Graphical probabilistic models – inference

Jiří Kléma

Department of Cybernetics,
FEE, CTU at Prague

http://ida.felk.cvut.cz

Agenda

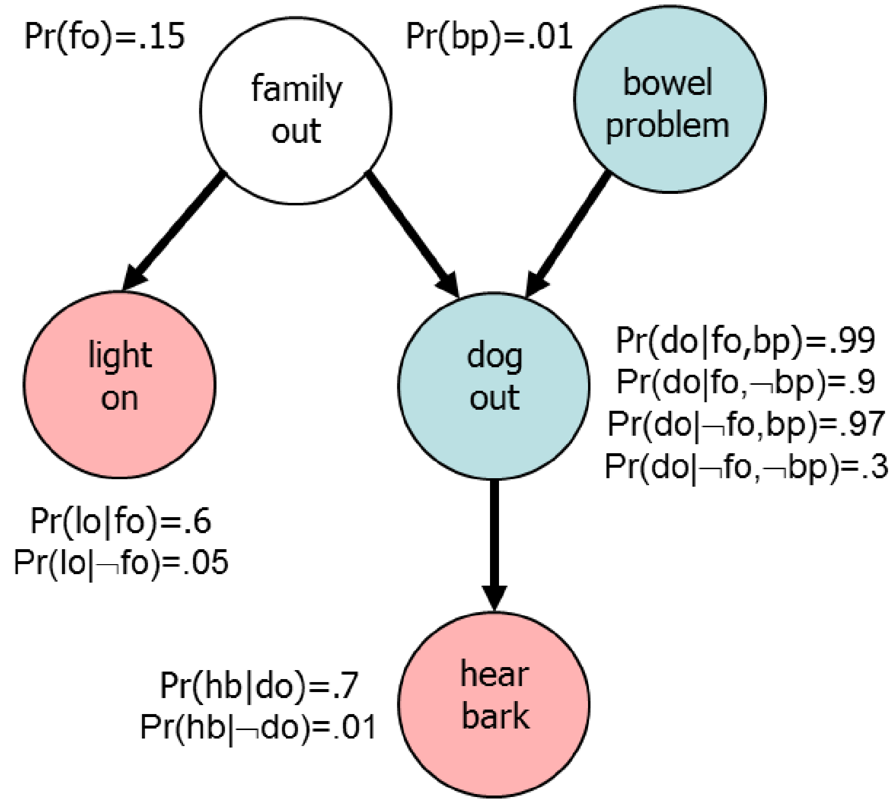
- Bayesian networks
  - fundamental tasks,
- **inference** and its complexity
  - straightforward enumeration
    - easy to understand but inefficient – computes joint probabilities,
    - descends to the level of atomic events,
    - acceleration by variable elimination,
  - limitations $\times$ efficiency of algorithms,
  - exact $\times$ approximate algorithms,
  - particular “fast” algorithms
    - belief propagation,
    - junction tree,
    - arc reversal,
    - Gibbs sampling.

Bayesian networks – fundamental tasks

- inference – reasoning, deduction
  - from observed events assumes on probability of other events,
  - observations (**E** – a set of evidence variables, **e** – a particular event),
  - target variables (**Q** – a set of query variables, Q – a particular query variable),
  - $Pr(\mathbf{Q}|\mathbf{e})$, resp. $Pr(Q \in \mathbf{Q}|\mathbf{e})$ to be found,
  - network is known (both graph and CPTs),
- learning network parameters from data
  - network structure (graph) is given,
  - "only" quantitative parameters (CPTs) to be optimized,
- learning network structure from data
  - propose an optimal network structure
    - which edges of the complete graph shall be employed?,
  - too many arcs $\longrightarrow$ complicated model,
  - too few arcs $\longrightarrow$ inaccurate model.

Probabilistic network – inference by enumeration

- Let us observe the following events:
  - no barking heard,
  - the door light is on.
- What is the prob of family being out?
  - searching for $Pr(fo|lo, \neg hb)$.
- Will observation influence the target event?
  - light on supports departure hypothesis,
  - no barking suggests dog inside,
  - the dog is in house when it is
    - rather healthy,
    - the family is at home.



Probabilistic network – inference by enumeration

- inference by enumeration
  - conditional probs calculated by summing the elements of joint probability table,
- how to find the joint probabilities (the table is not given)?
  - BN definition suggests:

    $$Pr(FO, BP, DO, LO, HB) =$$
    $$= Pr(FO)Pr(BP)Pr(DO|FO, BP)Pr(LO|FO)Pr(HB|DO)$$
- answer to the question?
  - conditional probability definition suggests:

    $Pr(fo|lo, \neg hb) = \frac{Pr(fo,lo,\neg hb)}{Pr(lo,\neg hb)}$
  - by joint prob marginalization we get:

    $Pr(fo, lo, \neg hb) = \sum_{BP,DO} Pr(fo, BP, DO, lo, \neg hb)$

    $Pr(fo, lo, \neg hb) = Pr(fo, bp, do, lo, \neg hb) + Pr(fo, bp, \neg do, lo, \neg hb)+$
    $+Pr(fo, \neg bp, do, lo, \neg hb) + Pr(fo, \neg bp, \neg do, lo, \neg hb) = .15 \times .01 \times .99 \times .6 \times .3 + .15 \times$
    $.01 \times .01 \times .6 \times .99 + .15 \times .99 \times .9 \times .6 \times .3 + .15 \times .99 \times .1 \times .6 \times .99 = .033$

    $Pr(lo, \neg hb) = Pr(fo, lo, \neg hb) + Pr(\neg fo, lo, \neg hb) = .066$

Probabilistic network – inference by enumeration

- after substitution:
  $Pr(fo|lo,\neg hb) = \frac{Pr(fo,lo,\neg hb)}{Pr(lo,\neg hb)} = \frac{.033}{.066} = 0.5$
- posterior probability $Pr(fo|lo,\neg hb)$ is higher then the prior $Pr(fo) = 0.15$.

