Machine learning: lecture 21

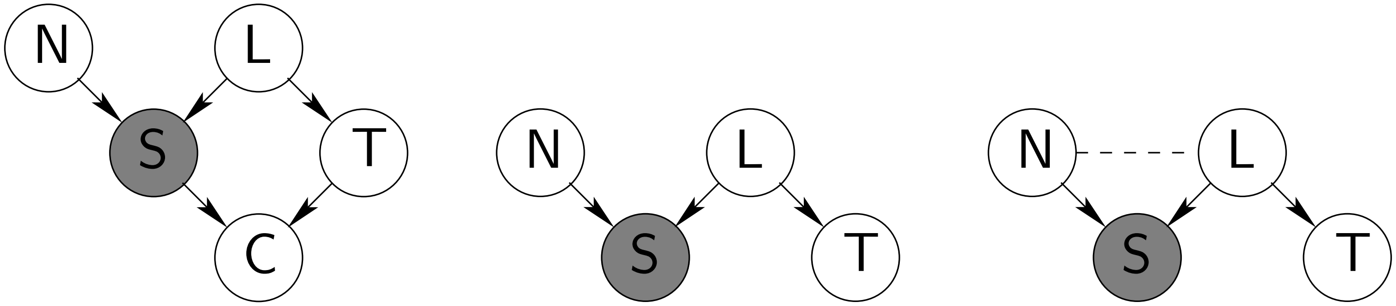
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Outline

- Bayesian networks, quantitative inference
  - medical diagnosis example
  - three inference problems
  - basic algorithms and problems

Bayesian networks: review

- Graph $\Rightarrow$ d-separation $\Rightarrow$ independence



  – conditional independence properties provide the basis for qualitative inferences

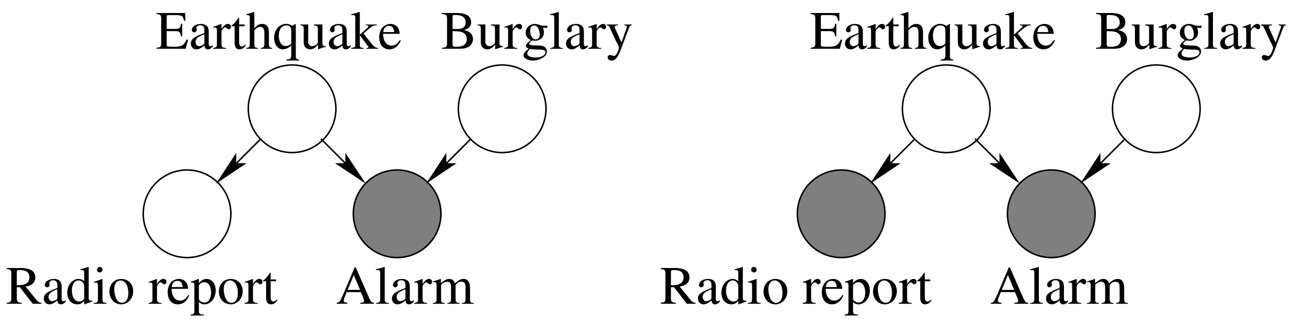
- Graph $\Rightarrow$ associated probability distribution

$$P(N)\,P(L)\,P(S|N,L)\,P(T|L)\,P(C|S,T)$$

(any distribution that factors in this manner is consistent with all the independence properties implied by the graph)

Bayesian networks: quantitative inferences

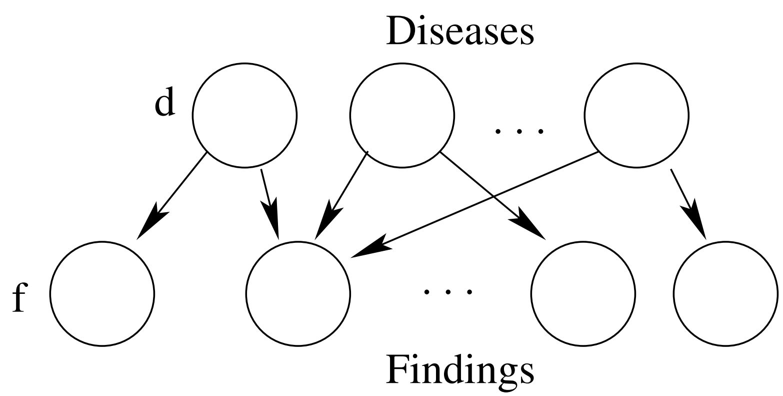
- In many cases the probabilities matter...

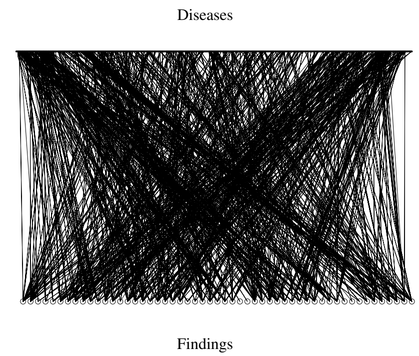


- We need to develop general purpose algorithms for making quantitative inferences with these models

Example setting: medical diagnosis

- The QMR-DT model (Shwe et al. 1991)

