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

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

- Also to:

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

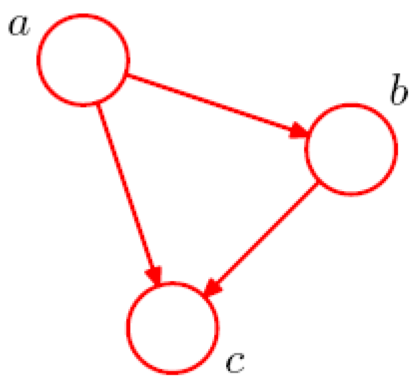
Directed - Bayesian Networks

- Representation
  - Which joint probability distributions does a graphical model represent?

For any arbitrary distribution,

Chain rule:

$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$



More generally:

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1},\ldots,X_1)$$

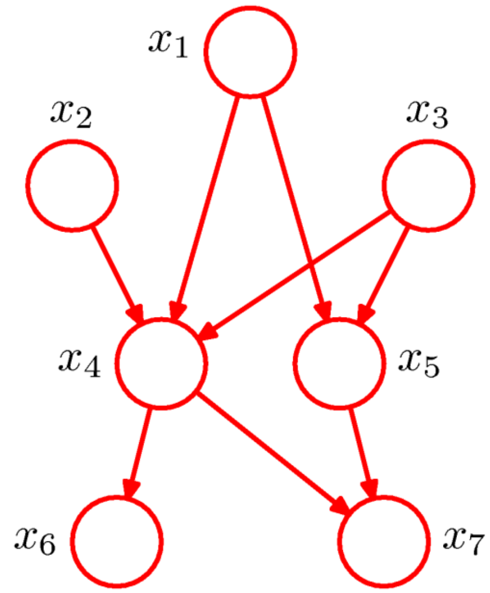
Fully connected directed graph between $X_1, \ldots, X_n$

Directed - Bayesian Networks

- Representation
  - Which joint probability distributions does a graphical model represent?

Absence of edges in a graphical model conveys useful information.

$$p(x_1, x_2, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



Directed - Bayesian Networks

- Representation
  - Which joint probability distributions does a graphical model represent?

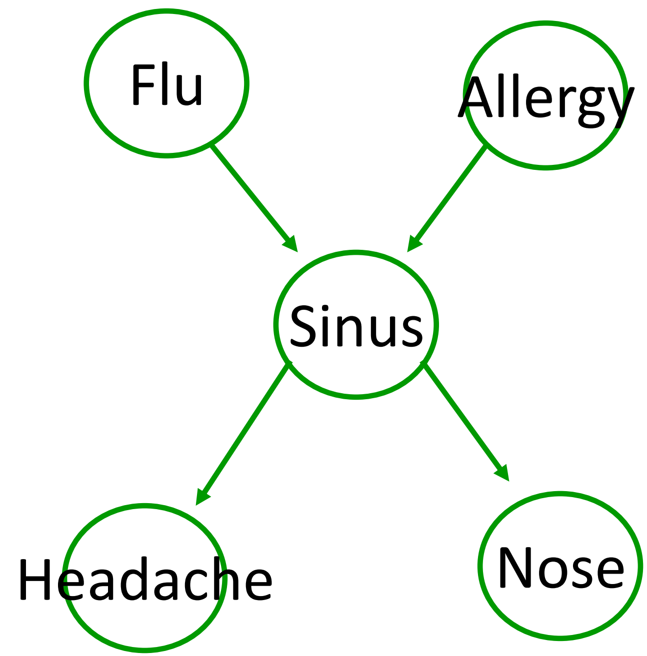
BN is a directed acyclic graph (DAG) that provides a compact representation for joint distribution

$$p(\mathbf{x}) = \boxed{\prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)}$$