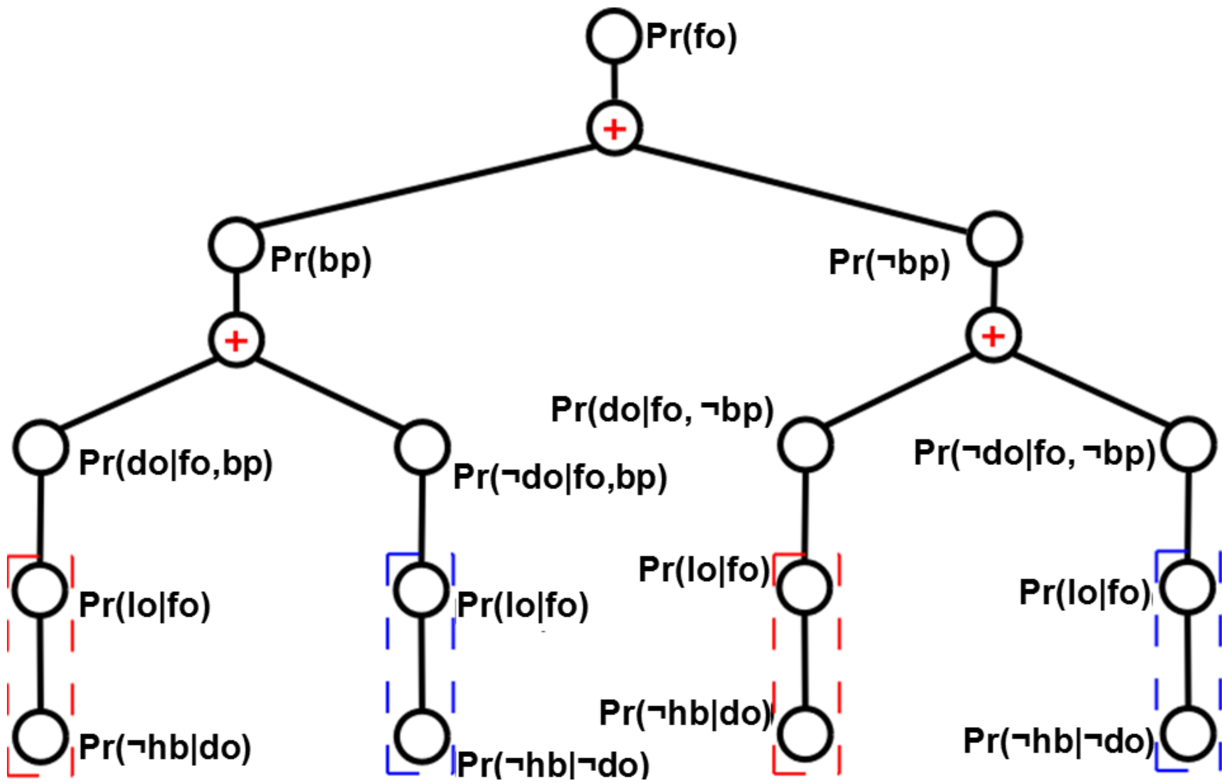
■ can we assume on complexity?

- instead of $2^5 - 1$=31 probs (either conditional or joint) 10 is needed only,
- however, joint probs are enumerated to answer the query
  - ∗ it is easy to show that inference remains a NP problem,
- to simply move summations right-to-left makes a difference, but not a principal one
  - ∗ see the evaluation tree on the next slide,

$$Pr(fo,lo,\neg hb) = \sum_{BP,DO} Pr(fo,BP,DO,lo,\neg hb) =$$
$$= Pr(fo)\sum_{BP} Pr(BP)\sum_{DO} Pr(DO|fo,BP)Pr(lo|fo)Pr(\neg hb|DO)$$

- inference by enumeration is an intelligible, but unfortunately inefficient procedure,
- solution: minimize recomputations, special network types or approximate inference.

Inference by enumeration – evaluation tree



- Complexity: time $\mathcal{O}(n2^d)$, memory $\mathcal{O}(n)$
  - n ... the number of variables, e ... the number of evidence variables, d=n-e,
- resource of inefficiency: recomputations ($Pr(lo|fo) \times Pr(\neg hb|DO)$ for each BP value)
  - variable ordering makes a difference – $Pr(lo|fo)$ shall be moved forward.

Inference by enumeration – straightforward improvements

- **variable elimination** procedure

  1. pre-computes **factors** to remove the inefficiency shown in the previous slide
     - factors serve for recycling the earlier computed intermediate results,
     - some variables are eliminated by summing them out,

       $\sum_P f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \times \sum_P f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{P}}$,
       assumes that $f_1, \ldots, f_i$ do not depend on P,

       when multiplying factors, the pointwise product is computed
       $f_1(x_1, ..., x_j, y_1, ..., y_k) \times f_2(y_1, ..., y_k, z_1, ..., z_l) = f(x_1, ..., x_j, y_1, ..., y_k, z_1, ..., z_l)$

       eventual enumeration over P_1 variable, which takes all (two) possible values
       $f_{\bar{P}_1}(P_2, ..., P_k) = \sum_{P_1} f_1(P_1, P_2, ..., P_k)$,

     - execution efficiency is influenced by the variable ordering when computing,
       (finding the best order is NP-complete problem, can be optimized heuristically too),

