Graphical Models

CS 6355: Structured Prediction

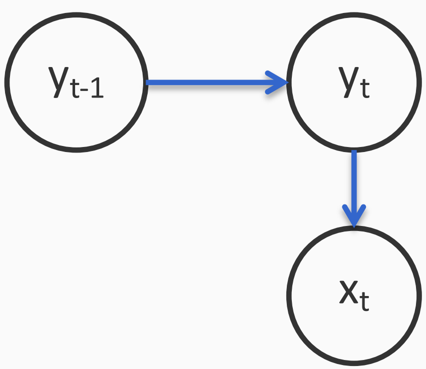
THE UNIVERSITY OF UTAH

So far…

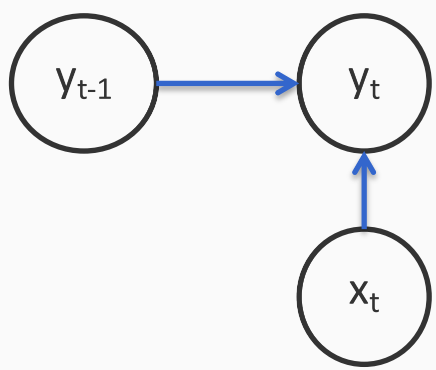
We discussed sequence labeling tasks:

- HMM: Hidden Markov Models
- MEMM: Maximum Entropy Markov Models
- CRF: Conditional Random Fields

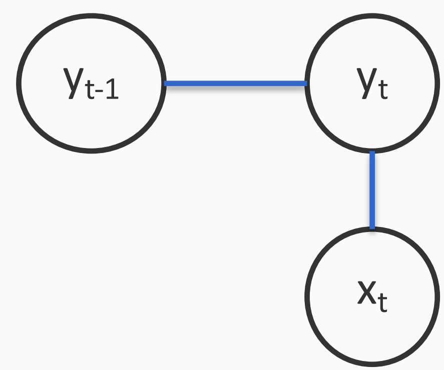
All these models use a linear chain structure to describe the interactions between random variables.



HMM



MEMM



CRF

This lecture

Graphical models

- Directed: Bayesian Networks
- Undirected: Markov Networks (Markov Random Field)

- Representations
- Inference
- Learning

Probabilistic Graphical Models

- Languages that represent probability distributions over multiple random variables
  - Directed or undirected graphs
- Encodes conditional independence assumptions
- Or equivalently, encodes factorization of joint probabilities.
- General machinery for
  - Algorithms for computing marginal and conditional probabilities
    - Recall that we have been looking at most probable states so far
    - Exploiting graph structure
  - An “inference engine”
  - Can introduce prior probability distributions
    - Because parameters are also random variables

Bayesian Network

Decompose joint probability via a directed acyclic graph

- Nodes represent random variables
- Edges represent conditional dependencies
- Each node is associated with a conditional probability table

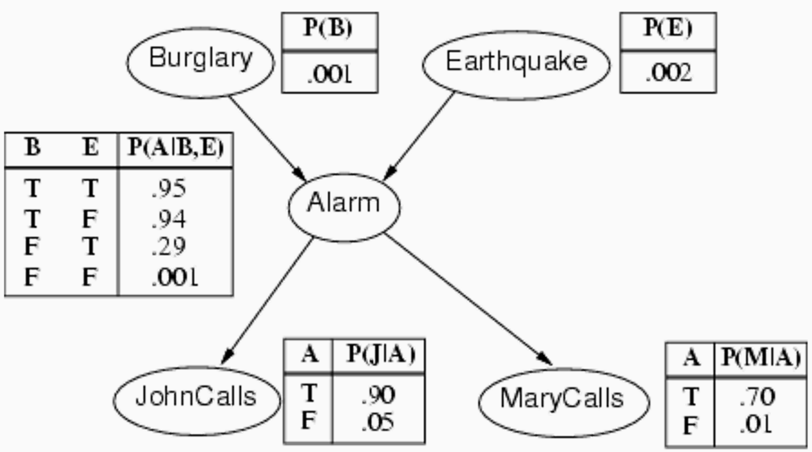
$$P(z_1, z_2, \cdots z_n) = \prod_i P\big(z_i \mid \mathrm{Parents}(z_i)\big)$$

Bayesian Network

Decompose joint probability via a directed acyclic graph

- Nodes represent random variables
- Edges represent conditional dependencies
- Each node is associated with a conditional probability table

$$P(z_1, z_2, \cdots z_n) = \prod_i P\big(z_i \mid \text{Parents}(z_i)\big)$$



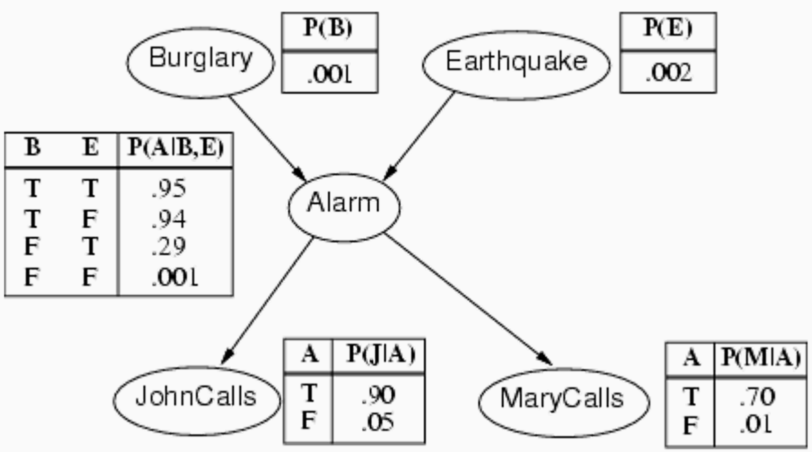
Example from Russell and Norvig

Bayesian Network

Decompose joint probability via a directed acyclic graph

- Nodes represent random variables
- Edges represent conditional dependencies
- Each node is associated with a conditional probability table

$$P(z_1, z_2, \cdots z_n) = \prod_i P\big(z_i \mid \text{Parents}(z_i)\big)$$



Joint probability

$$P(B, E, A, J, M) =$$
$$P(B) \cdot P(E) \cdot P(A \mid B, E) \cdot P(J \mid A) \cdot P(M \mid A)$$

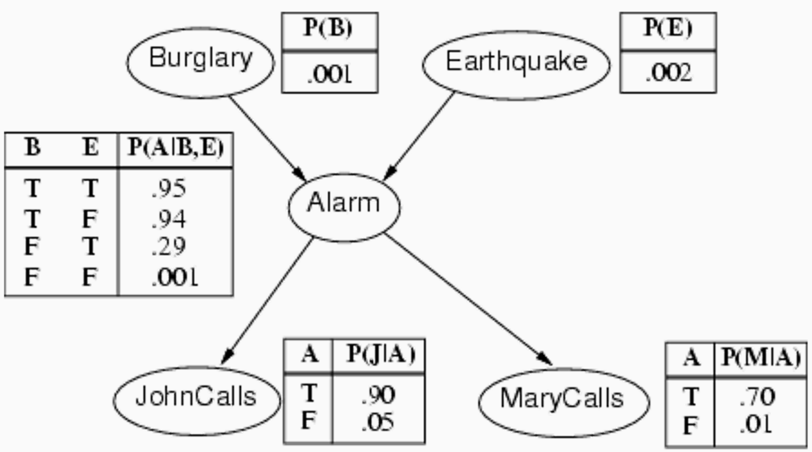
Example from Russell and Norvig

Bayesian Network

Decompose joint probability via a directed acyclic graph

- Nodes represent random variables
- Edges represent conditional dependencies
- Each node is associated with a conditional probability table