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

Bayesian Networks Example

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches



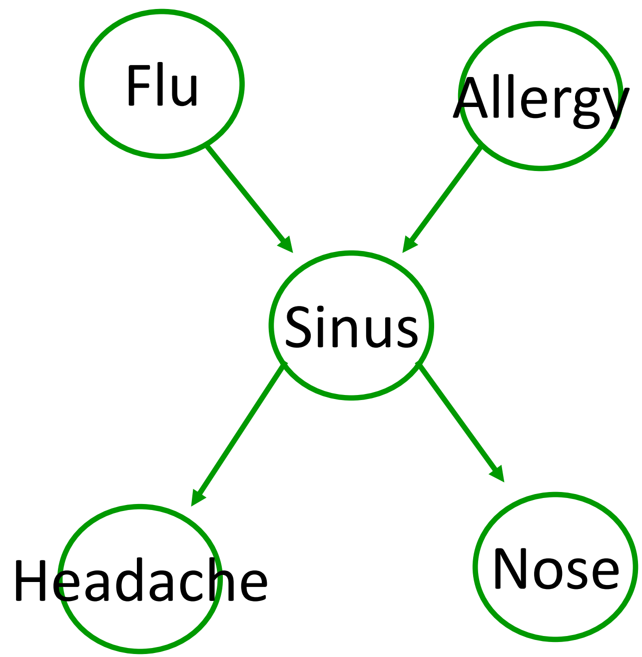
- Causal Network

- Local Markov Assumption: If you have no sinus infection, then flu has no influence on headache (flu causes headache but only through sinus)

Markov independence assumption

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

	parents	non-desc	assumption
S	F,A	-	-
H	S	F,A,N	$H \perp \{F,A,N\} \mid S$
N	S	F,A,H	$N \perp \{F,A,H\} \mid S$
F	-	A	$F \perp A$
A	-	F	$A \perp F$



Markov independence assumption

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

Joint distribution:

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

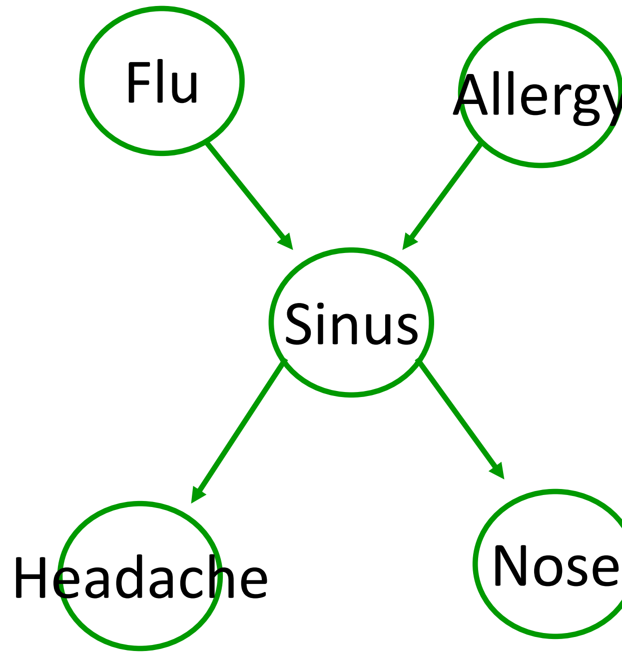
$= P(F)\ P(F|A)\ P(S|F,A)\ P(H|S,F,A)\ P(N|S,F,A,H)$

Chain rule

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

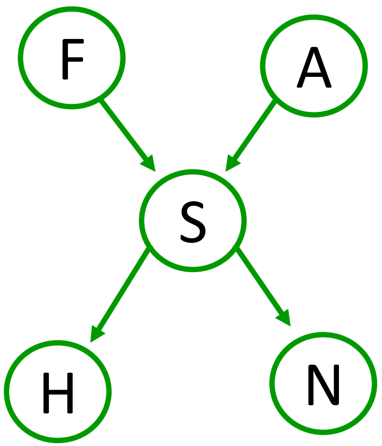
Markov Assumption

$F \perp A, \quad H \perp \{F,A\}|S, \quad N \perp \{F,A,H\}|S$



How many parameters in a BN?