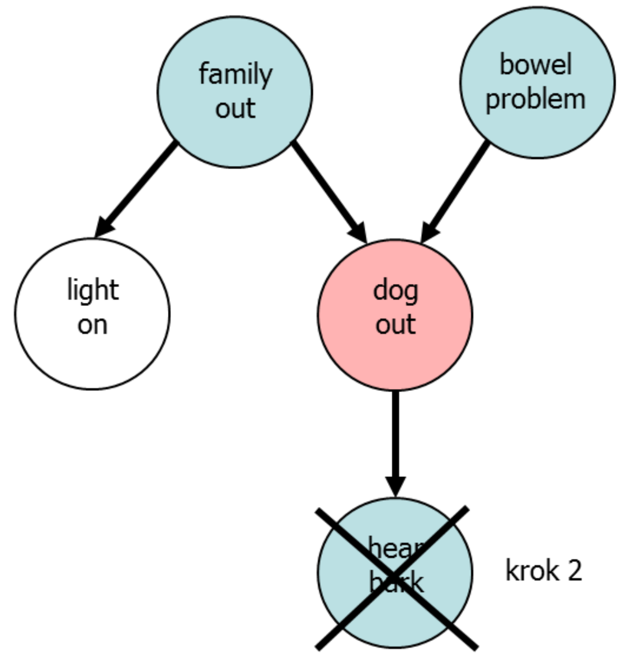
Inference by enumeration – straightforward improvements

- **variable elimination** procedure

  2. does not consider variables irrelevant to the query
     - all the leaves that are neither query nor evidence variable,
     - the rule can be applied recursively.

- example: $Pr(lo|do)$
  - what is prob that the door light is shining if the dog is in the garden?
  - we will enumerate $Pr(LO, do)$, since:

$$Pr(lo|do) = \frac{Pr(lo,do)}{Pr(do)} = \frac{Pr(lo,do)}{Pr(lo,do)+Pr(\neg lo,do)}$$



Inference by enumeration – variable elimination

- HB is irrelevant to the particular query, why?

$$\sum_{HB} Pr(HB|do) = 1$$

$$Pr(LO, do) = \sum_{FO,BP,HB} Pr(FO)Pr(BP)Pr(do|FO, BP)Pr(LO|FO)Pr(HB|do) =$$
$$= \sum_{HB} Pr(HB|do) \sum_{FO} Pr(FO)Pr(LO|FO) \sum_{BP} Pr(BP)Pr(do|FO, BP)$$

- after omitting the first invariant, **factorization** may take place

$$Pr(LO, do) = \sum_{FO} Pr(FO)Pr(LO|FO) \sum_{BP} Pr(BP)Pr(do|FO, BP) =$$
$$= \sum_{FO} Pr(FO)Pr(LO|FO) f_{\overline{BP}}(do|FO) = \sum_{FO} f_{\overline{BP},do}(FO)Pr(LO|FO) =$$
$$= f_{\overline{FO},\overline{BP},do}(LO)$$

- having the last factor (a table of two elements), one can read

$$Pr(lo|do) = \frac{f_{\overline{FO},\overline{BP},do}(lo)}{f_{\overline{FO},\overline{BP},do}(lo)+f_{\overline{FO},\overline{BP},do}(\neg lo)} = \frac{0.0941}{0.0941+0.3017} = \frac{0.0941}{0.3958} = 0.24$$

Variable elimination – factor computations

- factors are enumerated from CPTs by summing out variables
  - sum out BP: $CPT(DO)$ & $CPT(BP) \rightarrow f_{\overline{BP}}(do|FO)$
  - reformulate into: $CPT(FO)$ & $f_{\overline{BP}}(do|FO) \rightarrow f_{\overline{BP},do}(FO)$
  - sum out FO: $f_{\overline{BP},do}(FO)$ & $CPT(LO) \rightarrow f_{\overline{FO},\overline{BP},do}(LO)$

BP	$Pr(BP)$
T	0.01
F	0.99

$\times$

FO	BP	$Pr(do\|FO,BP)$
T	T	0.99
T	F	0.9
F	T	0.97
F	F	0.3

$\Rightarrow$

FO	$f_{\overline{BP}}(do\|FO)$
T	$0.9009{=}0.99{\times}0.01{+}0.9{\times}0.99$
F	$0.3067{=}0.97{\times}0.01{+}0.99{\times}0.3$

FO	$Pr(FO)$
T	0.15
F	0.85

$\times$

FO	$f_{\overline{BP}}(do\|FO)$
T	0.9009
F	0.3067

$\Rightarrow$

FO	$f_{\overline{BP},do}(FO)$
T	$0.1351{=}0.15{\times}0.9009$
F	$0.2607{=}0.85{\times}0.3067$