$$P(z_1, z_2, \cdots z_n) = \prod_i P\big(z_i \mid \text{Parents}(z_i)\big)$$



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

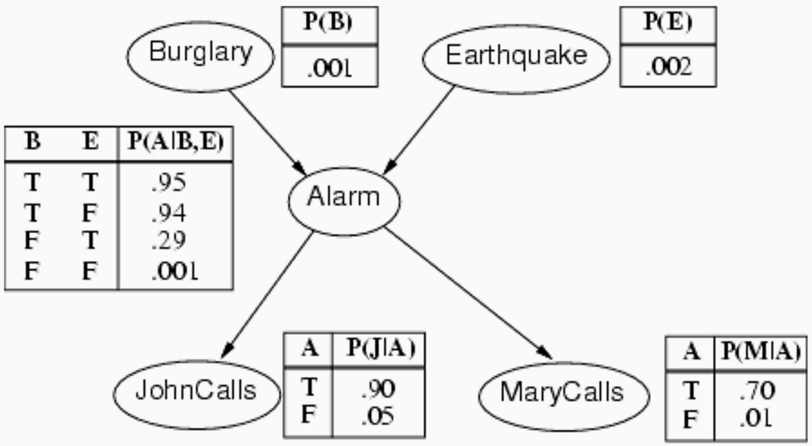
A	P(M|A)
T	.70
F	.01

Joint probability

$$P(B, E, A, J, M) =$$
$$P(B) \cdot P(E) \cdot P(A \mid B, E) \cdot P(J \mid A) \cdot P(M \mid A)$$

The network and its parameters are a compact representation of the joint probability distribution

Bayesian Network



Joint probability

$$P(B,E,A,J,M) = P(B) \cdot P(E) \cdot P(A \mid B,E) \cdot P(J \mid A) \cdot P(M \mid A)$$

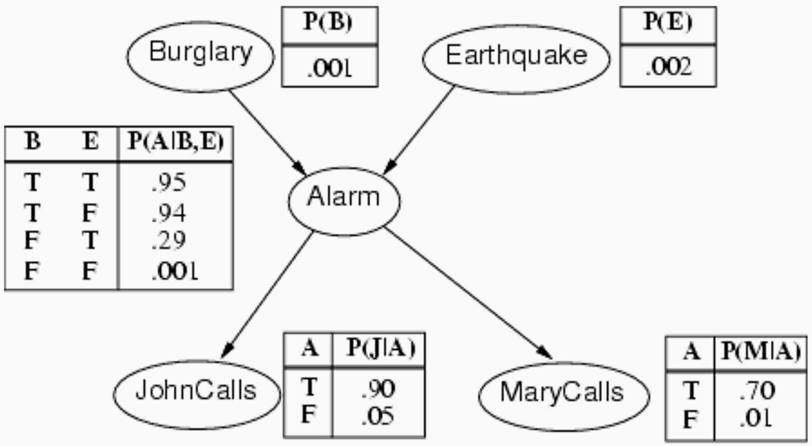
The network and its parameters are a compact representation of the joint probability distribution

We can query the model about any of the variables now

- "What is the probability that Mary calls if there is an earthquake?"
- "If John called and Mary did not call, what is the probability that there was a burglary?"

Example from Russell and Norvig

Bayesian Network



Joint probability

$$P(B,E,A,J,M) =$$
$$P(B) \cdot P(E) \cdot P(A \mid B,E) \cdot P(J \mid A) \cdot P(M \mid A)$$

The network and its parameters are a compact representation of the joint probability distribution

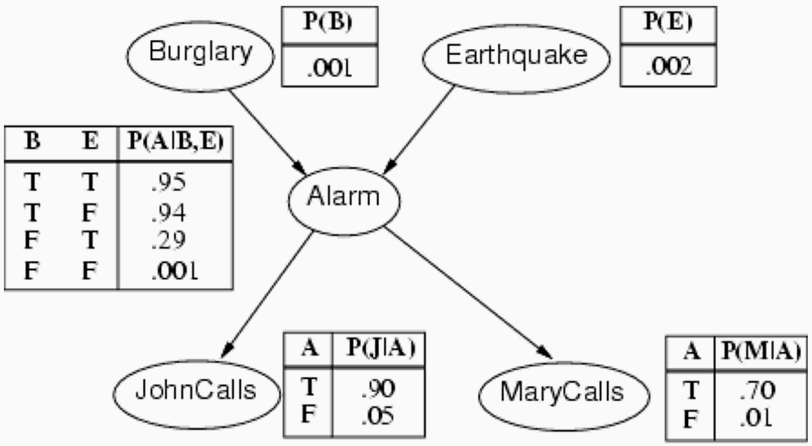
We can query the model about any of the variables now

- “What is the probability that Mary calls if there is an earthquake?”
- “If John called and Mary did not call, what is the probability that there was a burglary?”

$$P(M \mid E) = \frac{P(M,E)}{P(E)} = \frac{\sum_{B,A,J} P(B,E,A,J,M)}{\sum_{B,A,J,M} P(B,E,A,J,M)}$$

Example from Russell and Norvig

Bayesian Network



Joint probability

$$P(B, E, A, J, M) =$$
$$P(B) \cdot P(E) \cdot P(A \mid B, E) \cdot P(J \mid A) \cdot P(M \mid A)$$

The network and its parameters are a compact representation of the joint probability distribution

We can query the model about any of the variables now

- “What is the probability that Mary calls if there is an earthquake?”

$$P(M \mid E) = \frac{P(M, E)}{P(E)} = \frac{\sum_{B,A,J} P(B, E, A, J, M)}{\sum_{B,A,J,M} P(B, E, A, J, M)}$$

- “If John called and Mary did not call, what is the probability that there was a burglary?”

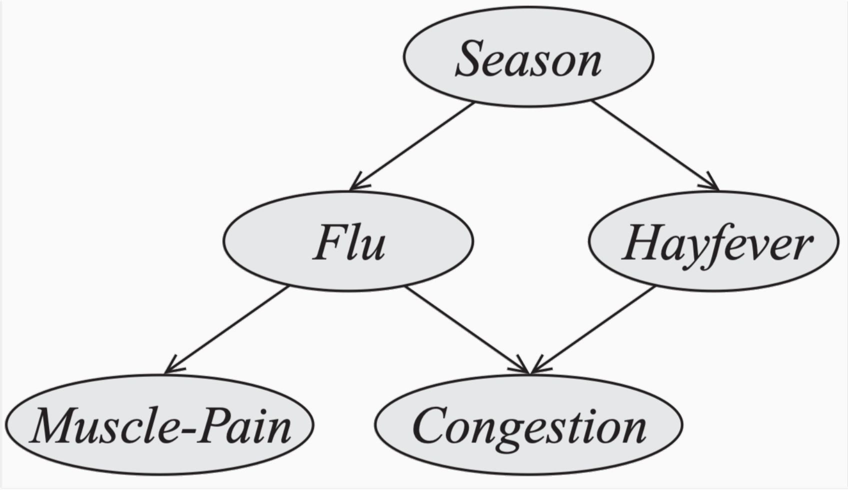
$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M)} = \frac{\sum_{E,A} P(B, E, A, J, \neg M)}{\sum_{B,E,A} P(B, E, A, J, \neg M)}$$

