2. Graphical Models

- Undirected graphical models
- Factor graphs
- Bayesian networks
- Conversion between graphical models

Graphical models

- There are three families of graphical models that are closely related, but suitable for different applications and different probability distributions:
  - Undirected graphical models (also known as Markov Random Fields)
  - Factor graphs
  - Bayesian networks

  we will learn what they are, how they are different and how to switch between them.

  consider a probability distribution over $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$

  $$\mu(x_1, x_2, \ldots, x_n)$$

  a **graphical model** is combination of a **graph** and a **set of functions** over a subset of random variables which define the probability distribution of interest

- graphical model is a marriage between probability theory and graph theory that allows compact representation and **efficient inference**, when the probability distribution of interest has special **independence** and **conditional independence structures**
- for example, consider a random vector $x = (x_1, x_2, x_3) \in \mathcal{X}^3$ and a given distribution $\mu(x_1, x_2, x_3)$
- we use (with a slight abuse of notations)

$$\begin{aligned} \mu(x_1) &\triangleq \sum_{x_2, x_3 \in \mathcal{X}^2} \mu(x_1, x_2, x_3) \ , \qquad \text{and} \\ \mu(x_1, x_2) &\triangleq \sum_{x_3 \in \mathcal{X}} \mu(x_1, x_2, x_3) \end{aligned}$$

  to denote the first order and the second order **marginals** respectively
- for this 3-variable case, we can list all possible independence structures

$$x_1 \perp (x_2, x_3) \;\Leftrightarrow\; \mu(x_1, x_2, x_3) = \mu(x_1)\mu(x_2, x_3) \quad (1)$$

$$x_1 \perp x_2 \;\Leftrightarrow\; \mu(x_1, x_2) = \mu(x_1)\mu(x_2) \quad (2)$$

$$x_1 \perp x_2 | x_3 \;\Leftrightarrow\; x_1 – x_3 – x_2 \;\Leftrightarrow\; \mu(x_1, x_2 | x_3) = \mu(x_1|x_3)\mu(x_2|x_3) \quad (3)$$

  and various permutations and combinations of these

- warm-up exercise
  - (1) $\Rightarrow$ (2)
    proof:
    $\mu(x_1,x_2) = \sum_{x_3} \mu(x_1,x_2,x_3) \overset{(1)}{=} \sum_{x_3} \mu(x_1)\mu(x_2,x_3) = \mu(x_1)\mu(x_2)$
  - (2) $\not\Rightarrow$ (3)
    counter example: $X_1 \perp X_2$ and $X_3 = X_1 + X_2$
  - (2) $\not\Leftarrow$ (3)
    counter example: Z_1, Z_2, X_3 are independent and $X_1 = X_3 + Z_1$, $X_2 = X_3 + Z_2$
- this hints that there are different notions of independence, and perhaps we need different types of graphical models to capture them

- all possible independencies for 3-variable distributions $\mu(x_1,x_2,x_3)$
  - $x_1 \perp (x_2,x_3), \quad x_2 \perp (x_1,x_3), \quad x_3 \perp (x_1,x_2)$
  - $x_1 \perp x_2, \quad x_1 \perp x_3, \quad x_2 \perp x_3,$
  - $x_1 \perp x_2 | x_3, \quad x_1 \perp x_3 | x_2, \quad x_2 \perp x_3 | x_1,$
- each $\mu(x_1,x_2,x_3)$ possesses a subset of these 9 independencies

- we can **categorize** all distributions, according to the independence they possess: e.g. $S = \{\mu(x_1, x_2, x_3) : x_1 \perp x_2, \text{ and } x_2 \perp x_3 | x_1\}$
- or we can also **partition** all distributions, according to the independence they possess: e.g. $S = \{\mu(x_1, x_2, x_3) : x_1 \perp x_2, \text{ and } x_2 \perp x_3 | x_1 \text{ but no other independencies}\}$
- there are 2^9 such possible combinations of independencies
- not all of them are feasible,
  e.g. $S = \{\mu(x_1, x_2, x_3) : x_1 \perp x_2, \text{ but } x_1 \not\perp (x_2, x_3)\}$ is an empty set