FO	$f_{\overline{BP},do}(FO)$
T	0.1351
F	0.2607

$\times$

LO	FO	$Pr(LO\|FO)$
T	T	0.6
T	F	0.05
F	T	0.4
F	F	0.95

$\Rightarrow$

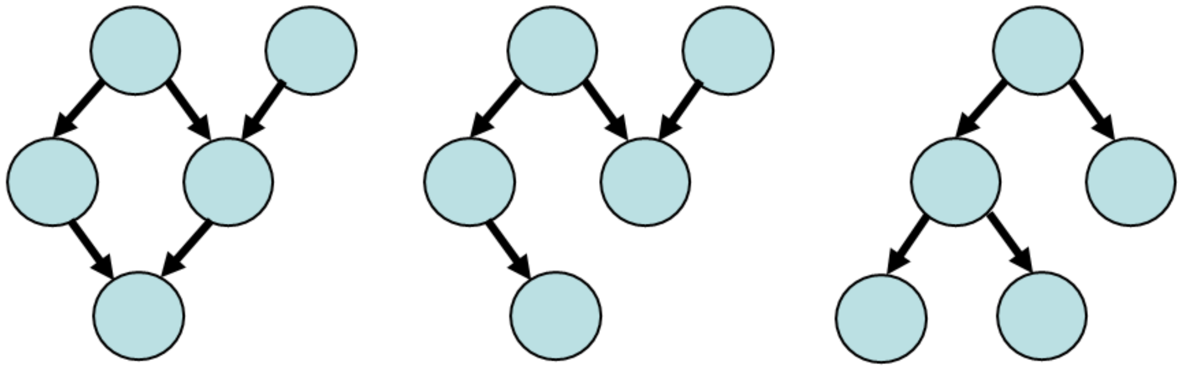
LO	$f_{\overline{FO},\overline{BP},do}(LO)$
T	$0.0941{=}0.1351{\times}0.6{+}0.2607{\times}0.05$
F	$0.3017{=}0.1351{\times}0.4{+}0.2607{\times}0.95$

Inference by enumeration – comparison of the number of operations

- let us take the last example
  - namely the total number of sums and products in $Pr(LO, do)$,
  - (the final $Pr(lo|do)$ enumaretion is identical for all procedures),
- naïve enumeration, no evaluation tree
  - 4 products (5 vars) $\times 2^4$ ($\#$ atomic events on unevidenced variables) $+\ 2^4 - 1$ sums,
  - in total 79 operations,
- using evaluation tree and a proper reordering of variables
  - in total 33 operations,
- with variable elimination on top of that
  - in total 14 operations (6 in Tab1, 2 in Tab2, 6 in Tab3).

Inference in networks without undirected cycles

- polytree (singly connected network, directed graph without undirected cycles),



DAG
(multiply connected net)

polytree
(singly connected net)

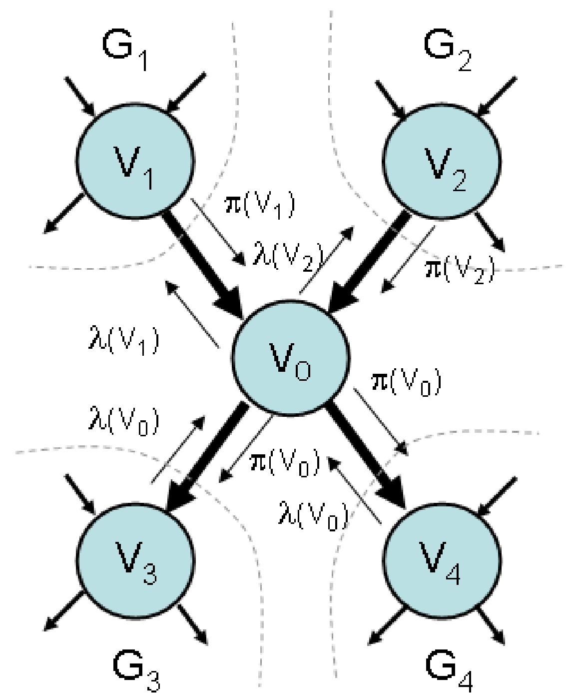
directed tree
(strict definition)

- **belief propagation** or polytree algorithm – J. Pearl
  - exact algorithm,
  - time complexity proportional with network diameter
    * and thus polynomial with the number of variables at worst,
    * and linear wrt the number of network parameters (CPTs).

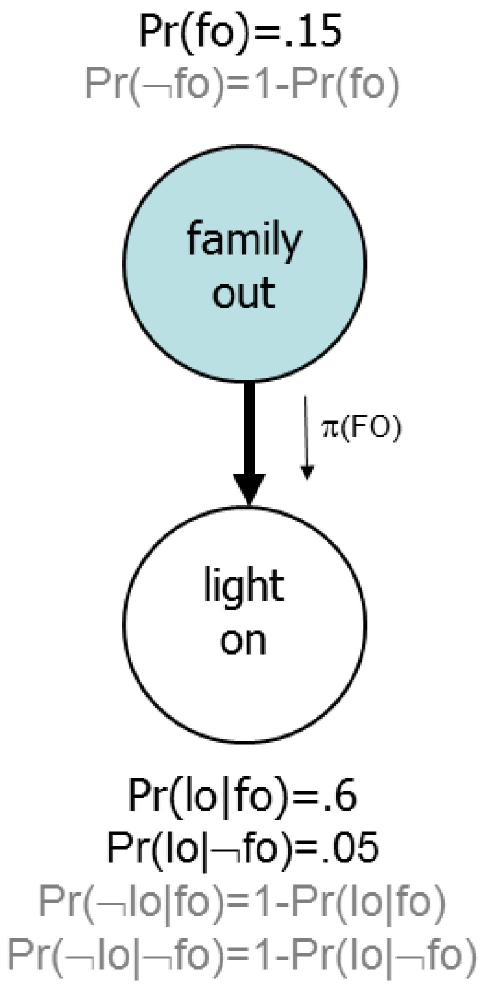
Belief propagation

- local evidence in V node changes $Pr(V)$ to $Pr^*(V)$, network needs to be updated,
- each evidence node sends a message about the evidence to its children and parents,
- every node updates its belief in all of its possible values based on the neighbor messages and further sends this evidence to its neighbors,

- two reasoning types can be distinguished
  - **causal** (π)
    evidence is propagated from ancestor nodes,
  - **diagnostic** (λ)
    evidence is propagated from descendant nodes,
- object-oriented method
  - nodes = objects, edges = communication channels,
  - time stamps on variable (node) states needed
    (an iteration index must be concerned).