Example from Russell and Norvig

Independence Assumptions of a BN

If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

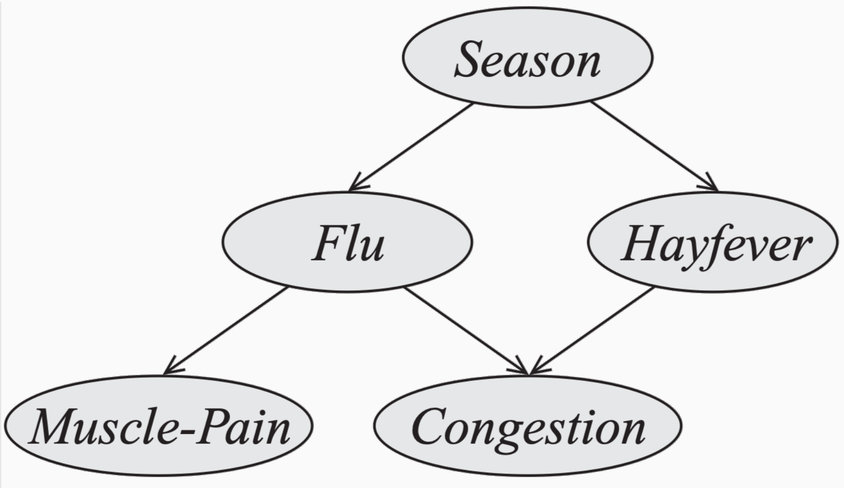


Example from Daphne Koller

Independence Assumptions of a BN

If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"



Example from Daphne Koller

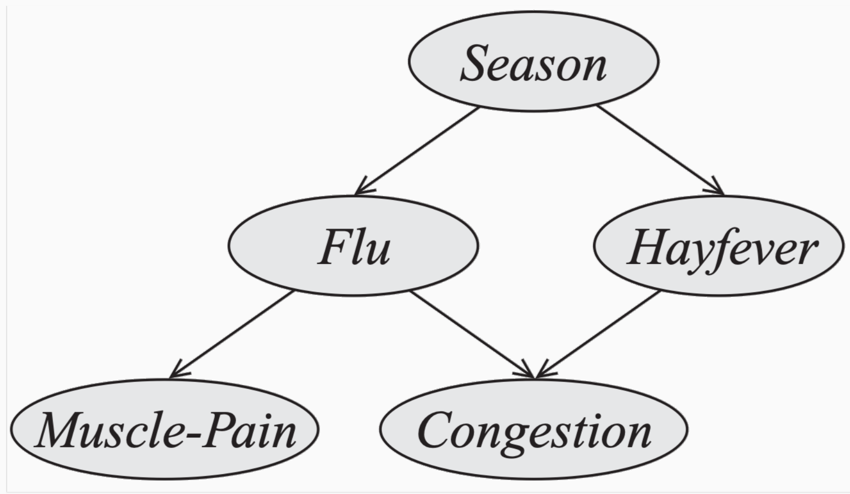
Independence Assumptions of a BN

If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Local independencies: A node is independent with its *non-descendants* given its parents

$$X_i \perp \mathrm{NonDescendants}(X_i) \mid \mathrm{Parents}(X_i)$$



Example from Daphne Koller

Independence Assumptions of a BN

If X, Y, Z are random variables, we write

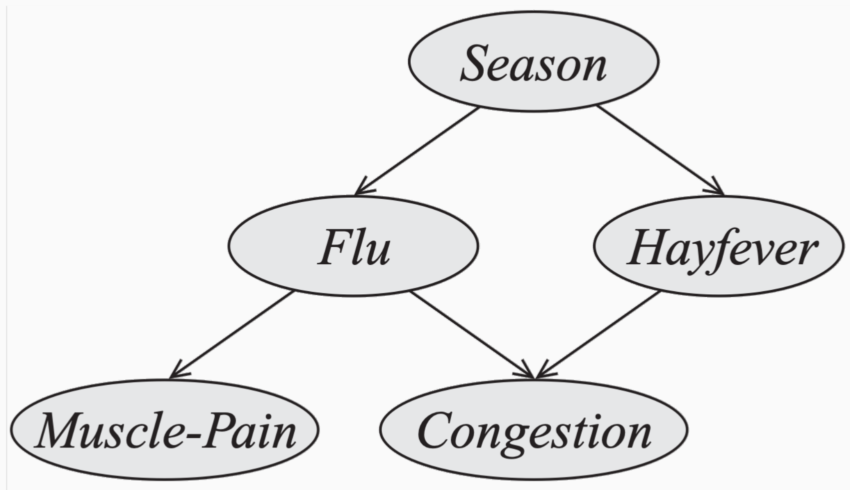
- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Local independencies: A node is independent with its non-descendants given its parents

$$X_i \perp \mathrm{NonDescendants}(X_i) \mid \mathrm{Parents}(X_i)$$

Examples:

- $Flu \perp Hayfever \mid Season$
- $Congestion \perp Season \mid Flu, Hayfever$



Example from Daphne Koller

Independence Assumptions of a BN

If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Local independencies: A node is independent with its non-descendants given its parents

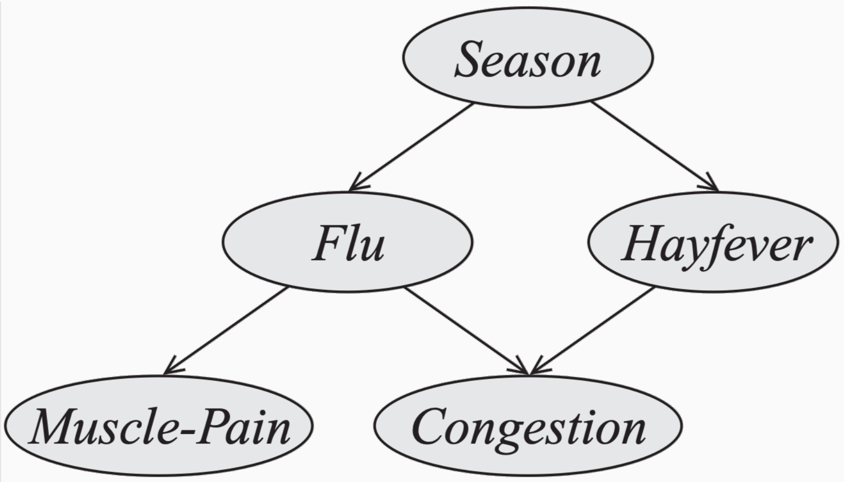
$$X_i \perp \text{NonDescendants}(X_i) \mid \text{Parents}(X_i)$$

Examples:

- $Flu \perp Hayfever \mid Season$
- $Congestion \perp Season \mid Flu, Hayfever$

Parents of a node shield it from influence of ancestors and non-descendants…

… but information about descendants can influence beliefs about a node.