- in fact, there are exponentially many possible independencies, resulting in doubly exponentially many possible independence structures in a distribution

- we want to use a graph to represent a set of distributions that share some independencies
- perhaps, one graph could represent one subset of independencies (either a inclusive category or a exclusive partition)
- however, there are only 2^{n^2} undirected graphs (4^{n^2} for directed)
- hence, graphical models only capture (important) subsets of possible independence structures

a **probabilistic graphical model** is a graph $G(V, E)$ representing a family of probability distributions

1. that share the same factorization of the probability distribution; and
2. that share the same independence structure.

we study 3 types of graphical models

- **undirected graphical model** = Markov Random Field (MRF)

$$\mu(x) \;=\; \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \psi_c(x_c)$$

where $\mathcal{C}(G)$ is the set of all maximal cliques in the undirected graph $G(V, E)$, $\psi_c(x_c)$ is a non-negative function over the variables $x_c = \{x_i : i \in c\}$, and $Z \in \mathbb{R}^+$ is called the **partition function** which normalizes the distribution to sum to one

- an **undirected graph** $G(V, E)$ is a collection of nodes $V = \{1, 2, \ldots, n\}$ for the variables $\{x_1, \ldots, x_n\}$ and undirected edges $E \subseteq V \times V$
- a **clique** c is a subset of nodes $c \subseteq V$ such that all pairs in c are connected via edges in E
- a clique c is said to be **maximal** if one cannot add any more node to c to make it a larger clique

- **factor graph model (FG)**

$$\mu(x) \;=\; \frac{1}{Z}\prod_{a\in F}\psi_a(x_{\partial a})$$

where F is the set of factor nodes in the undirected bipartite graph $G(V,F,E)$, ∂a is the set of neighbors of the node a, and $\psi_a(x_{\partial a})$ are no-negative functions called the **factors**

- an undirected graph $G(V,F,E)$ is bipartite if there are no edges between a node in V and a node in F
- a node in F is called a **factor node**, and a node in V is called a **variable node**
- each factor node $a\in F$ is associated with a factor $\psi_a(x_{\partial a})$, where ∂a are the variable nodes adjacent to factor a, and $x_{\partial a}$ are the set of corresponding variables

- **directed graphical model** = Bayesian Network (BN)

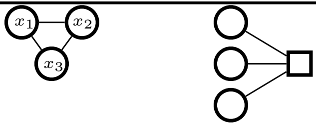
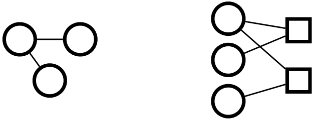
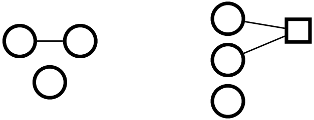
$$\mu(x) \quad = \quad \prod_{i \in V} \mu(x_i | x_{\pi(i)})$$

where $\pi(i)$ is the set of parent nodes in the directed acyclic graph (DAG) $G(V, E)$

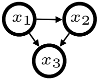
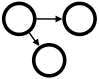
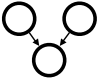
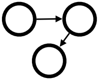
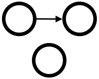
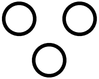
  - in a **directed graph**, an edge (i, j) is different from an edge (j, i)
  - an undirected graph is called **acyclic** if it does not have cycles
  - a cycle in a directed graph is a sequence of nodes $c = (i_1, i_2, \ldots, i_k)$ such that $i_1 = i_k$ and $(i_\ell, i_{\ell+1}) \in E$ for all $\ell \in [k-1]$
  - we use $[N] = \{1, 2, \ldots, N\}$ to denote the first N integers
  - **parent nodes** of a node i in a directed graph is the set of nodes $\pi(i)\{j \in V : (j, i) \in E\}$