- Discrete variables $X_1, \ldots, X_n$
- Directed Acyclic Graph (DAG)
  - Defines parents of X_i, $\mathbf{Pa}_{X_i}$
- CPTs (Conditional Probability Tables)
  - $P(X_i \mid \mathbf{Pa}_{X_i})$



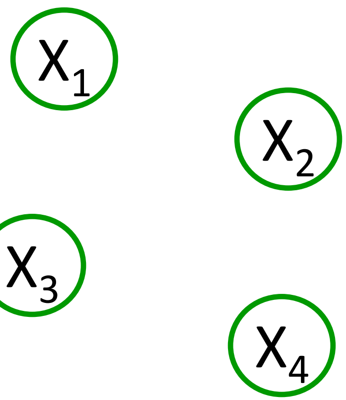
E.g. $X_i = S$, $\mathbf{Pa}_{X_i} = \{F, A\}$

	F=f, A=f	F=t, A=f	F=f, A=t	F=t,A=t
S=t	0.9	0.8	0.7	0.3
S=f	0.1	0.2	0.3	0.7

n variables, K values, max d parents/node $O(nK \times K^d)$

Two (trivial) special cases

Fully disconnected graph



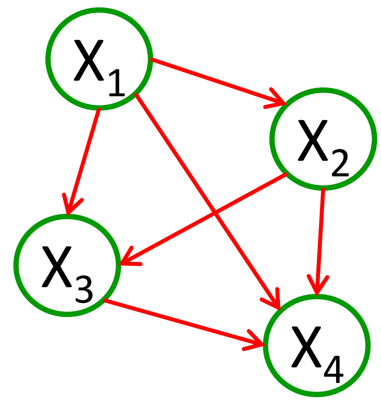
X_i

parents: ϕ

non-descendants: $X_1,\ldots,X_{i-1}, X_{i+1},\ldots, X_n$

$X_i \perp X_1,\ldots,X_{i-1},X_{i+1},\ldots, X_n$

Fully connected graph



X_i

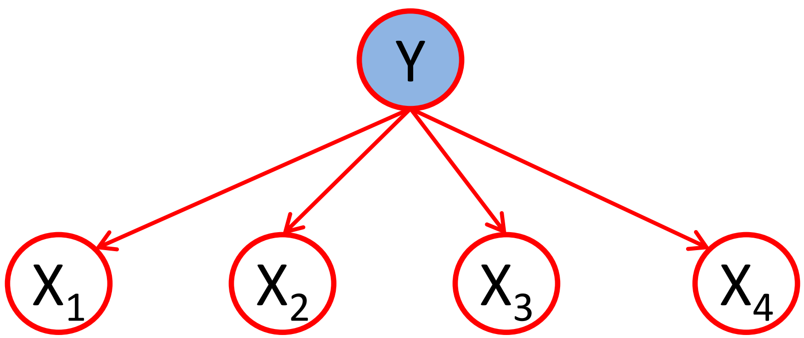
parents: $X_1, \ldots, X_{i-1}$

non-descendants: ϕ

No independence assumption

Bayesian Networks Example

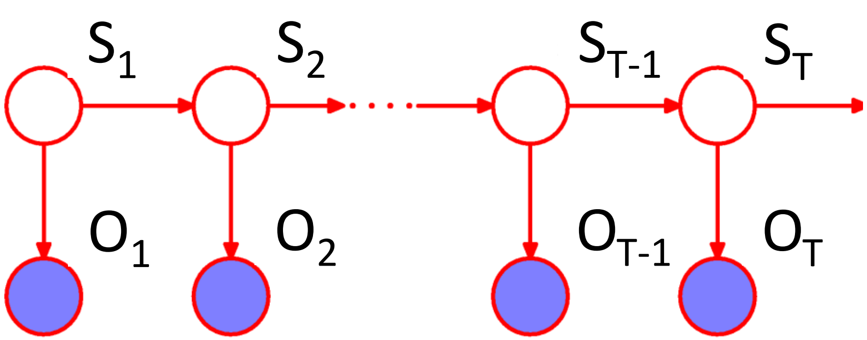
- Naïve Bayes $X_i \perp X_1,\ldots,X_{i-1},X_{i+1},\ldots, X_n | Y$



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

$P(Y)P(X_1|Y)\ldots P(X_1|Y)$

- HMM



$$p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) \quad = \\ p(S_1) \prod_{t=2}^T p(S_t|S_{t-1}) \prod_{t=1}^T p(O_t|S_t)$$