Example from Daphne Koller

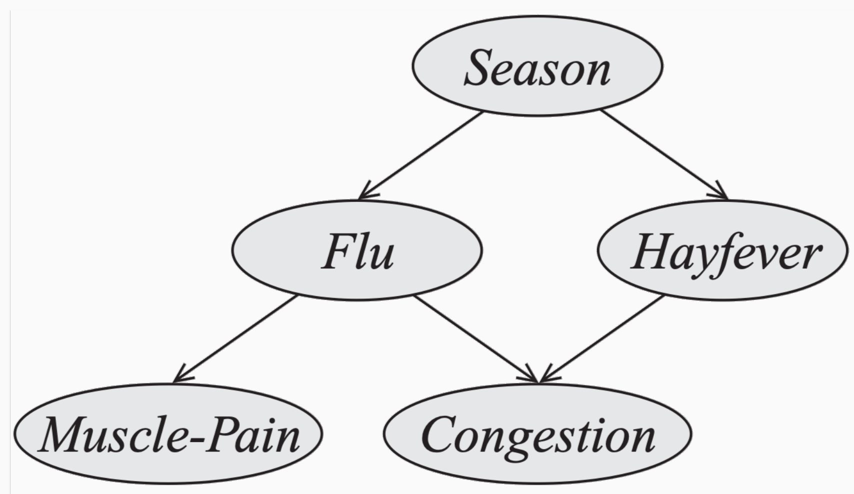
Independence Assumptions of a BN

If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Topological independencies: A node is independent of all other nodes given its *parents, children and children's parents,* together called the node's **Markov Blanket**

$$X_i \perp X_j \mid \text{MarkovBlanket}(X_i)$$



Example from Daphne Koller

Independence Assumptions of a BN

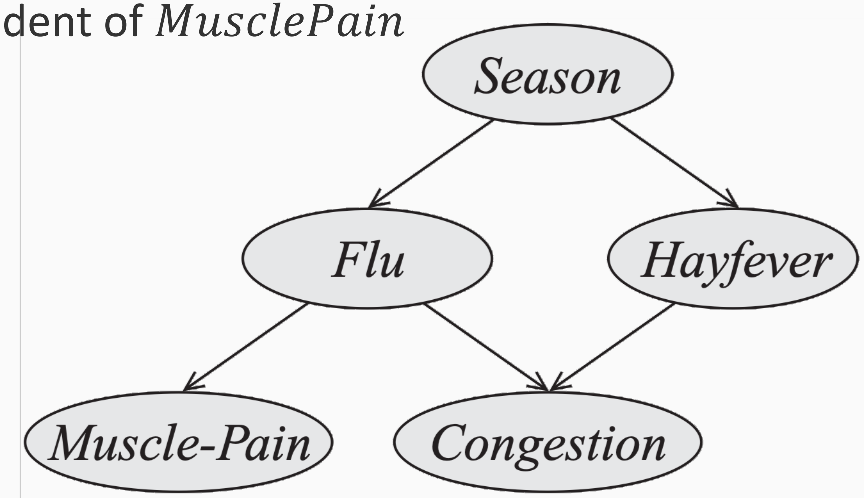
If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Topological independencies: A node is independent of all other nodes given its *parents, children and children's parents,* together called the node's **Markov Blanket**

$$X_i \perp X_j \mid \text{MarkovBlanket}(X_i)$$

Example: The Markov blanket of $Hayfever$ is the set $\{Season, Congestion, Flu\}$. If we know these variables, $Hayfever$ is independent of $MusclePain$



Example from Daphne Koller

Independence Assumptions of a BN

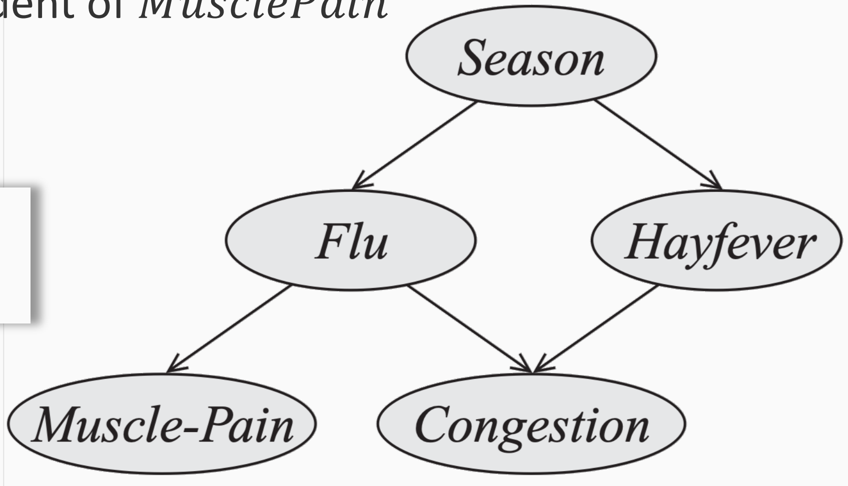
If X, Y, Z are random variables, we write

- $X \perp Y$ to say "X is independent of Y" and
- $X \perp Y \mid Z$ to say "X is independent of Y given Z"

Topological independencies: A node is independent of all other nodes given its *parents, children and children's parents,* together called the node's **Markov Blanket**

$$X_i \perp X_j \mid \text{MarkovBlanket}(X_i)$$

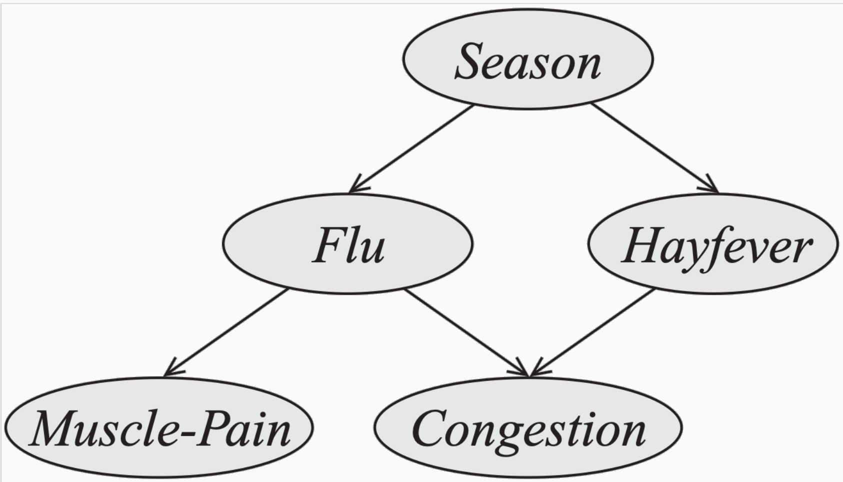
Example: The Markov blanket of $Hayfever$ is the set $\{Season, Congestion, Flu\}$. If we know these variables, $Hayfever$ is independent of $MusclePain$



The Markov blanket of a node shields it from influence of any other node

Example from Daphne Koller

Independence Assumptions of a BN



$$(F \perp H \mid S)$$
$$(C \perp S \mid F, H)$$
$$(M \perp H, C \mid F)$$
$$(M \perp C \mid F)$$

- *Local independencies*: A node is independent with its non-descendants given its parents.

$$(X_i \perp \mathrm{NonDescendants}(X_i) | \mathrm{Parents}(X_i))$$

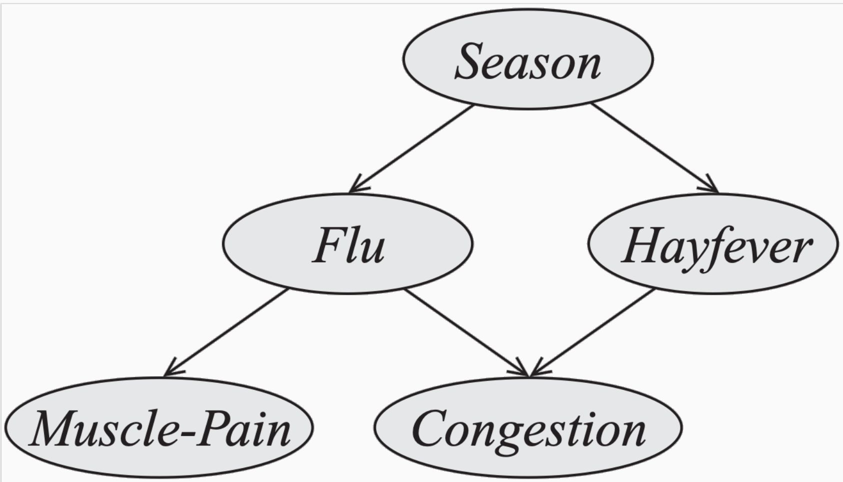
- Topological *independencies*: A node is independent of all other nodes given its *parents, children and children's parents*—that is given its Markov Blanket.

$$(X_i \perp X_j | \mathrm{MB}(X_i)) \quad \text{for all} \quad j \neq i$$

- More general notions of independencies exist.