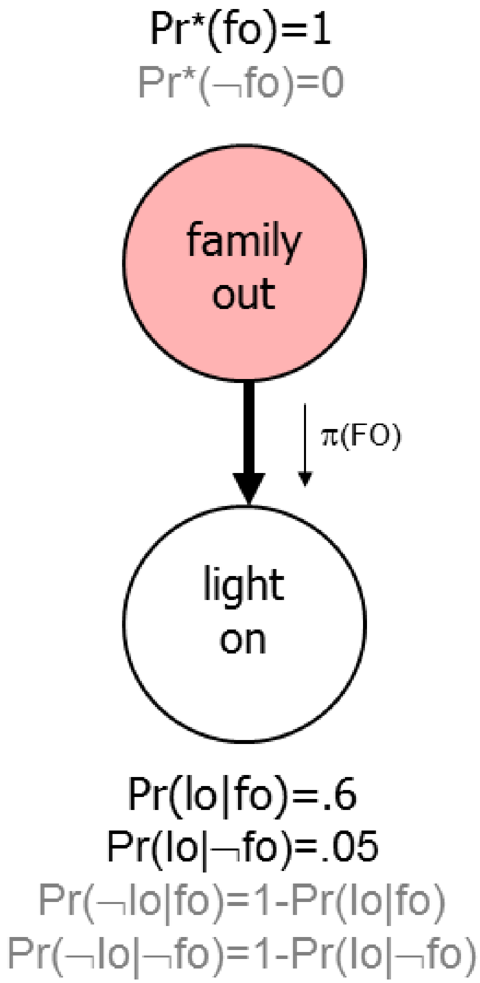


Belief propagation – causal evidence message passing



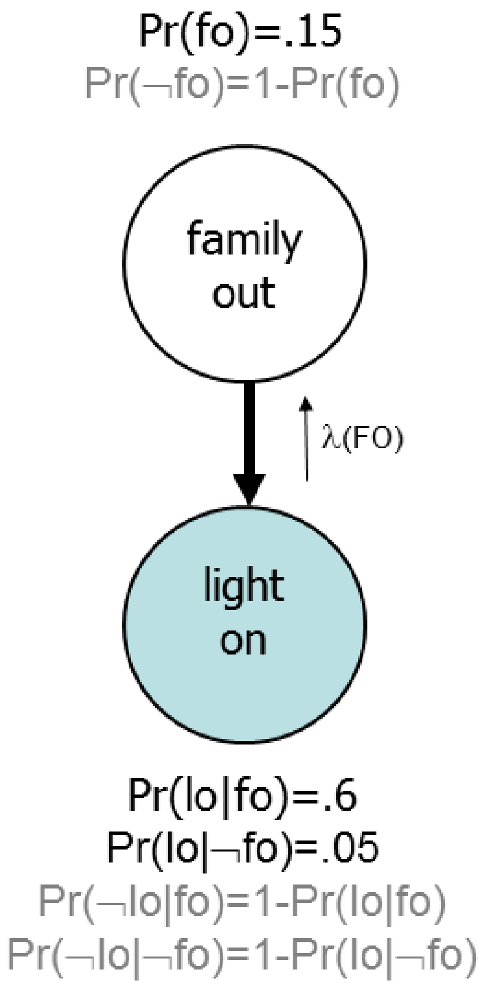
- Aim: find prob, that light is on in the trivial network on the left
  - neither $Pr(lo)$ (nor $\neg Pr(lo)$) can be computed from local data purely,
  - $Pr(lo) = Pr(lo|fo) \times Pr(fo) + Pr(lo|\neg fo) \times Pr(\neg fo)$,
- FO node must pass a message to LO: $\pi_{LO}^{FO}(FO) = Pr(FO)$
  - provided FO is not an evidence node, it sends its prior prob
  - $\pi_{LO}^{FO}(fo) = Pr(fo)$, $\pi_{LO}^{FO}(\neg fo) = Pr(\neg fo)$,
  - $Pr(lo) = Pr(lo|fo)\pi_{LO}^{FO}(fo) + Pr(lo|\neg fo)\pi_{LO}^{FO}(\neg fo)$
  - $Pr(lo) = .6 \times .15 + .05 \times .85 = .1325$

Belief propagation – causal evidence message passing



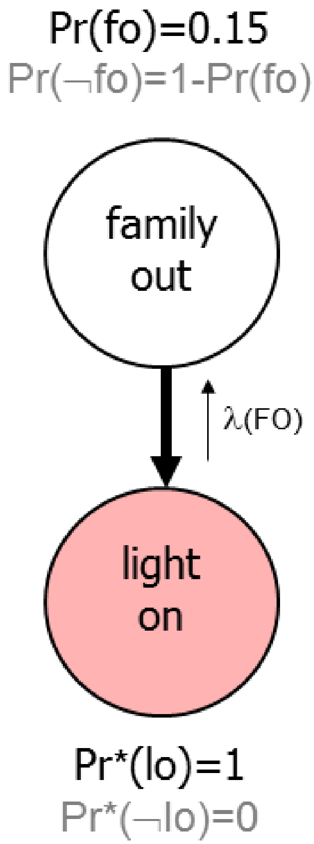
- Aim: find prob, that light is on in the trivial network on the left
  - neither $Pr(lo)$ (nor $\neg Pr(lo)$) computable from local data purely,
  - $Pr(lo) = Pr(lo|fo) \times Pr(fo) + Pr(lo|\neg fo) \times Pr(\neg fo)$,
- FO node must pass a message to LO: $\pi_{LO}^{FO}(FO) = Pr(FO)$
  - knowing that family left the house, FO node sends $\pi_{LO}^{FO}(fo) = Pr^*(fo) = 1, \pi_{LO}^{FO}(\neg fo) = Pr^*(\neg fo) = 0$
  - $Pr^*(lo) = Pr(lo|fo) = .6$
  - $Pr^*(\neg lo) = Pr(\neg lo|fo) = .4$

Belief propagation – diagnostic evidence message passing



- Aim: find prob that family is out in the trivial net on the left
- let us take the situation when the child node is not observed
  - neither $Pr(fo)$ (nor $\neg Pr(fo)$) is a function of $Pr(lo)$,
  - LO passes $\lambda_{LO}^{FO}(FO) = 1$ (invariant in further computations).