Explaining Away

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

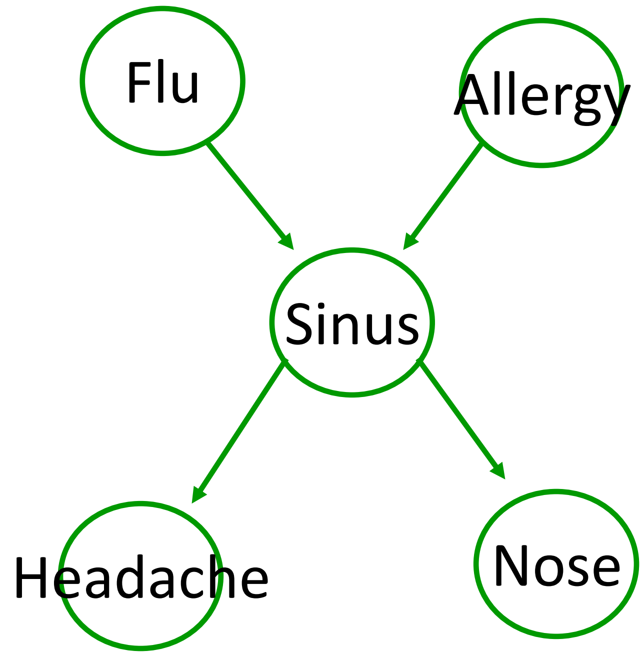
$F \perp A$ $P(F|A=t) = P(F)$

$F \perp A | S$?
$P(F|A=t,S=t) = P(F|S=t)$? **No!**

$P(F=t|S=t)$ is high,
but $P(F=t|A=t,S=t)$ not as high
since $A = t$ explains away $S=t$

Infact, $P(F=t|A=t,S=t) < P(F=t|S=t)$

$F \perp A | N$? **No!**



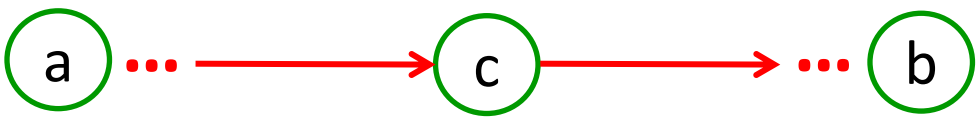
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp \text{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i})$
- But then we talked about other (in)dependencies
  - e.g., explaining away

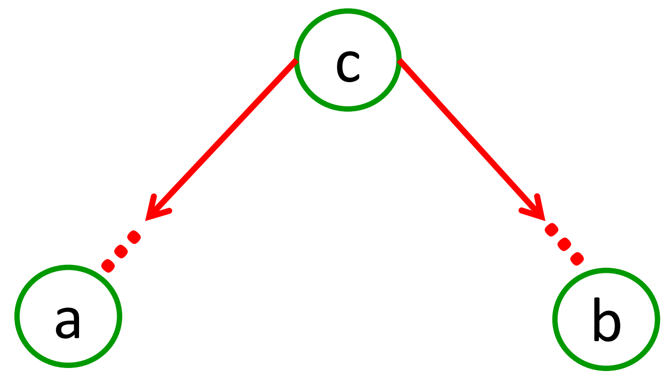
- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!