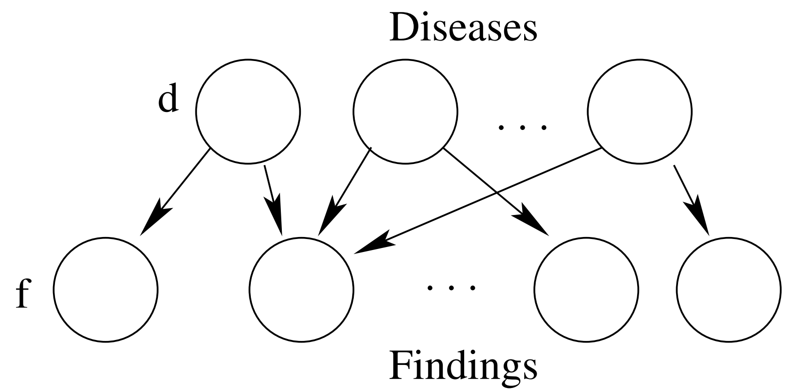




  - about 600 binary (0/1) disease variables representing diseases that are "present" or "absent"
  - about 4000 associated binary (0/1) findings; findings may be either "positive" or "negative"

Example cont'd

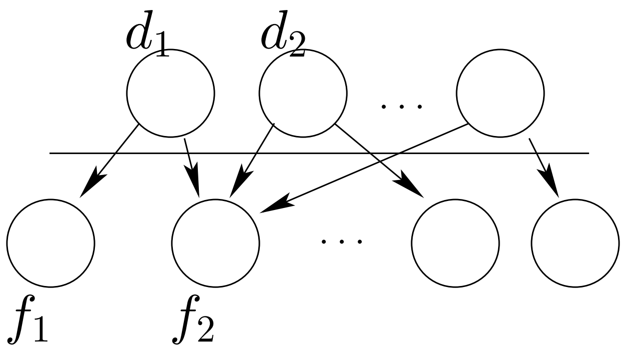
- The model is based on a number of simplifying assumptions



- Assumptions explicit in the graph:
  - relevant variables
  - marginal independence of diseases
  - conditional independence of findings
- Further assumptions about the probability distribution:
  - causal independence

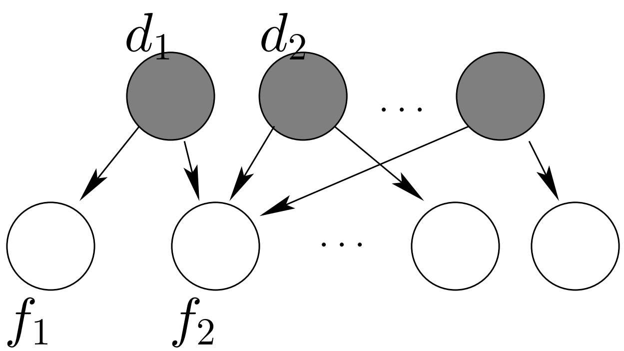
Assumptions in detail

- Diseases are marginally independent



$$
\begin{aligned}
d_1 &= \text{Hodgkins disease} \\
d_2 &= \text{Plasma cell myeloma} \\
d_3 &= \dots
\end{aligned}
$$

- Findings are conditionally independent given the diseases

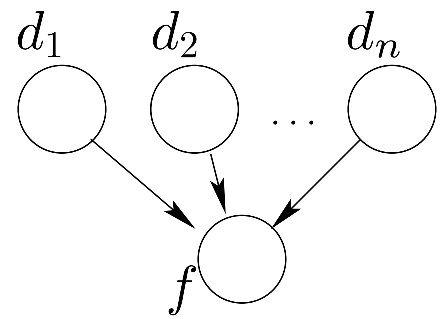


$f_1 = $ Bone X-ray fracture

$f_2 = \dots$

Assumptions in detail

- We have to specify how n (potentially 100 or more) underlying diseases conspire to influence any finding

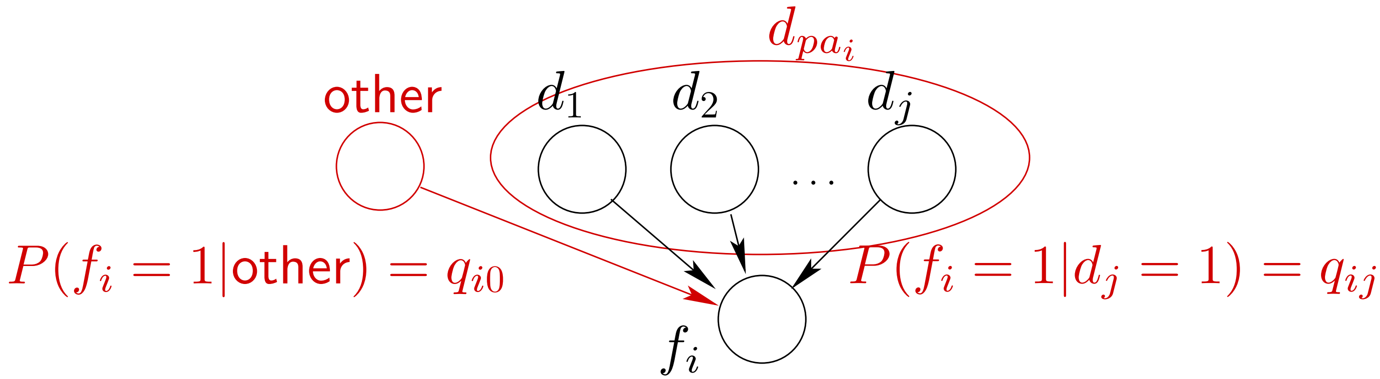


The size of the conditional probability table for $P(f|d_1, d_2, d_3, \dots)$ would increase exponentially with the number of associated diseases

$\Rightarrow$ e.g, causal independence assumption

Causal independence: noisy-or

- We assume that each finding is negative if all the associated diseases (if present) *independently* fail to produce a positive outcome



$$P(f_i = 0|d_{pa_i}) = P(f_i = 0|\text{other}) \prod_{j \in pa_i} P(f_i = 0|d_j)$$

$$= (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j}$$

and $P(f_i = 1|d_{pa_i}) = 1 - P(f_i = 0|d_{pa_i})$.