Belief propagation – diagnostic evidence message passing



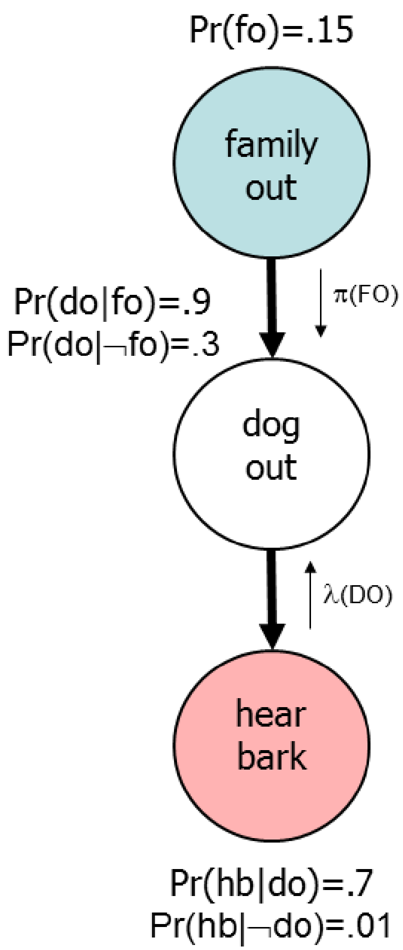
- Aim: find prob that family left the house in the trivial net on the left
- provided that $Pr^*(lo) = 1$ is given
  - we search for $Pr^*(fo) = Pr(fo|lo)$,
  - from Bayes theorem $Pr(fo|lo) = \frac{Pr(lo|fo)Pr(fo)}{Pr(lo)}$
  - two values are actually needed: $Pr(lo|fo)$ and $Pr(lo)$
  - however, $Pr(lo)$ is unknown (only $Pr^*(lo) = 1$)
  - LO passes $\lambda_{LO}^{FO}(fo) = Pr(lo|fo)$ and $\lambda_{LO}^{FO}(\neg fo) = Pr(lo|\neg fo)$
  - and makes use of normalization $Pr^*(fo) + Pr^*(\neg fo) = 1$

$$Pr^*(fo) = \alpha\lambda_{LO}^{FO}(fo)Pr(fo) = \alpha \times .6 \times .15 = .09\alpha$$
$$Pr^*(\neg fo) = \alpha\lambda_{LO}^{FO}(\neg fo)Pr(\neg fo) = \alpha \times .05 \times .85 = .0425\alpha$$

$$Pr^*(fo) + Pr^*(\neg fo) = 1 \longrightarrow .09\alpha + .0425\alpha = 1$$
$$\alpha = 1/.1325 \cong 7.55$$

  - it can be inferred that $\alpha = \frac{1}{Pr(lo)}$
  - $Pr^*(fo) = 0.68,\ Pr^*(\neg fo) = 0.32$

Belief propagation – combined propagation

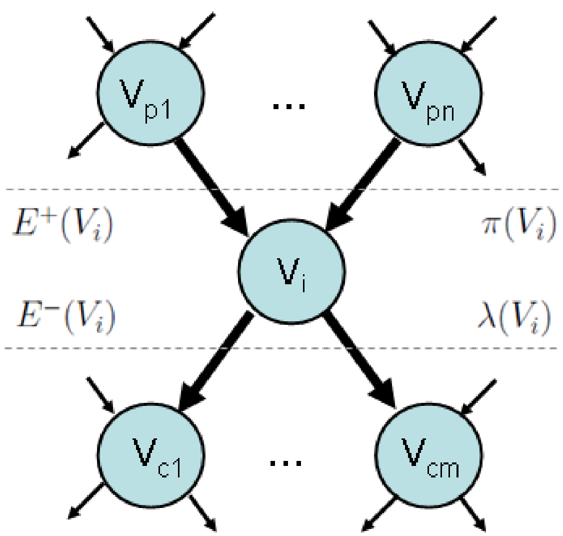


- Aim: find prob that the dog is out knowing it barks,
- child is observed, parent is unobserved,
- finding $Pr^*(DO)$ asks both for causal and diagnostic inference
  $\pi^{FO}_{DO}(fo) = Pr(fo)$, $\pi^{FO}_{DO}(\neg fo) = Pr(\neg fo)$
  $\lambda^{DO}_{HB}(do) = Pr(hb|do)$, $\lambda^{DO}_{HB}(\neg do) = Pr(hb|\neg do)$

  $$\begin{aligned} Pr^*(do) &= \alpha\lambda^{DO}_{HB}(do)[Pr(do|fo)\pi^{FO}_{DO}(fo) + \\ &\quad + Pr(do|\neg fo)\pi^{FO}_{DO}(\neg fo)] = \\ &= .7\alpha[.9 \times .15 + .3 \times .85] = \alpha \times .7 \times .39 = .273\alpha \end{aligned}$$

  $Pr^*(\neg do) =$ analogically $= 6.1 \times 10^{-3}\alpha$
  $\alpha \cong 3.58$, $Pr^*(do) \cong .98$, $Pr^*(\neg do) \cong .02$
- if we generalize
  - $Pr^*(DO) = Pr(DO|Evidence) = \alpha \times \pi(DO) \times \lambda(DO)$
  - α – normalization constant,
  - $\pi(DO)$ – **compound** causal parameter,
  - $\lambda(DO)$ – **compound** diagnostic parameter.