- note that missing edges represent simpler distributions with more independence structures
- also, factor graphs are strictly more general than MRFs
- FGs cannot represent all BNs and BNs cannot represent all FGs

- warm-up example: Markov Random Fields (MRF) and Factor Graphs (FG)

MRF / FG	factorization	independence
	$\mu(x_1, x_2, x_3)$	none
	$\psi(x_1, x_2)\psi(x_1, x_3)$	$x_2 \perp x_3 \| x_1$ [Exercise 2.1]
	$\psi(x_1, x_2)\psi(x_3)$	$x_3 \perp (x_1, x_2)$
	$\psi(x_1)\psi(x_2)\psi(x_3)$	all indep.

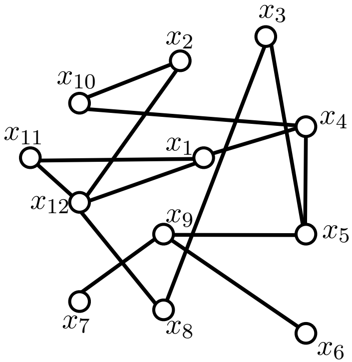
- warm-up example: Bayesian Network (BN) of ordering $(x_1 \to x_2 \to x_3)$

BN	factorization	independence
	$\mu(x_1)\mu(x_2\|x_1)\mu(x_3\|x_1, x_2)$	none
	$\mu(x_1)\mu(x_2\|x_1)\mu(x_3\|x_1)$	$x_2 \perp x_3\|x_1$
	$\mu(x_1)\mu(x_2)\mu(x_3\|x_1, x_2)$	$x_1 \perp x_2$
	$\mu(x_1)\mu(x_2\|x_1)\mu(x_3\|x_2)$	$x_1 \perp x_3\|x_2$
	$\mu(x_1)\mu(x_2\|x_1)\mu(x_3)$	$x_3 \perp (x_1, x_2)$
	$\mu(x_1)\mu(x_2)\mu(x_3)$	all indep.

Family #1: Undirected Pairwise Graphical Models

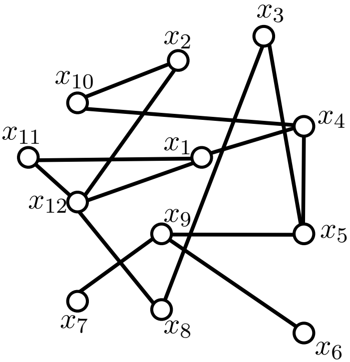
Family #1: Undirected Pairwise Graphical Models

(a.k.a. Pairwise MRF)



$G = (V, E)$, $V = [n] \triangleq \{1, \ldots, n\}$, $x = (x_1, \ldots, x_n)$, $x_i \in \mathcal{X}$
if we say a joint distribution $\mu(x)$ has the above graphical model, then

- $\mu(x)$ can be decomposed as prescribed by the graph G:
  $\mu(x) = (1/Z) \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$
- which implies a certain set of independencies encoded in G



Undirected pairwise graphical models are specified by

- Graph $G = (V, E)$
- Alphabet $\mathcal{X}$
- **Compatibility function** $\psi_{ij} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, for all $(i, j) \in E$

$$\mu(x) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$$

- **pairwise** MRF only allow compatibility functions over **two** variables

Undirected Pairwise Graphical Models

- Graph $G(V, E)$
- Alphabet $\mathcal{X}$
  - Typically $|\mathcal{X}| < \infty$
  - Occasionally $\mathcal{X} = \mathbb{R}$ and

$$\mu(dx) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)\, dx$$

(all formulae interpreted as densities [it is okay if you don't understand the above notation for now])

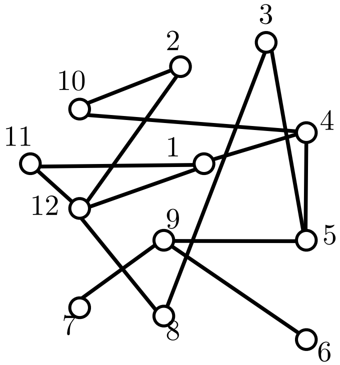
- Compatibility function $\psi_{ij} : \mathcal{X}^2 \to \mathbb{R}^+$

$$\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$$

- Partition function Z plays a crucial role!