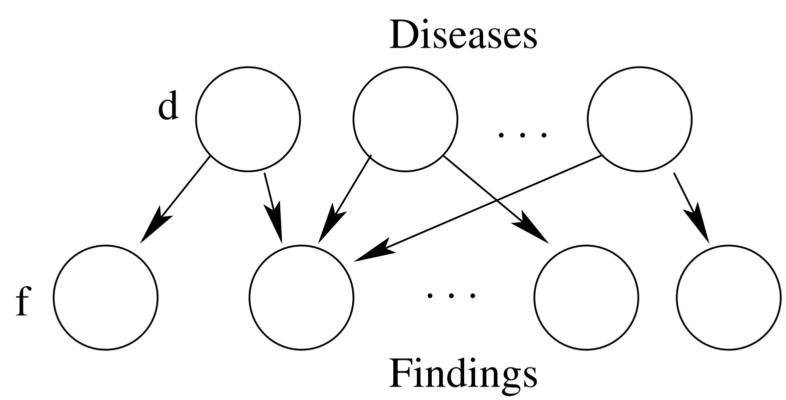
Joint distribution

- After all these assumptions, we can write down the following joint distribution over n diseases and m findings

$$P(f, d) = \left[\prod_{i=1}^{m} P(f_i|d_{pa_i})\right]\left[\prod_{j=1}^{n} P(d_j)\right]$$

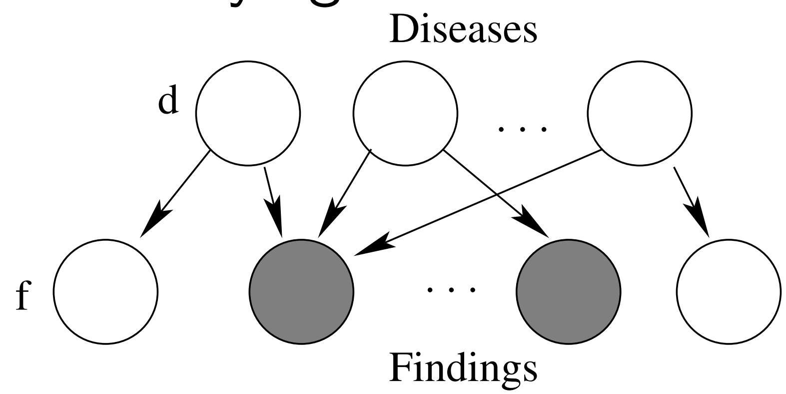
where $$P(f_i = 0|d_{pa_i}) = (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j}$$

The only adjustable parameters in this model are q_{ij} and $P(d_j)$



Three inference problems

- Given a set of observed findings $f^* = \{f_1^*, \ldots, f_k^*\}$, we wish to infer what the underlying diseases are



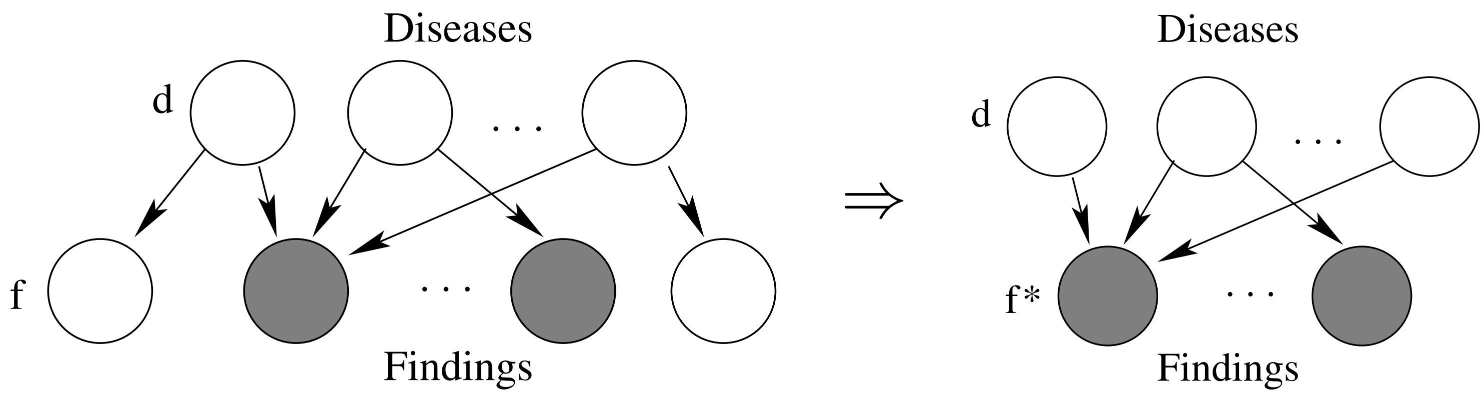
1. What are the marginal posterior probabilities over the diseases?