Example from Daphne Koller

Independence Assumptions of a BN



$$(F \perp H \mid S)$$
$$(C \perp S \mid F, H)$$
$$(M \perp H, C \mid F)$$
$$(M \perp C \mid F)$$

- *Local independencies*: A node is independent with its non-descendants given its parents.

$$(X_i \perp \mathrm{NonDescendants}(X_i) | \mathrm{Parents}(X_i))$$

- Topological *independencies*: A node is independent of all other nodes given its *parents, children and children's parents*—that is given its Markov Blanket.

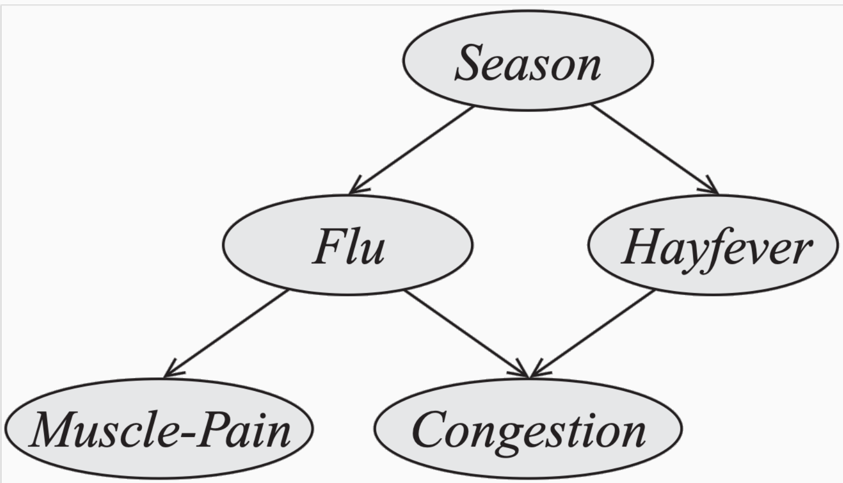
$$(X_i \perp X_j | \mathrm{MB}(X_i)) \quad \text{for all} \quad j \neq i$$

- More general notions of independencies exist.

Where do the independence assumptions come from?

Example from Daphne Koller

Independence Assumptions of a BN



$$(F \perp H \mid S)$$
$$(C \perp S \mid F, H)$$
$$(M \perp H, C \mid F)$$
$$(M \perp C \mid F)$$

- *Local independencies*: A node is independent with its non-descendants given its parents.

$$(X_i \perp \mathrm{NonDescendants}(X_i) | \mathrm{Parents}(X_i))$$

- Topological *independencies*: A node is independent of all other nodes given its *parents, children and children's parents*—that is given its Markov Blanket.

$$(X_i \perp X_j | \mathrm{MB}(X_i)) \quad \text{for all} \quad j \neq i$$

- More general notions of independencies exist.

Where do the independence assumptions come from?

Domain knowledge

Example from Daphne Koller

We have seen Bayesian networks before

- The naïve Bayes model is a simple Bayesian Network
  - The naïve Bayes assumption is an example of an independence assumption

- The hidden Markov model is another Bayesian network

Inference with Bayesian networks

Edges in a BN are typically interpreted as being causal, i.e., the parents of a node causally influencing them

Inference with Bayesian networks

Edges in a BN are typically interpreted as being causal, i.e., the parents of a node causally influencing them

The general inference problem with Bayesian networks: Find the probability of unknown variables, having observed values of some others.

Inference with Bayesian networks

Edges in a BN are typically interpreted as being causal, i.e., the parents of a node causally influencing them

The general inference problem with Bayesian networks: Find the probability of unknown variables, having observed values of some others.

Example: If we have a BN with variables X_1, X_2, X_3 and we wish to compute the probability of X_1 given X_3

$$P(X_1 \mid X_3) = \frac{P(X_1, X_3)}{P(X_3)}$$

Inference with Bayesian networks

Edges in a BN are typically interpreted as being causal, i.e., the parents of a node causally influencing them

The general inference problem with Bayesian networks: Find the probability of unknown variables, having observed values of some others.

Example: If we have a BN with variables X_1, X_2, X_3 and we wish to compute the probability of X_1 given X_3

$$P(\, X_1 \mid X_3\,) = \frac{P(X_1, X_3)}{P(X_3)} = \frac{\sum_{X_2} P(X_1, X_2, X_3)}{\sum_{X_1, X_3} P(X_1, X_2, X_3)}$$

Inference with Bayesian networks

Edges in a BN are typically interpreted as being causal, i.e., the parents of a node causally influencing them

The general inference problem with Bayesian networks: Find the probability of unknown variables, having observed values of some others.

Example: If we have a BN with variables X_1, X_2, X_3 and we wish to compute the probability of X_1 given X_3

$$P(X_1 \mid X_3) = \frac{P(X_1, X_3)}{P(X_3)} = \frac{\sum_{X_2} P(X_1, X_2, X_3)}{\sum_{X_1, X_3} P(X_1, X_2, X_3)}$$

Bad News: Inference in a Bayesian network is $\#P$ hard (i.e., as hard as counting the number of satisfying solutions of a CNF formula)

More bad news: Even approximate inference in a Bayesian network is NP-hard!

Good news: Efficient algorithms exist for networks with special structures.

Causality may not be easy to determine

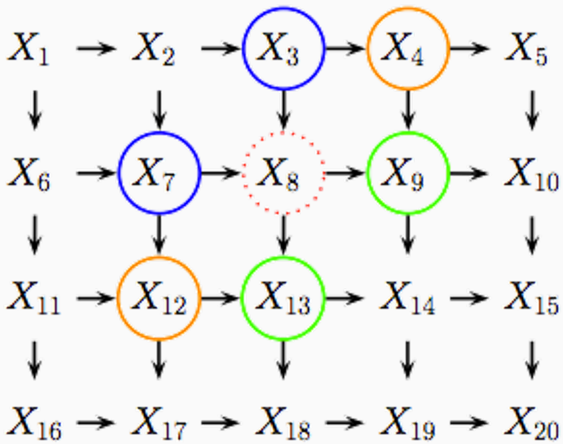
Sometimes Bayes nets cannot represent the independence relations we want *conveniently*

- Eg: Segmenting an image by assigning a label to each pixel

Causality may not be easy to determine

Sometimes Bayes nets cannot represent the independence relations we want *conveniently*

- Eg: Segmenting an image by assigning a label to each pixel
  - Say, we want adjacent labels to influence each other

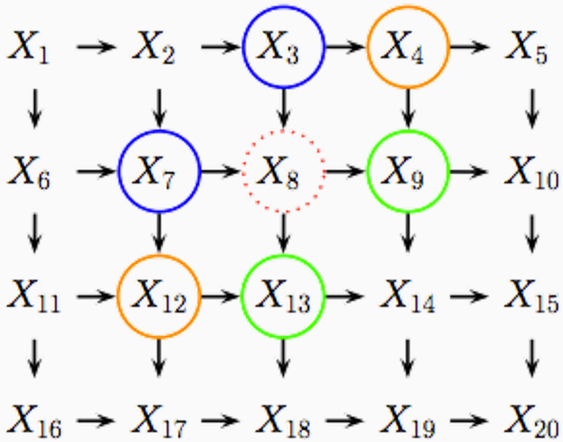


Example from Kevin Murphy

Causality may not be easy to determine

Sometimes Bayes nets cannot represent the independence relations we want *conveniently*