D-separation

- a is D-separated from b by c ≡ a ⊥ b|c
- Three important configurations

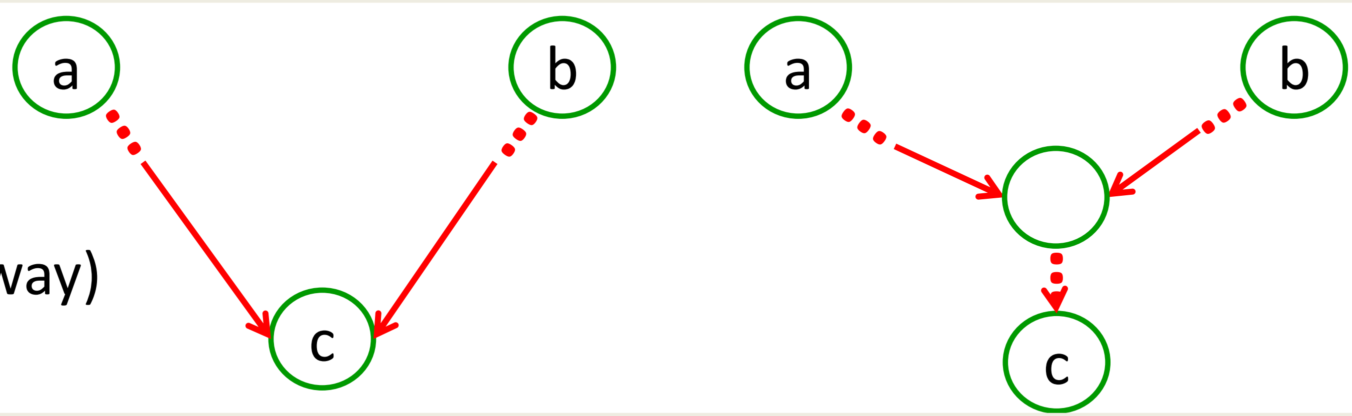


Causal direction



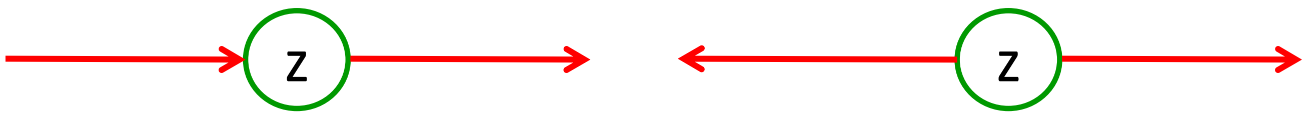
Common cause

V-structure
(Explaining away)

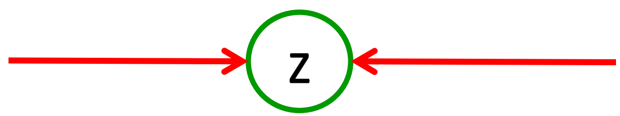


D-separation

- A, B, C – non-intersecting set of nodes
- A is D-separated from B by $C \equiv A \perp B|C$
  if all paths between nodes in A & B are "blocked"
  i.e. path contains a node z such that either



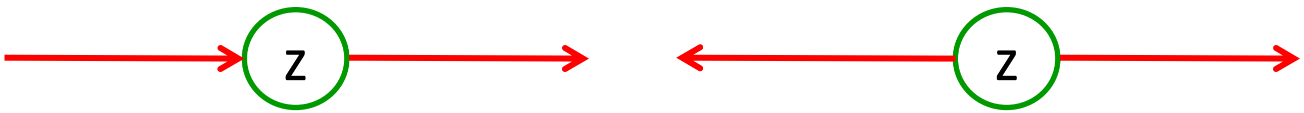
and z in C, OR



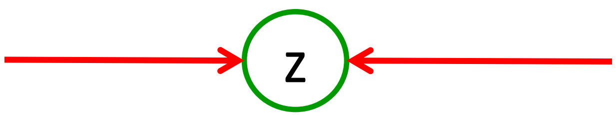
and neither z nor any of its descendants is in C.

D-separation Example

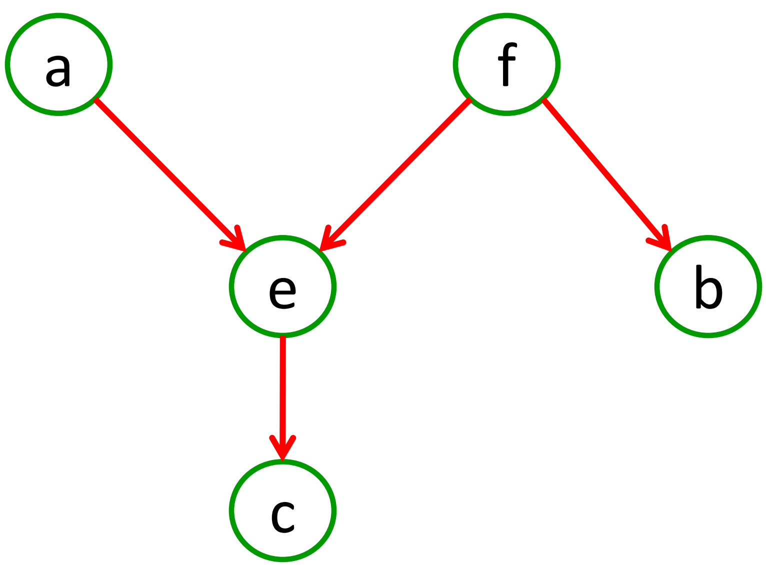
A is D-separated from B by C if every path between A and B contains a node z such that either