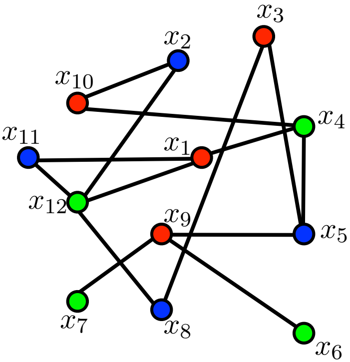
$$Z = \sum_{x\in\mathcal{X}^n} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$$

Graph notation

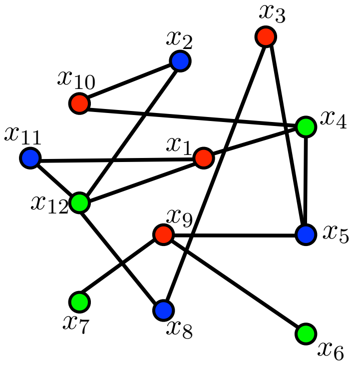


- $\partial i \equiv \{$neighborhood of node $i\}$, $\quad \partial 9 = \{5, 6, 7\}$
- deg$(i) = |\partial i|$, $\quad$ deg$(9) = 3$
- $x_U \equiv (x_i)_{i \in U}$, $\quad x_{\{1,5\}} = (x_1, x_5)$
- $x_{-i} \equiv x_{V \setminus \{i\}}$ $\quad x_{\partial 9} = (x_5, x_6, x_7)$
- Complete graph $\quad x_{-9} = (x_1, \dots, x_8, x_{10}, x_{11}, x_{12})$
- Clique

Example



- Coloring (e.g. ring tone)
- Given graph $G = (V, E)$ and a set of colors $\mathcal{X} = \{R, G, B\}$
- Find a coloring of the vertices such that no two adjacent vertices have the same color
- Fundamental question: Chromatic number
- our goal: translate this into an **inference on graphical models**, so that we can use the techniques from the mature field of **probabilistic graphical models**



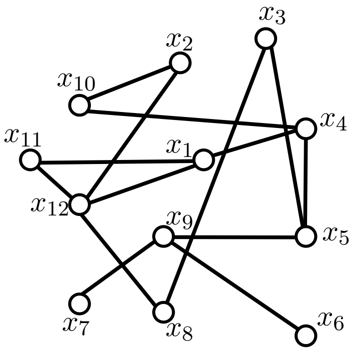
A (joint) probability of interest is uniform measure over all possible colorings:

$$\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} \mathbb{I}(x_i \neq x_j)$$

$\mathbb{I}(x_i \neq x_j)$ is an indicator, which is one if $x_i \neq x_j$ and zero otherwise

- $Z =$ total number of colorings
- Sampling from this distribution is equivalent to finding a coloring
- similarly, independent set problem [Exercise 2.3, 2.4]

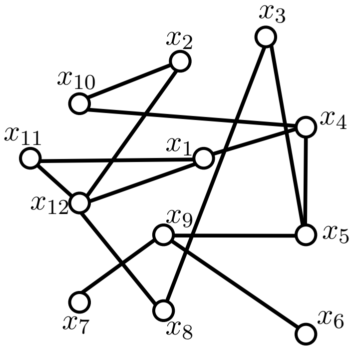
(General) Undirected Graphical Model



Undirected graphical models are specified by

- Graph $G = (V, E)$
- Alphabet $\mathcal{X}$
- Compatibility function $\psi_c : \mathcal{X}^c \to \mathbb{R}_+$, for all maximal cliques $c \in \mathcal{C}$