- Eg: Segmenting an image by assigning a label to each pixel
  - Say, we want adjacent labels to influence each other



Two problems:

1. What is the correct direction of arrows?

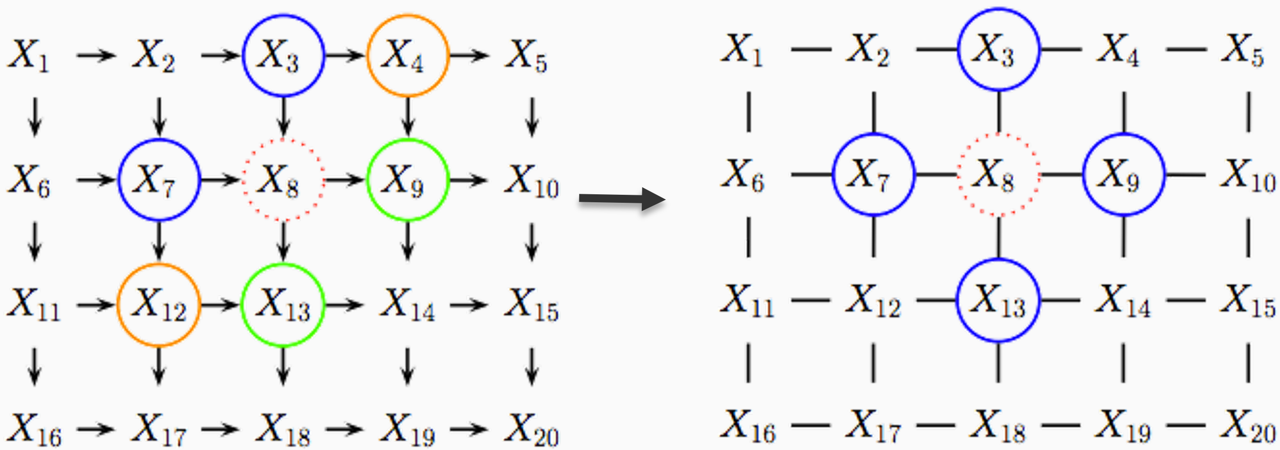
2. For any choice of the arrows, strange dependencies show up. X_8 is independent of everything given its Markov blanket (other circled nodes here)

Example from Kevin Murphy

From directed to undirected networks

Sometimes Bayes nets cannot represent the independence relations we want conveniently.

- Eg: Segmenting an image by assigning a label to each pixel
  - Say, we want adjacent labels to influence each other



Example from Kevin Murphy

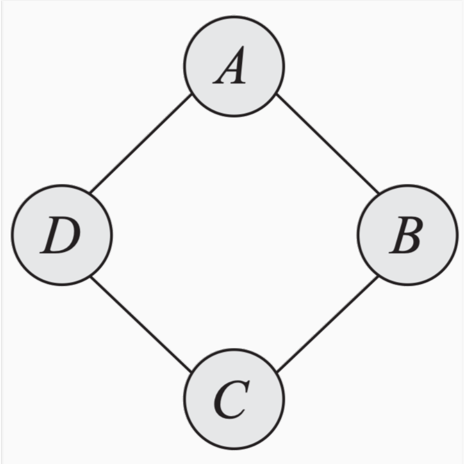
Undirected Graphical Models

a.k.a Markov Random Fields / Markov Networks

- Another way of defining conditional independence

- General structure
  - Nodes are random variables
  - Edges (hyper-edges) define dependencies

- The nodes in a *complete* subgraph form a *clique.*



Cliques:
{AB}, {BC}, {CD}, {AD}

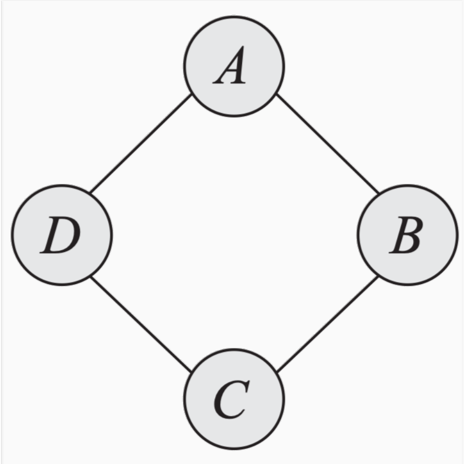
Undirected Graphical Models

a.k.a Markov Random Fields / Markov Networks

- Another way of defining conditional independence

- General structure
  - Nodes are random variables
  - Edges (hyper-edges) define dependencies

- The nodes in a *complete* subgraph form a *clique.*



Cliques:
{AB}, {BC}, {CD}, {AD}

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}} f_c(\mathbf{x}_c)$$

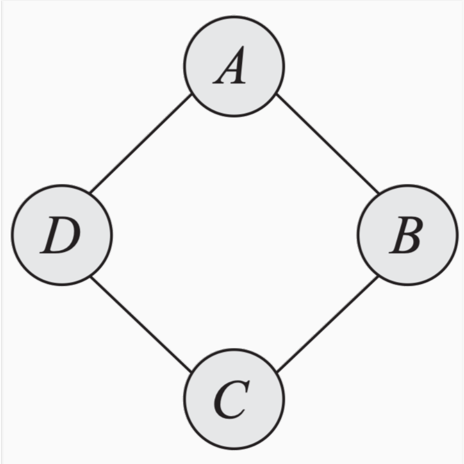
$$P(A,B,C,D) = \frac{1}{Z} f_1(A,B) f_2(B,C) f_3(C,D) f_4(A,D)$$

This is a Gibbs distribution if all factors are *positive*

Undirected Graphical Models

a.k.a Markov Random Fields / Markov Networks

- Another way of defining conditional independence
- General structure
  - Nodes are random variables
  - Edges (hyper-edges) define dependencies
- The nodes in a *complete* subgraph form a *clique.*



Cliques:
{AB}, {BC}, {CD}, {AD}

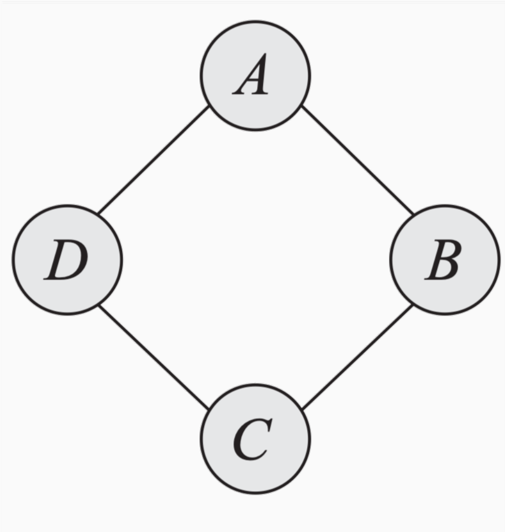
The joint probability decouples over cliques. Every clique x_c associated with a *potential function* $f(x_c)$

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}} f_c(\mathbf{x}_c)$$

$$P(A,B,C,D) = \frac{1}{Z} f_1(A,B) f_2(B,C) f_3(C,D) f_4(A,D)$$

This is a Gibbs distribution if all factors are *positive*

Independence Assumptions of a MRF

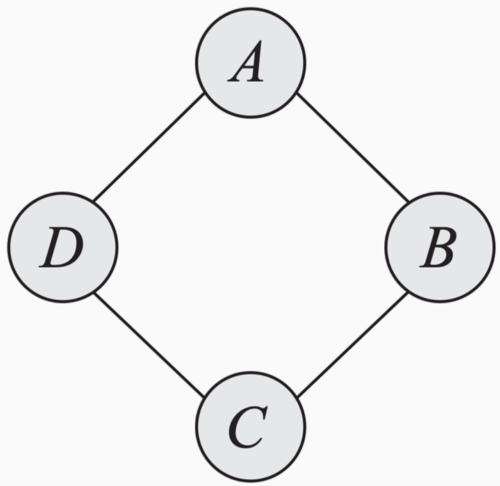


$$(A \perp C \mid B, D)$$
$$(B \perp D \mid A, C)$$

- *Local independencies*: A node is independent of all other nodes given its neighbors.