2. What is the most likely setting of all the underlying disease variables?

3. Which test should we carry out next in order to get the most information about the diseases?

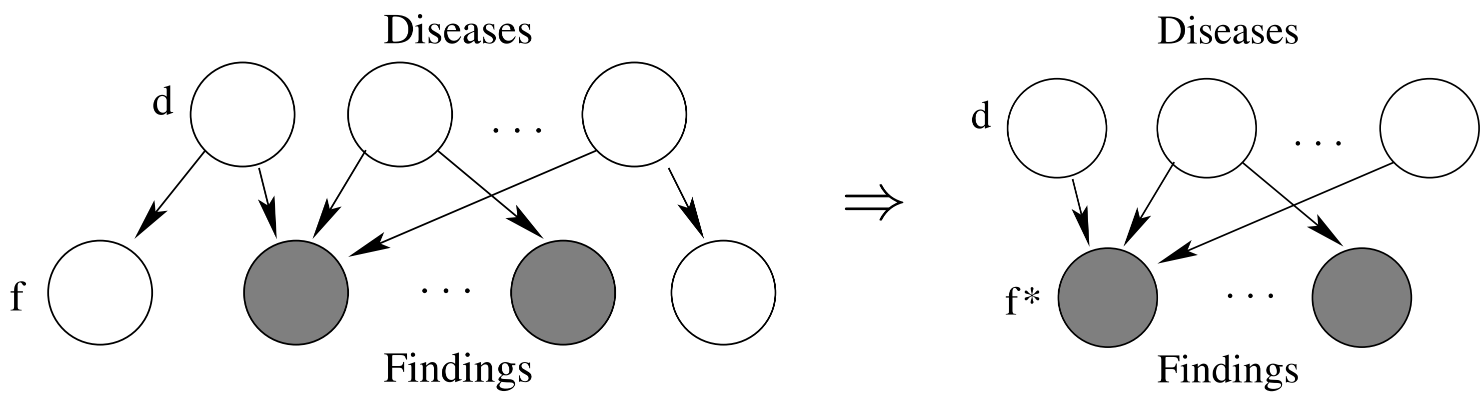
Inference problem cont'd

- For the purposes of inferring the presence or absence of the underlying diseases, we can ignore any findings that remain unobserved (as if they were not in the model to begin with)

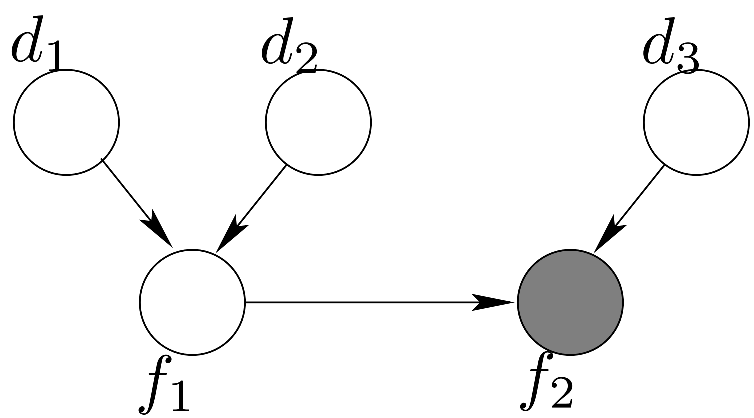


Inference problem cont'd

- For the purposes of inferring the presence or absence of the underlying diseases, we can ignore any findings that remain unobserved (as if they were not in the model to begin with)

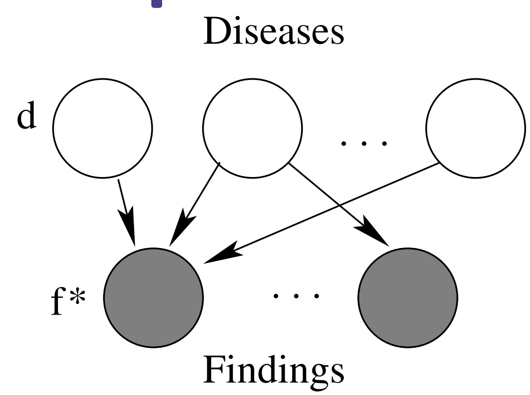


- What if the findings were not conditionally independent?

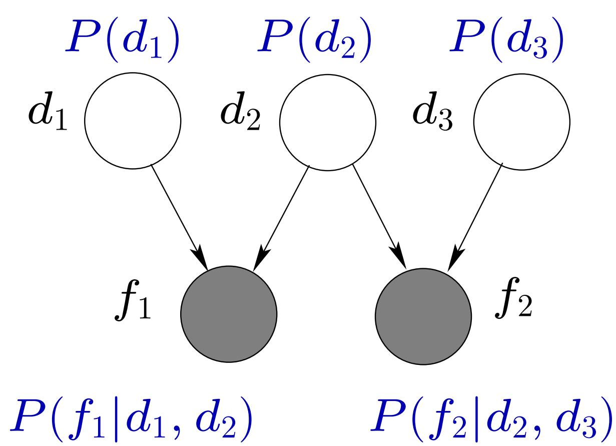


First inference problem: posterior marginals

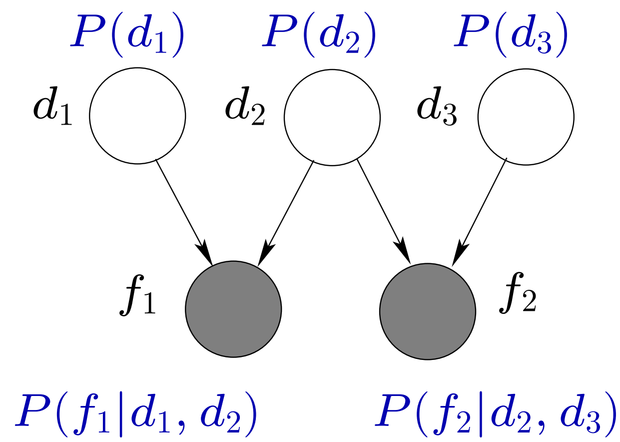
- Given the observations we already have all the information, only implicitly