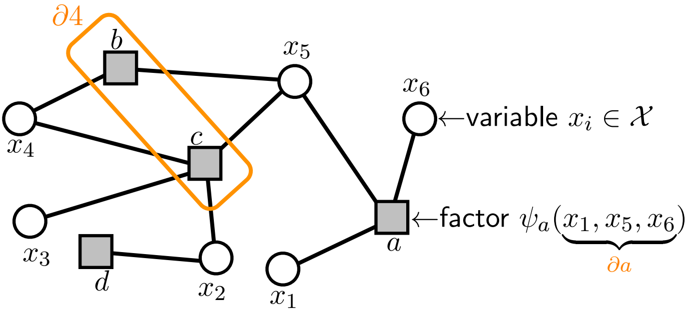
$$\mu(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$



- consider a fixed graph $G(V, E)$
  - the **factorizations** implied by the graph under MRF and pairwise MRF are different, e.g. (x_1, x_{11}, x_{12})
  - however, **independencies** implied by the graph under MRF or pairwise MRF are the same
  - by choosing the right compatibility functions any model represented by pairwise MRF can be represented by MRFs, but not the other way around

Family #2: Factor Graph Models

Family #2: Factor graph models



Factor graph $G = (V, F, E)$

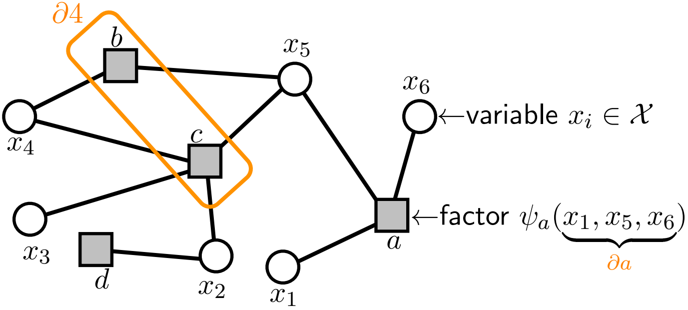
- Variable nodes $i, j, k, \cdots \in V$
- Function nodes $a, b, c, \cdots \in F$

Variable node $x_i \in \mathcal{X}$, for all $i \in V$

Function node $\psi_a : \mathcal{X}^{|\partial a|} \to \mathbb{R}_+$, for all $a \in F$

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a})$$

Factor graph models



Factor graph model is specified by

- Factor graph $G = (V, F, E)$
- Alphabet $\mathcal{X}$
- Compatibility function $\psi_a : \mathcal{X}^{\partial a} \to \mathbb{R}_+$, for $a \in F$

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a})$$

Partition function: $Z = \sum_{x \in \mathcal{X}^V} \prod_{a \in F} \psi_a(x_{\partial a})$

Conversion between factor graphs and pairwise models

From pairwise model to factor graph

A pairwise model on $G(V, E)$ with alphabet $\mathcal{X}$ can be represented by a factor graph $G'(V', F', E')$ with $V' = V$, $F' \simeq E$, $|E'| = 2|E|$, $\mathcal{X}' = \mathcal{X}$.

- Put a factor node on each edge

From factor graph to a general undirected graphical model (MRF)

A factor model on $G(V, F, E)$ with alphabet $\mathcal{X}$ can be represented by a MRF on $G'(V', E')$ with $V' = V$, $E' \simeq \sum_{a \in F} |\partial a|^2$, $\mathcal{X}' = \mathcal{X}$.

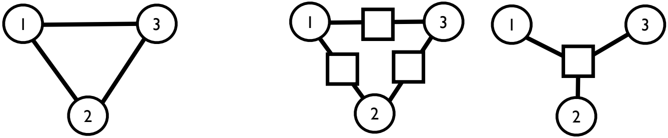
- A factor node is turned into a clique

From factor graph to a pairwise model

A factor model on $G(V, F, E)$ can be represented by a pairwise model on $G'(V', E')$ with $V' = V \cup F$, $E' = E$, $\mathcal{X}' = \mathcal{X}^{\Delta}$, $\Delta = \max_{a \in F} \deg(a)$.