And z in C

or 

And neither z nor its descendants are in C



$a \perp b \mid f$?

Yes, Consider z = f or z = e

$a \perp b \mid c$?

No, Consider z = e

Representation Theorem

- Set of distributions that factorize according to the graph - F

- Set of distributions that respect conditional independencies implied by d-separation properties of graph – I

$$F \Rightarrow I$$

Important because: **Given independencies of *P* can get BN structure *G***

$$I \Leftarrow F$$

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

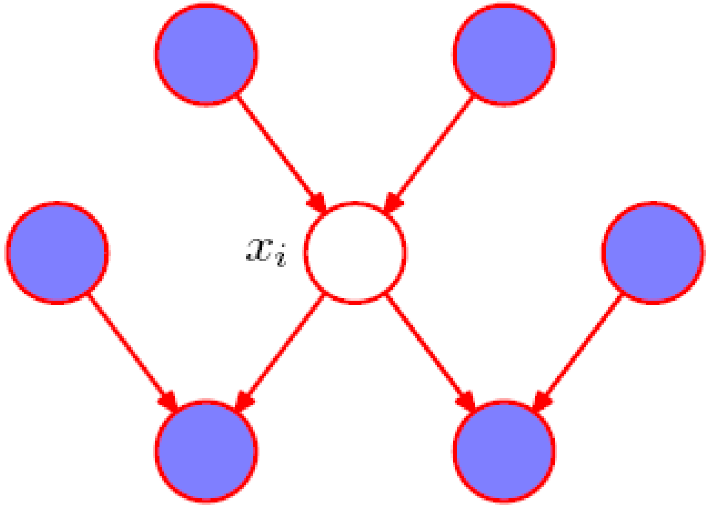
Markov Blanket

- Conditioning on the Markov Blanket, node i is independent of all other nodes.

$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(x_1,\ldots,x_n)}{\sum_i p(x_1,\ldots,x_n)} = \frac{\prod_k p(x_k|pa(x_k))}{\sum_i \prod_k p(x_k|pa(x_k))}$$

Only terms that remain are the ones which involve i

$$p(x_i|pa(x_i)) \qquad p(x_k|pa(x_k) \ni i)$$



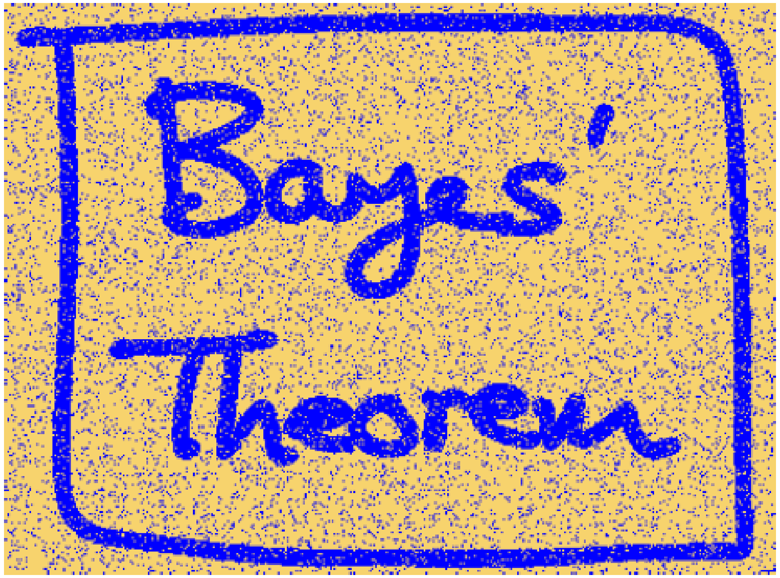
- Markov Blanket of node i - Set of parents, children and co-parents of node i