- Global independencies: If X, Y, Z are sets of nodes, X is conditionally independent of Y given Z if removing all nodes of Z removes all paths from X to Y

Independence Assumptions of a MRF

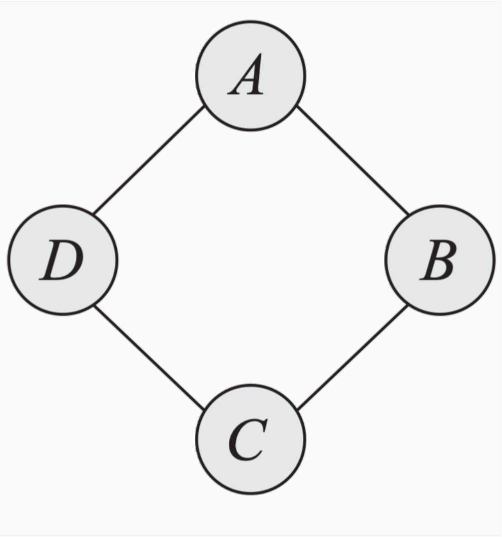


$$(A \perp C \mid B, D)$$
$$(B \perp D \mid A, C)$$

- *Local independencies*: A node is independent of all other nodes given its neighbors.
- Global independencies: If X, Y, Z are sets of nodes, X is conditionally independent of Y given Z if removing all nodes of Z removes all paths from X to Y

Where do the independence assumptions come from?

Independence Assumptions of a MRF



$$(A \perp C \mid B, D)$$
$$(B \perp D \mid A, C)$$

- *Local independencies*: A node is independent of all other nodes given its neighbors.

- Global independencies: If X, Y, Z are sets of nodes, X is conditionally independent of Y given Z if removing all nodes of Z removes all paths from X to Y

Where do the independence assumptions come from?

Domain knowledge

MRF to Factor graph

$$P_\theta(\mathbf{x}) \propto \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$$

Normalize:

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$$

where $Z(\theta) = \sum_{\mathbf{x}} \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$

Z: Called the partition function, sum over all assignments to the random variables

f($\mathbf{x}_c$, μ) is often written as exp($\mu^T \mathbf{x}_c$)
Log-linear model

Factor graphs

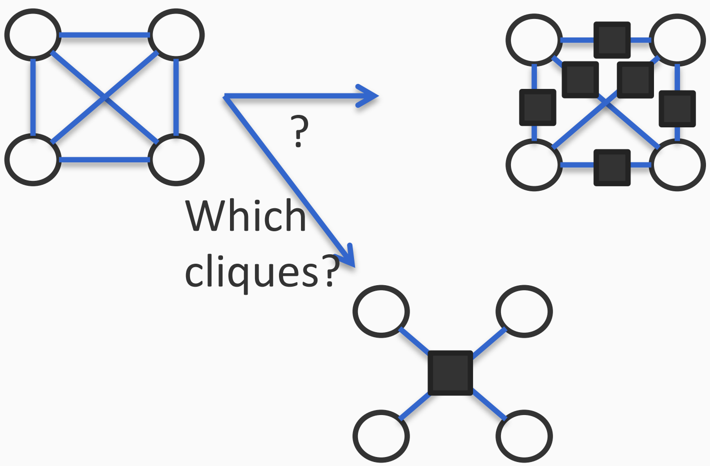
$$P_\theta(\mathbf{x}) \propto \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$$

Normalize:

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$$

where $Z(\theta) = \sum_{\mathbf{x}} \prod_{c \in \text{Cliques}} f(\mathbf{x}_c, \theta)$

Factor graph: Makes the factorization explicit, factors instead of cliques



Z: Called the partition function, sum over all assignments to the random variables

$f(\mathbf{x}_c, \mu)$ is often written as $\exp(\mu^T \mathbf{x}_c)$
Log-linear model

Factor graphs

$$P_\theta(\mathbf{x}) \propto \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$$

Normalize:

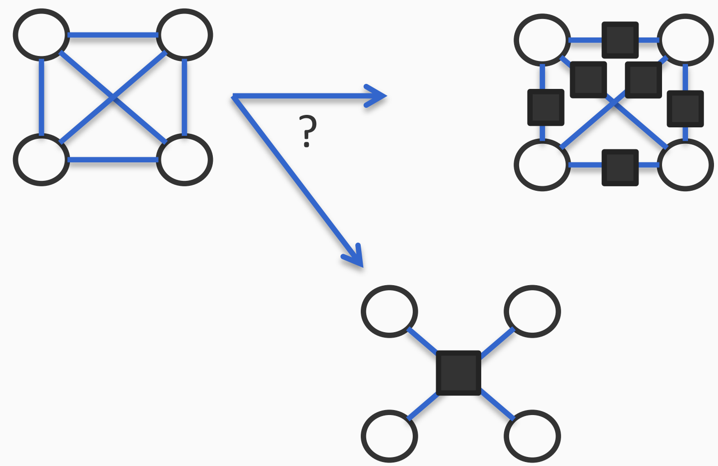
$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$$

where $Z(\theta) = \sum_{\mathbf{x}} \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$

Z: Called the partition function, sum over all assignments to the random variables

f($\mathbf{x}_c$, μ) is often written as exp($\mu^T \mathbf{x}_c$)
Log-linear model

Factor graph: Makes the factorization explicit, factors instead of cliques



Factor graphs

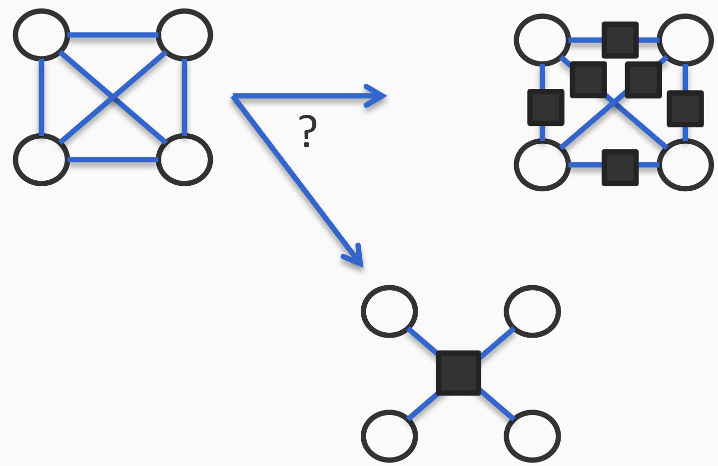
$$P_\theta(\mathbf{x}) \propto \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$$

Normalize:

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$$

where $Z(\theta) = \sum_{\mathbf{x}} \prod_{c \in \cancel{\text{Cliques}}\ \text{Factors}} f(\mathbf{x}_c, \theta)$