- A factor node is represented by a large variable node

Factor graphs are more 'fine grained' than undirected graphical models



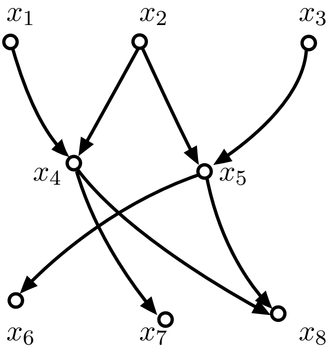
$\psi(x_1, x_2, x_3)$ $\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{31}(x_3, x_1)$ $\psi_{123}(x_1, x_2, x_3)$

all three encodes same independencies, but different factorizations (in particular the degrees of freedom in the compatibility functions are $3|\mathcal{X}|^2$ vs. $|\mathcal{X}|^3$)

- set of independencies represented by MRF is the same as FG
- but FG can represent a larger set of factorizations

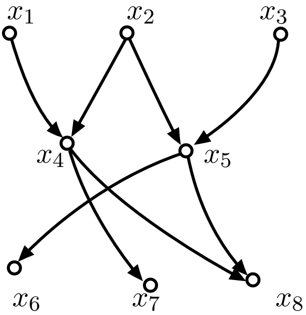
Family #3: Bayesian Networks

Family #3: Bayesian networks



DAG: Directed Acyclic Graph $G = (V, D)$
Variable nodes $V = [n]$, $x_i \in \mathcal{X}$, for all $i \in V$
Define $\pi(i) \equiv \{\text{parents of } i\}$
Set of directed edges D

$$\mu(x) = \prod_{i \in V} \mu_i(x_i | x_{\pi(i)})$$



Bayesian network is specified by

- directed **acyclic** graph $G = (V, D)$
- alphabet $\mathcal{X}$
- conditional probability $\mu_i(\cdot|\cdot) : \mathcal{X} \times \mathcal{X}^{\pi(i)} \to \mathbb{R}_+$, for $i \in V$

$$\mu(x) = \prod_{i \in V} \mu_i(x_i|x_{\pi(i)})$$

- we do not need normalization $(1/Z)$ since

$$\sum_{x_i \in \mathcal{X}} \mu_i(x_i|x_{\pi(i)}) = 1 \;\; \Rightarrow \;\; \sum_{x \in \mathcal{X}^V} \mu(x) \; = \; 1$$

Conversion between Bayesian networks and factor graphs

from Bayesian network to factor graph

A Bayes network $G = (V, D)$ with alphabet $\mathcal{X}$ can be represented by a factor graph model on $G' = (V', F', E')$ with $V' = V$, $|F'| = |V|$, $|E'| = |D| + |V|$, $\mathcal{X}' = \mathcal{X}$.

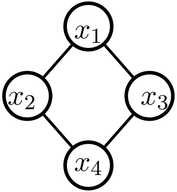
- represent by a factor node each conditional probability
- **moralization** for conversion from BN to MRF (we will learn this)

from factor graph to Bayesian network

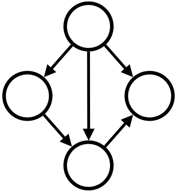
A factor model on $G = (V, F, E)$ with alphabet $\mathcal{X}$ can be represented by a Bayes network $G' = (V', D')$ with $V' = V$ and $\mathcal{X}' = \mathcal{X}$.