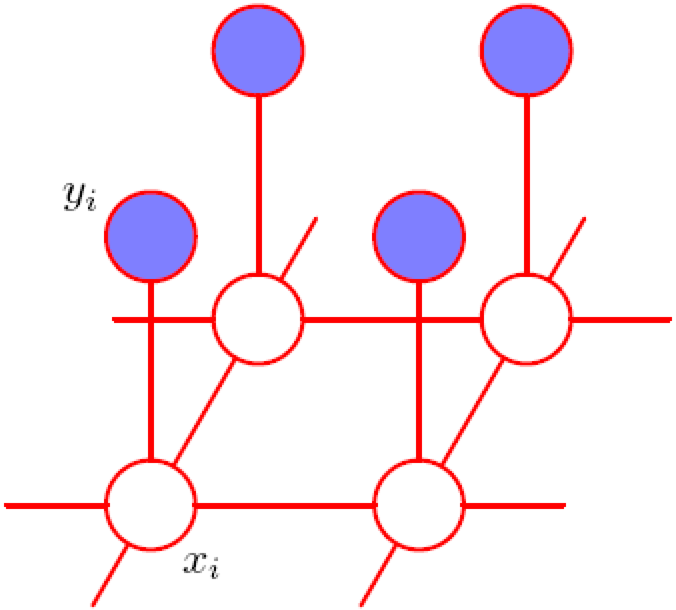
Undirected – Markov Random Fields

- Popular in statistical physics and computer vision communities

- Example – Image Denoising

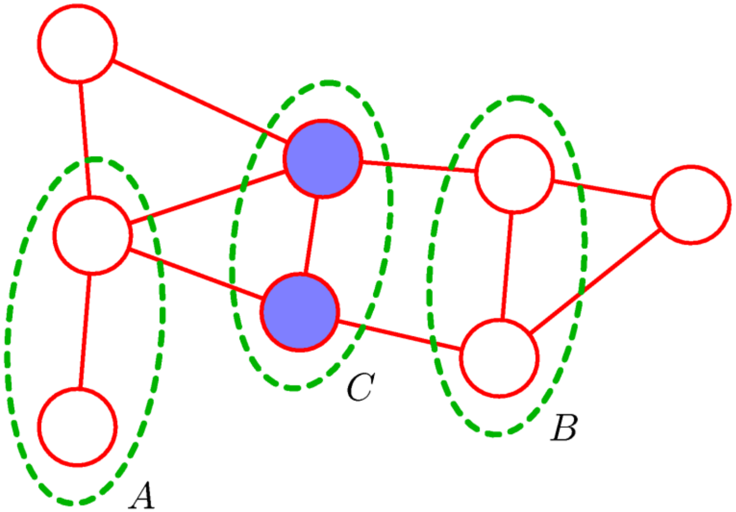
x_i – value at pixel i
y_i – observed noisy value





Conditional Independence properties

- No directed edges
- Conditional independence ≡ graph separation
- A, B, C – non-intersecting set of nodes
- $A \perp B | C$ if all paths between nodes in A & B are "blocked"
  i.e. path contains a node z in C.



Factorization

- Joint distribution factorizes according to the graph

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

$\mathcal{C}$ is the set of maximal cliques in the graph

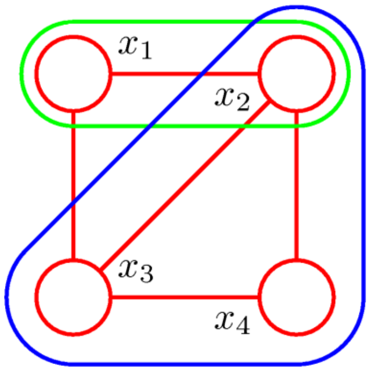
$\psi_C(x_C)$ is a potential function on the clique x_C