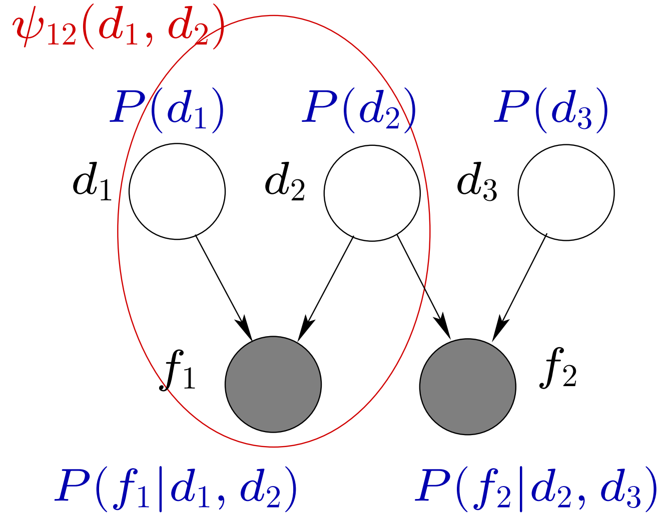
- What messages (if any) do the disease variables have to share for them to be able to compute the posterior marginals locally?



Inference: graph transformation

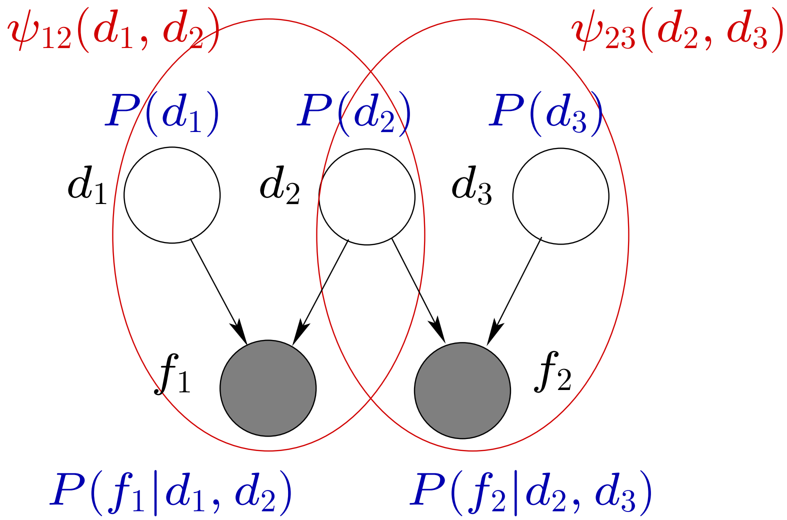


Inference: graph transformation



$$\psi_{12}(d_1, d_2) \;=\; P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$