- take a topological ordering, e.g. $x_1, \ldots, x_n$
- for each node i, starting from the first node, find a minimal set $U \subseteq \{1, \ldots, i-1\}$ such that x_i is conditionally independent of $x_{\{1,\ldots,i-1\}\setminus U}$ given x_U. (we will learn how to do this)
- in general the resulting Bayesian network is dense

Because MRF and BN are incomparable, some independence structure is lost in conversion

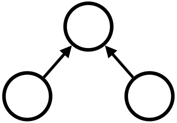


$\mu(x) = \psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)\psi(x_3, x_4)$
$x_1 \perp x_4 | (x_2, x_3)$
$x_2 \perp x_3 | (x_1, x_4)$

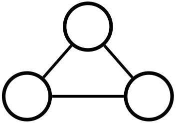


ordering:(x_1, x_2, x_4, x_3)

$x_2 \perp x_3 | (x_1, x_4)$



$\mu(x) = \mu(x_2)\mu(x_3)\mu(x_1 | x_2, x_3)$
$x_2 \perp x_3$



no independence

- undirected graphical models can be represented by factor graphs
  - we can go from MRF to FG without losing any information on the independencies implies by the model
- Bayesian networks are not compatible with undirected graphical models or factor graphs
  - if we go from one model to the other, and then back to the original model, then we will not, in general, get back the same model as we started out with
  - we lose any information on the independencies implies by the model, when switching from one model to the other

Bayes networks with observed variables

$V = H \cup O$

Hidden variables: $x = (x_i)_{i \in H}$

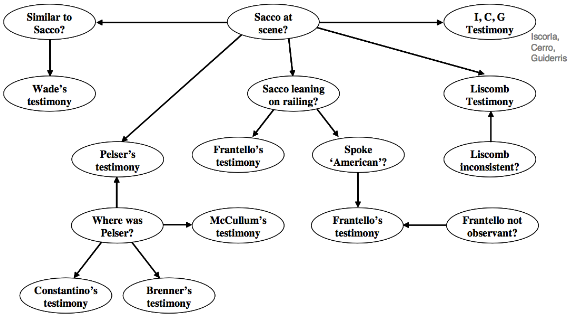
Observed variables: $y = (y_i)_{i \in O}$

$$\mu(x, y) = \prod_{i \in H} \mu(x_i | x_{\pi(i) \cap H}, y_{\pi(i) \cap O}) \prod_{i \in O} \mu(y_i | x_{\pi(i) \cap H}, y_{\pi(i) \cap O})$$

Typically interested in $\mu_y(x) \equiv \mu(x|y)$ and

$$\arg\max_x \quad \mu_y(x)$$

Example

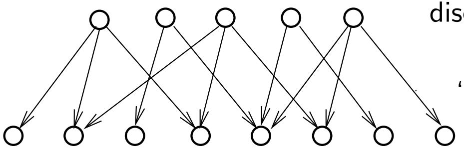


Forensic Science

[Kadane, Shum, *A probabilistic analysis of the Sacco and Vanzetti evidence*, 1996]

[Taroni et al., *Bayesian Networks and Probabilistic Inference in Forensic Science*, 2006]

Example



Medical Diagnosis

[M. Shwe, et al., Methods of Information in Medicine, 1991]

Roadmap

Cond. Indep. $\mu(x)$	Factorization $\mu(x)$	Graphical Model	Graph G	Cond. Indep. implied by G
x_1–$\{x_2, x_3\}$–x_4;	$\frac{1}{Z}\prod \psi_a(x_{\partial a})$	FG	Factor	Markov
x_4–$\{\}$–x_7;	$\frac{1}{Z}\prod \psi_C(x_C)$	MRF	Undirected	Markov
$\vdots$	$\prod \psi_i(x_i \vert x_{\pi(i)})$	BN	Directed	Markov

- A $\mu(x)$ can be represented by multiple {FG,MRF,BN} with multiple graphs (but same $\mu(x)$)
- We want a 'simple' graph representation (sparse, small alphabet size)
  - Memory to store the graphical model
  - Computations for inference
- $\mu(x)$ with some conditional independence structure can be represented by simple {FG,MRF,BN}