Arbitrary positive function

normalization factor

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

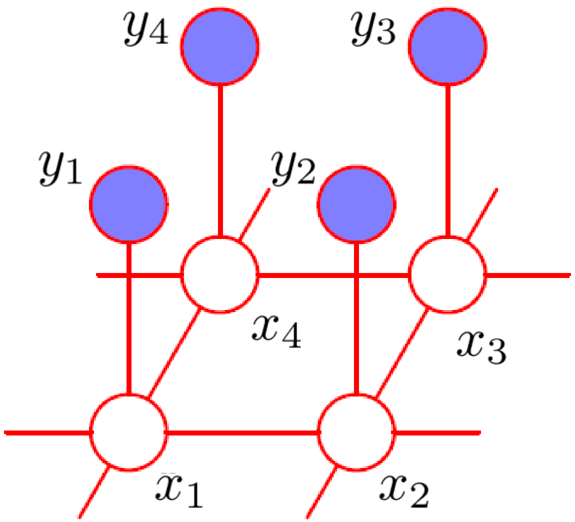
typically NP-hard to compute



Clique, $x_C = \{x_1, x_2\}$

Maximal clique
$x_C = \{x_2, x_3, x_4\}$

MRF Example



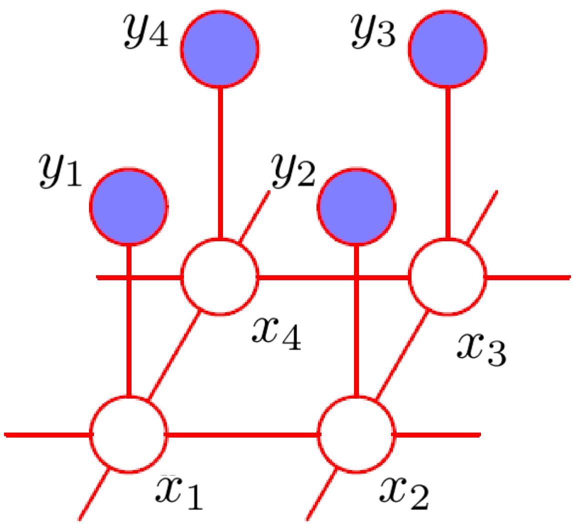
$$P(x,y) \propto \Psi(x_1,x_2)\Psi(x_1,x_3)\Psi(x_2,x_4)\Psi(x_3,x_4)\prod_{i=1}^{4}\Psi(x_i,y_i)$$

Often $\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$

Energy of the clique (e.g. lower if variables in clique take similar values)

$$p(\mathbf{x}) = \prod_{C\in\mathcal{C}} \exp\{-E(\mathbf{x}_C)\} = \exp\{-\sum_{C\in\mathcal{C}} E(\mathbf{x}_C)\}$$

MRF Example



Ising model:

cliques are edges $x_C = \{x_i, x_j\}$
binary variables $x_i \in \{-1,1\}$

$$\psi_C(\mathbf{x}_C) = \exp\{\beta \underbrace{x_i x_j}\}$$

1 if $x_i = x_j$
-1 if $x_i \neq x_j$

$$p(\mathbf{x}) = \prod_{(i,j)\in E} \exp\{\beta x_i x_j\} = \exp\{\sum_{(i,j)\in E} \beta x_i x_j\}$$

Probability of assignment is higher if neighbors x_i and x_j are same

Hammersley-Clifford Theorem

- Set of distributions that factorize according to the graph - $\mathcal{F}$

- Set of distributions that respect conditional independencies implied by graph-separation – $\mathcal{I}$

$$\mathcal{F} \Rightarrow \mathcal{I}$$

Important because: **Given independencies of *P* can get MRF structure *G***

$$\mathcal{I} \Leftarrow \mathcal{F}$$

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

What you should know…

- Graphical Models: Directed Bayesian networks, Undirected Markov Random Fields
  - A compact **representation** for large probability distributions
  - Not an algorithm
- Representation of a BN, MRF
  - Variables
  - Graph
  - CPTs
- Why BNs and MRFs are useful
- D-separation (conditional independence) & factorization