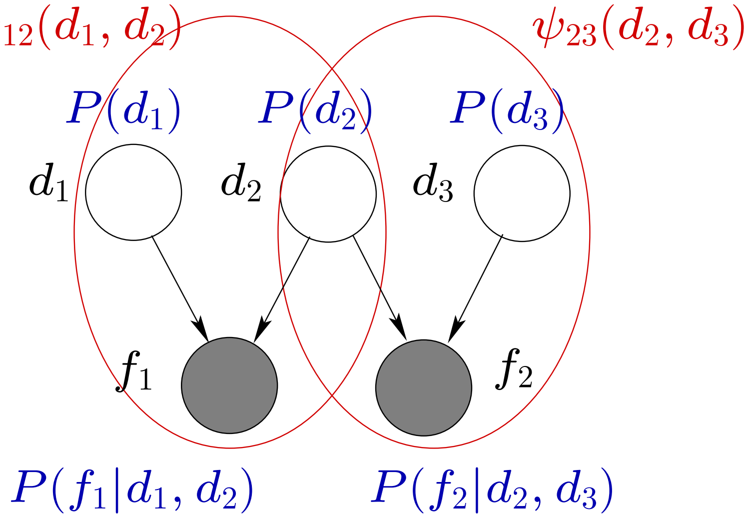
Inference: graph transformation



$$\psi_{12}(d_1, d_2) = P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$

$$\psi_{23}(d_2, d_3) = P(d_3)P(f_2^*|d_2, d_3)$$

Inference: graph transformation



$$\psi_{12}(d_1, d_2) = P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$

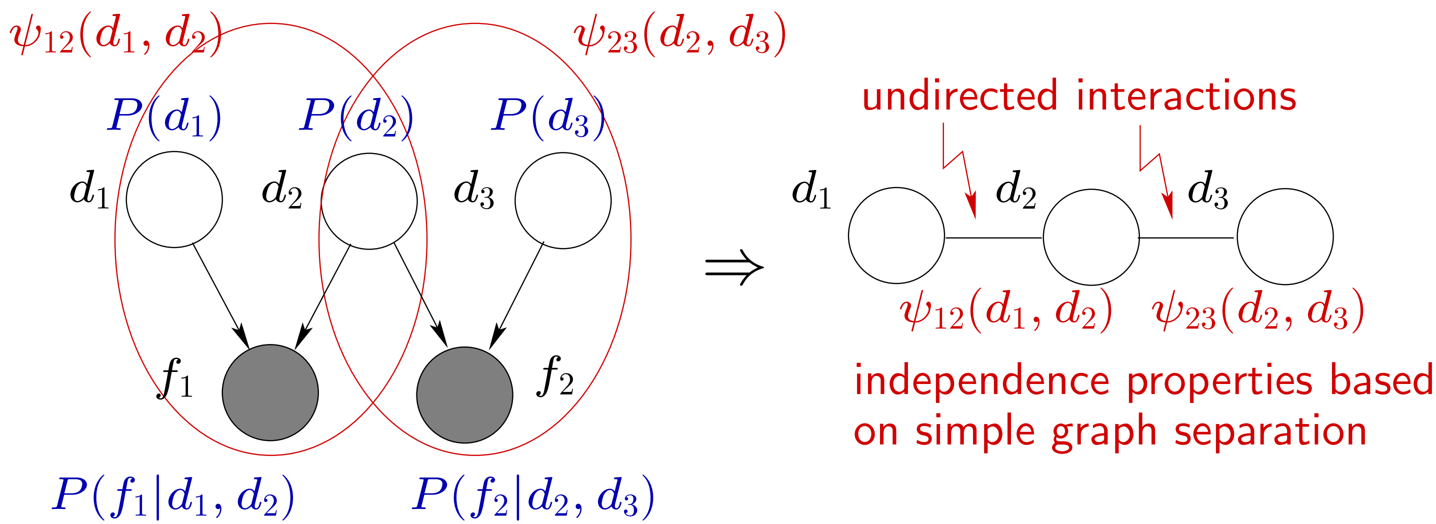
$$\psi_{23}(d_2, d_3) = P(d_3)P(f_2^*|d_2, d_3)$$

- Joint distribution as a product of "interaction potentials"

$$P(d_1, d_2, d_3, \text{data}) = \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3)$$

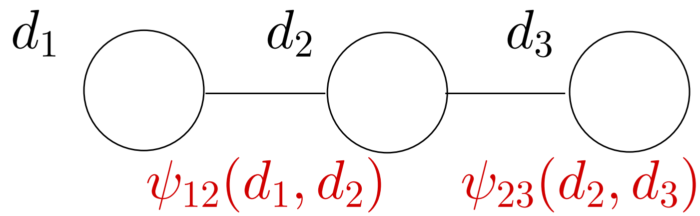
Inference: graph transformation

- We have transformed the Bayesian network into an undirected graph model (Markov random field):



$$P(d_1, d_2, d_3, \text{data}) = \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3)$$

Marginalization



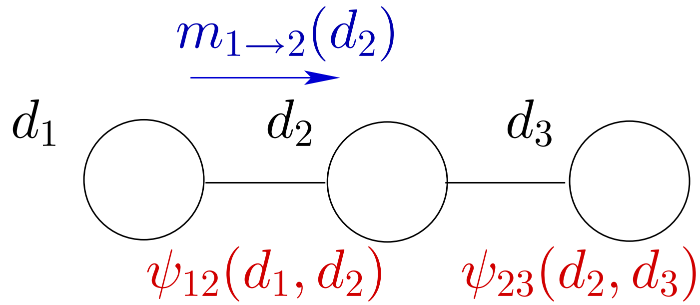
- It suffices to evaluate the following probabilities

$$
\begin{aligned}
P(d_1, \text{data}) &= \sum_{d_2, d_3} P(d_1, d_2, d_3, \text{data}) \\
P(d_2, \text{data}) &= \sum_{d_1, d_3} P(d_1, d_2, d_3, \text{data}) \\
P(d_3, \text{data}) &= \sum_{d_1, d_2} P(d_1, d_2, d_3, \text{data})
\end{aligned}
$$

These will readily yield the posterior probabilities of interest:

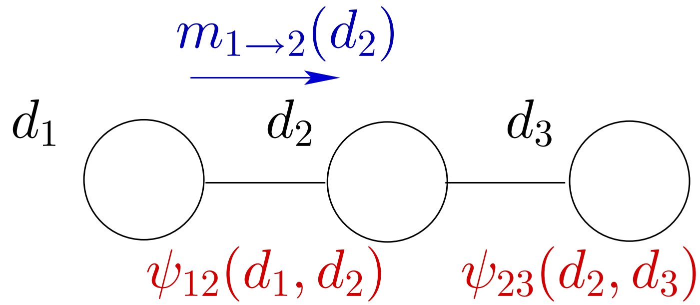
$$P(d_1|\text{data}) = P(d_1, \text{data}) / \sum_{d_1'} P(d_1', \text{data})$$

Marginalization and messages



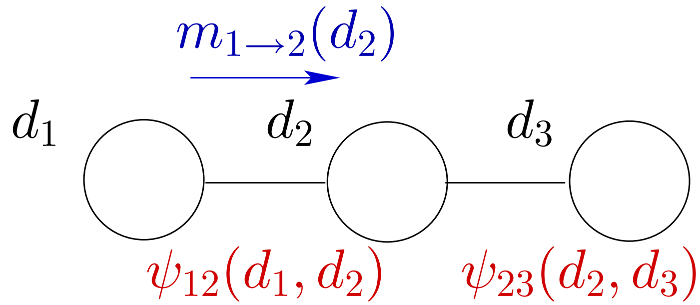
$$P(d_2, d_3, \mathsf{data}) \quad = \quad \sum_{d_1} P(d_1, d_2, d_3, \mathsf{data})$$