Belief propagation – combined propagation



- the evidence set with respect to V_i: $E = E^+(V_i) \cup E^-(V_i)$
  - causal ($E^+(V_i)$) and diagnostic ($E^-(V_i)$) observations,
  - polytree $\longrightarrow$ it holds $E^+(V_i) \perp\!\!\!\perp E^-(V_i)|V_i$,
  - the only path connecting a causal and a diagnostic node leads through V_i.
- this separation can be used when computing probs $Pr^*(V_i)$

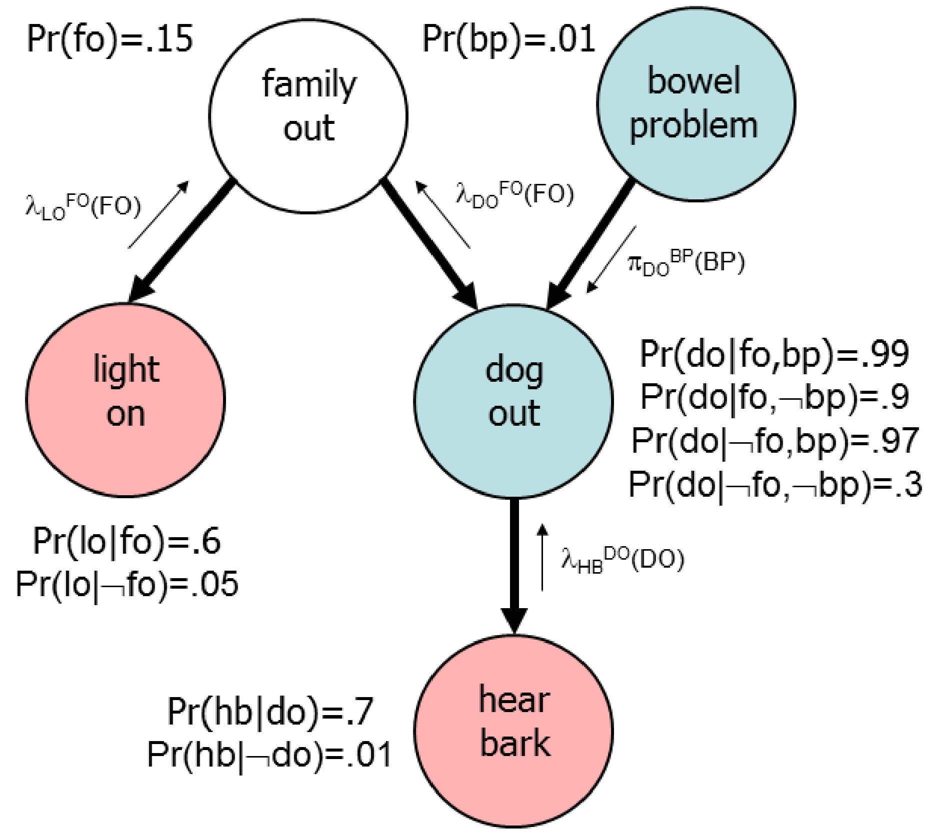
$$\begin{aligned} Pr^*(V_i) = Pr(V_i|E) &= Pr(V_i|E^+(V_i), E^-(V_i)) = \\ &= \alpha' \times Pr(E^+(V_i), E^-(V_i)|V_i) \times Pr(V_i) = \\ &= \alpha' \times Pr(E^-(V_i)|V_i) \times Pr(E^+(V_i)|V_i) \times Pr(V_i) = \\ &= \alpha \times Pr(E^-(V_i)|V_i) \times Pr(V_i|E^+(V_i)) = \\ &= \alpha \times \lambda(V_i) \times \pi(V_i) = bel(V_i) \end{aligned}$$

- compound causal $\pi(V_i)$ and diagnostic $\lambda(V_i)$ parameter

$$\pi(V_i) = \sum_{V_{p1},\dots,V_{pn}} Pr(V_i|V_{p1},\dots,V_{pn}) \prod_{j=1}^{n} \pi_{V_i}^{V_{pj}}(V_{pj}) \qquad \lambda(V_i) = \prod_{j=1}^{m} \lambda_{V_{cj}}^{V_i}(V_i)$$

Belief propagation – combined propagation

- Let us search for $Pr^*(FO)$ again: $Pr^*(fo) = Pr(fo|lo, \neg hb)$ a $Pr^*(\neg fo) = Pr(\neg fo|lo, \neg hb)$
  – $Pr^*(FO) = \alpha \times \lambda(FO) \times \pi(FO) = \alpha \times \lambda_{LO}^{FO}(FO) \times \lambda_{DO}^{FO}(FO) \times Pr(FO)$,
- λ messages from evidence nodes:
  – simple, follows from earlier examples,
  – light on – $Pr^*(lo) = 1$
    * $\lambda_{LO}^{FO}(fo) = Pr(lo|fo) = 0.6$,
    * $\lambda_{LO}^{FO}(\neg fo) = Pr(lo|\neg fo) = 0.05$,
  – no barking heard – $Pr^*(hb) = 0$,
    * $\lambda_{HB}^{DO}(do) = Pr(\neg hb|do) = 0.3$,
    * $\lambda_{HB}^{DO}(\neg do) = Pr(\neg hb|\neg do) = 0.99$,
- π message from BP node carries the priors:
  – $\pi_{DO}^{BP}(bp) = 0.01$, $\pi_{DO}^{BP}(\neg bp) = 0.99$,
- it is more difficult to quantify $\lambda_{DO}^{FO}(FO)$.
  – it equals $Pr^*(DO|FO)$.