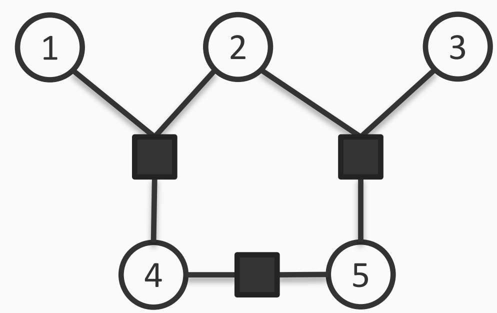
Factor graph: Makes the factorization explicit, factors instead of cliques



Z: Called the partition function, sum over all assignments to the random variables

$f(\mathbf{x}_c, \mu)$ is often written as $\exp(\mu^T \mathbf{x}_c)$

Log-linear model



$$P(\mathbf{x}) = \frac{1}{Z} f_a(x_1, x_2, x_4) f_b(x_2, x_3, x_5) f_c(x_4, x_5)$$

Comments about MRFs

- Connection to statistical physics
  - Identical to Boltzmann distribution in energy based models
  - Probability of a system existing in a state:

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(\mathbf{x}_c)\right)$$

Z: *Zustandssumme,* “sum over states”, more commonly called the partition function

Comments about MRFs

- Connection to statistical physics
  - Identical to Boltzmann distribution in energy based models
  - Probability of a system existing in a state:

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(\mathbf{x}_c)\right)$$

Energy of clique c existing in state $\mathbf{x}_c$

Z: *Zustandssumme,* "sum over states", more commonly called the partition function

History of the Markov random field

Ernst Ising [1925] introduced a model to explain permanent ferromagnetism in ferromagnets below a certain temperature

- Early versions of the idea by Lenz [1920]

Ising's original model:

- Suppose we have a chain of points, each of which can be associated with a certain spin (either up or down)

- The goal: To describe a probability measure over configurations of spins at a specified temperature

- Ising defined the energy of a configuration as being locally factorized over neighboring points

Comments about MRFs

Connection to statistical physics

- Identical to Boltzmann distribution in energy based models
- Probability of a system existing in a state:

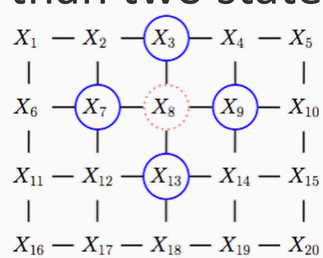
$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(\mathbf{x}_c)\right)$$

Energy of clique c existing in state $\mathbf{x}_c$

Z: *Zustandssumme,* "sum over states", more commonly called the partition function

If **x** is dependent on all its neighbors:

- If x can be in one of two states (binary), *Ising model*
- If x can be in one of more than two states (multiclass), Potts model



Bayesian Networks vs. Markov Networks

- Both networks represent
  - A set of conditional independence relations
  - i.e, a skeleton that shows how a joint probability distribution is factorized

- Both networks have theorems about equivalence between *conditional independence and joint probability factorization*

- *Converting between these representations*
  - A BN can be converted into an MRF with a normalization constant one
  - A MRF can also be converted into a BN, but this may lead to a very large network

See the chapter on undirected graphical models in Koller and Friedman's book

Computational questions

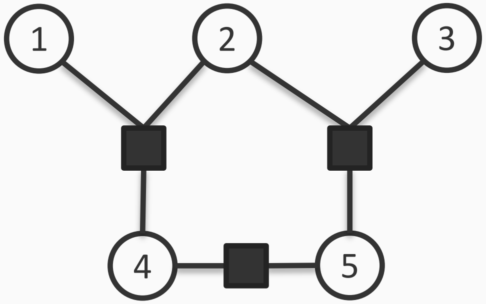
- Learning model parameters
- Learning independence assumptions
- Inference

Learning questions

Two kinds of learning questions:

1. Structure learning: Given data, find independence assumptions to design an MRF (or a BN)
   – A difficult problem, we will not see a lot of this
2. Learning model parameters: Given data and a structure, find the parameters that define the factor potentials
   – We will see more of this as we go along

Inference in graphical models

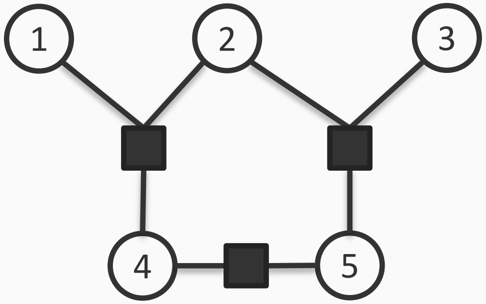


(more on this in future classes)

In general, compute probability of a subset of states

– $P(\mathbf{x}_A)$, for some subsets of random variables $\mathbf{x}_A$
  - Note: So far, we have generally considered the equivalent of $\text{argmax}_x P(\mathbf{x})$

Inference in graphical models



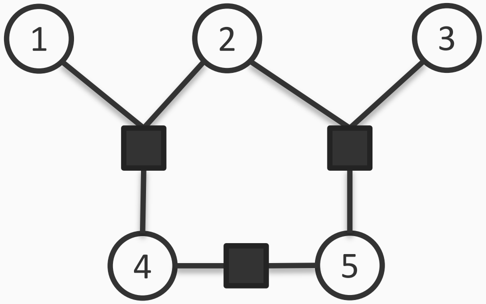
(more on this in future classes)

In general, compute probability of a subset of states

– $P(\mathbf{x}_A)$, for some subsets of random variables $\mathbf{x}_A$
  - Note: So far, we have generally considered the equivalent of $\text{argmax}_x\ P(\mathbf{x})$

- Exact inference
- "Approximate" inference

Inference in graphical models



(more on this in future lectures)

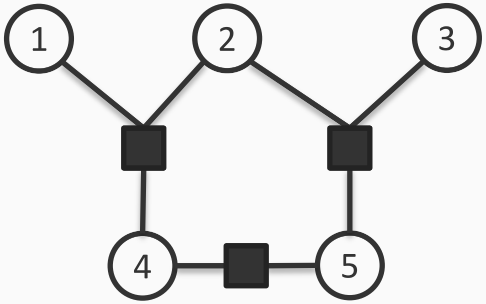
In general, compute probability of a subset of states

– $P(\mathbf{x}_A)$, for some subsets of random variables $\mathbf{x}_A$

- Exact inference
  - – Variable elimination
    - Marginalize by summing out variables in a "good" order
    - Think about what we did for Viterbi *What makes an ordering good?*
  - – Belief propagation (exact only for graphs without loops)
    - Nodes pass messages to each other about their estimate of what the neighbor's state should be
  - – Generally efficient for trees, sequences (and maybe other graphs too)

- "Approximate" inference

Inference in graphical models



(more on this in future lectures)

In general, compute probability of a subset of states

- $P(\mathbf{x}_A)$, for some subsets of random variables $\mathbf{x}_A$

- Exact inference *NP-hard in general, works for simple graphs*
- "Approximate" inference
  - Markov Chain Monte Carlo
    - Gibbs Sampling/Metropolis-Hastings
  - Variational algorithms
    - Frame inference as an optimization problem, perturb it to an approximate one and solve the approximate problem
  - Loopy Belief propagation
    - Run BP and hope it works!
  - The not-so-good news: Approximate inference is also intractable!

Summary

- Graphical models are languages that represent independence assumptions
  - We saw Bayesian networks and Markov networks
  - So far, both networks represent joint distributions

- We will use the factor graph notation across the rest of the semester

- Coming up:
  - Markov logic: A language for defining Markov networks
  - Conditional models