Marginalization and messages



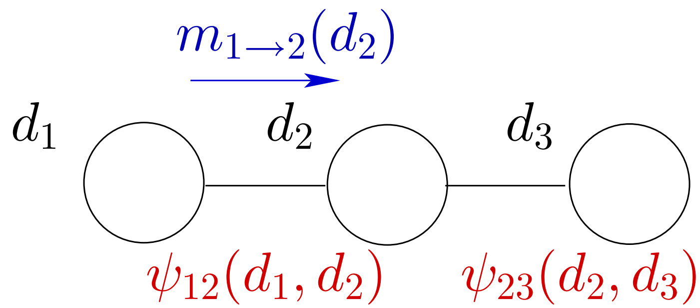
$$
\begin{aligned}
P(d_2, d_3, \mathsf{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \mathsf{data}) \\
&= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



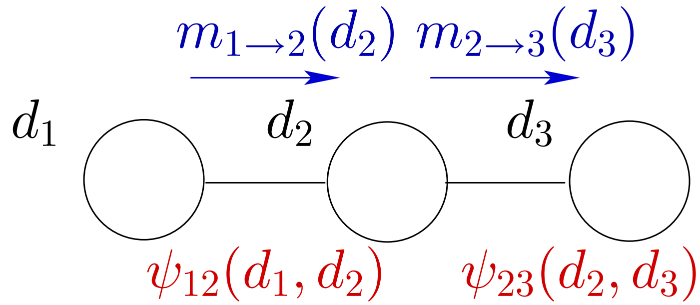
$$
\begin{aligned}
P(d_2, d_3, \text{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \text{data}) \\
&= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3) \\
&= \left[\sum_{d_1} \psi_{12}(d_1, d_2) \right] \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



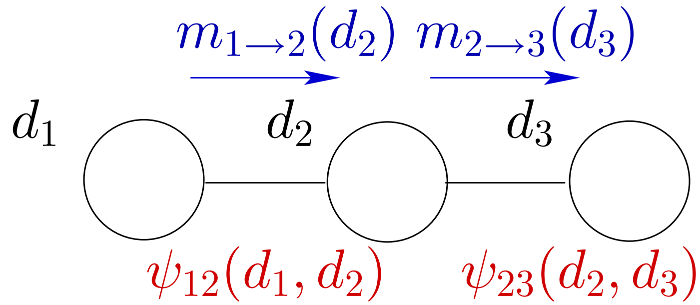
$$
\begin{aligned}
P(d_2, d_3, \mathsf{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \mathsf{data}) \\
&= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3) \\
&= \left[\sum_{d_1} \psi_{12}(d_1, d_2) \right] \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



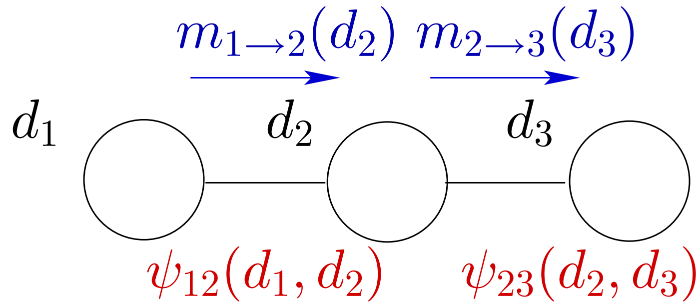
$$P(d_3, \mathsf{data}) \;\; = \;\; \sum_{d_2} P(d_2, d_3, \mathsf{data})$$

Marginalization and messages



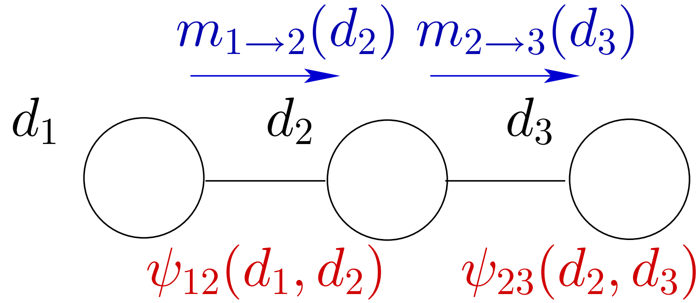
$$
\begin{aligned}
P(d_3, \mathsf{data}) &= \sum_{d_2} P(d_2, d_3, \mathsf{data}) \\
&= \sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \cdot 1
\end{aligned}
$$

Marginalization and messages



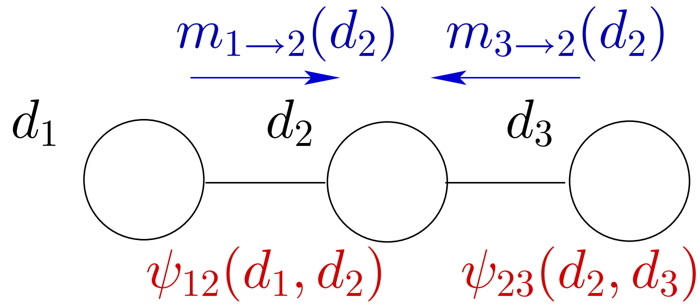
$$
\begin{aligned}
P(d_3, \mathsf{data}) &= \sum_{d_2} P(d_2, d_3, \mathsf{data}) \\
&= \sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \cdot 1 \\
&= \left[\sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \right] \cdot 1
\end{aligned}
$$

Marginalization and messages



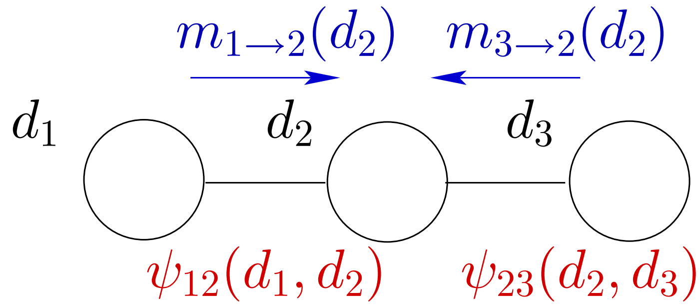
$$
\begin{aligned}
P(d_3, \mathsf{data}) &= \sum_{d_2} P(d_2, d_3, \mathsf{data}) \\
&= \sum_{d_2} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \cdot 1 \\
&= \left[\sum_{d_2} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \right] \cdot 1 \\
&= m_{2\to3}(d_3) \cdot 1
\end{aligned}
$$

Marginalization and messages



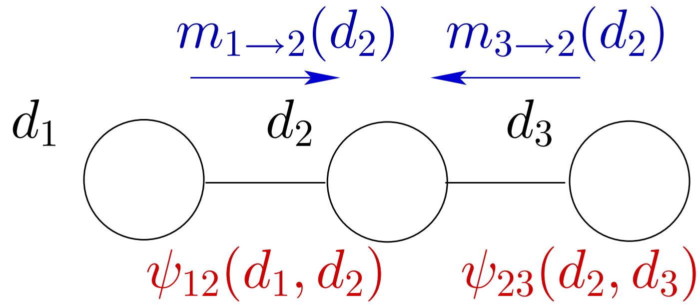
$$P(d_2, \text{data}) \quad = \quad \sum_{d_3} P(d_2, d_3, \text{data})$$

Marginalization and messages



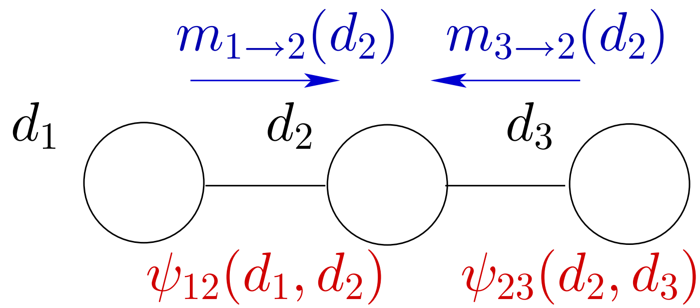
$$
\begin{aligned}
P(d_2, \mathsf{data}) &= \sum_{d_3} P(d_2, d_3, \mathsf{data}) \\
&= \sum_{d_3} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right]
\end{aligned}
$$

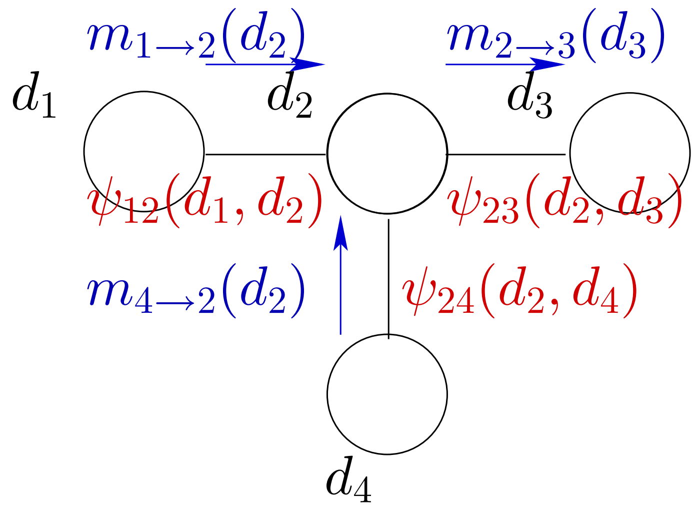
Marginalization and messages



$$
\begin{aligned}
P(d_2, \textsf{data}) &= \sum_{d_3} P(d_2, d_3, \textsf{data}) \\
&= \sum_{d_3} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to 2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right] \\
&= m_{1\to 2}(d_2) \cdot m_{3\to 2}(d_2)
\end{aligned}
$$

Message passing and trees

- The same message passing approach (belief propagation) works for any tree structured model

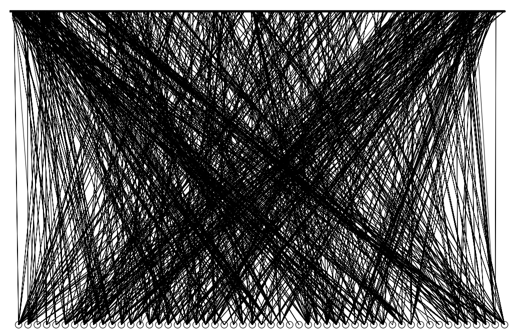


$$m_{2\to 3}(d_3) = \sum_{d_2} m_{1\to 2}(d_2) m_{4\to 2}(d_2) \psi_{23}(d_2, d_3)$$

$$P(d_2, \text{data}) = ?$$

Back to the diagnosis problem

- This does not look like a tree...



  - clusters of variables ⇒ tree over clusters
  - approximate inference