Belief propagation – combined propagation

- In general, V_i sends messages as follows
  - $\pi^{V_i}_{V_{cj}}(V_i) = \alpha\pi(V_i)\prod^{k\neq j}_{k=1\ldots m}\lambda^{V_{ck}}_{V_i}(V_i)$,
  - $\lambda^{V_{pj}}_{V_i}(V_{pj}) = \sum_{v_i}\lambda(V_i)\sum_{V_{p1},\ldots,V_{pn}} Pr(V_i|V_{p1},\ldots,V_{pn})\prod^{k\neq j}_{k=1\ldots n}\pi^{V_k}_{V_i}(V_k)$,
- DO node passes to FO node

$$\begin{aligned}\lambda^{FO}_{DO}(fo) &= \lambda^{DO}_{HB}(do)\big[Pr(do|fo,bp)\pi^{BP}_{DO}(bp) + Pr(do|fo,\neg bp)\pi^{BP}_{DO}(\neg bp)\big]\\ &+ \lambda^{DO}_{HB}(\neg do)\big[Pr(\neg do|fo,bp)\pi^{BP}_{DO}(bp) + Pr(\neg do|fo,\neg bp)\pi^{BP}_{DO}(\neg bp)\big] =\\ &= .3(.99\times .01 + .9\times .99) + .99(.01\times .01 + .1\times .99) = .27 + .098 = 0.368\end{aligned}$$

$$\begin{aligned}\lambda^{FO}_{DO}(\neg fo) &= \lambda^{DO}_{HB}(do)\big[Pr(do|\neg fo,bp)\pi^{BP}_{DO}(bp) + Pr(do|\neg fo,\neg bp)\pi^{BP}_{DO}(\neg bp)\big]\\ &+ \lambda^{DO}_{HB}(\neg do)\big[Pr(\neg do|\neg fo,bp)\pi^{BP}_{DO}(bp) + Pr(\neg do|\neg fo,\neg bp)\pi^{BP}_{DO}(\neg bp)\big] =\\ &= .3(.97\times .01 + .3\times .99) + .99(.03\times .7 + .1\times .99) \cong .092 + .686 = 0.778\end{aligned}$$

- next, $Pr^*(FO)$ can be computed

$$\begin{aligned}Pr^*(fo) &= \alpha\times\lambda^{FO}_{LO}(fo)\times\lambda^{FO}_{DO}(fo)\times Pr(fo) =\\ &= \alpha\times .6\times .368\times .15 = .033\alpha\end{aligned}$$

$$\begin{aligned}Pr^*(\neg fo) &= \alpha\times\lambda^{FO}_{LO}(\neg fo)\times\lambda^{FO}_{DO}(\neg fo)\times Pr(\neg fo) =\\ &= \alpha\times .05\times .778\times .85 = .033\alpha\end{aligned}$$

- $Pr^*(fo) + Pr^*(\neg fo) = 1 \rightarrow \alpha = \frac{1}{.066} = 15.15 \rightarrow Pr^*(fo) = Pr^*(\neg fo) = 0.5$

Belief propagation – summary

- Initialization step
  - each observed node sends its causal and diagnostic parameters
    - π is either 0 or 1, λ carries conditional node probs (both according to observations),
  - each unobserved root passes its causal π equal to its prior prob distribution,
  - each unobserved leaf passes its diagnostic $\lambda = 1$.
- Iteration steps
  - carried out until any change occurs,
  - each node V_i which:
    - received the causal $\pi_{V_i}^{V_{pj}}(V_{pj})$ of all parents
      - $\Rightarrow$ computes its compound $\pi(V_i)$,
    - received the diagnostic $\lambda_{V_{cj}}^{V_i}(V_i)$ of all its children
      - $\Rightarrow$ computes its compound $\lambda(V_i)$,
    - knows its compound $\pi(V_i)$ and received diagnostic $\lambda_{V_{cj}}^{V_i}$ of all its children excepted V_c
      - $\Rightarrow$ passes $\pi_{V_c}^{V_i}(V_i)$ to V_c child,
    - knows its compound $\lambda(V_i)$ and received causal $\pi_{V_i}^{V_{pj}}(V_{pj})$ of all its parents excepted V_p
      - $\Rightarrow$ passes its $\lambda_{V_i}^{V_p}(V_p)$ to V_p parent.

Other inference algorithms

- **Clique or junction tree** – Lauritzen & Spigelhalter
  - apparently the most commonly used method for exact inference in general DAGs,
  - complexity is exponential in the size of the largest clique in transformed undirected graph,
  - applicable namely in sparse networks,
- **arc reversal** – R. D. Shachter
  - another exact inference method for general DAGs,
- stochastic sampling
  - approximate inference method for general DAGs,
  - instead of exact distribution $Pr(Q|\mathrm{e})$ makes its estimation by stochastic simulation,
  - although it does not have lower than NP complexity in general, time may be obtained at the expense of accuracy,
  - particular algorithms
    - **rejection sampling** – Henrion,
    - **likelihood weighting** – Fung & Chang,
    - **Gibbs sampling** – Geman & Geman, Pearl.

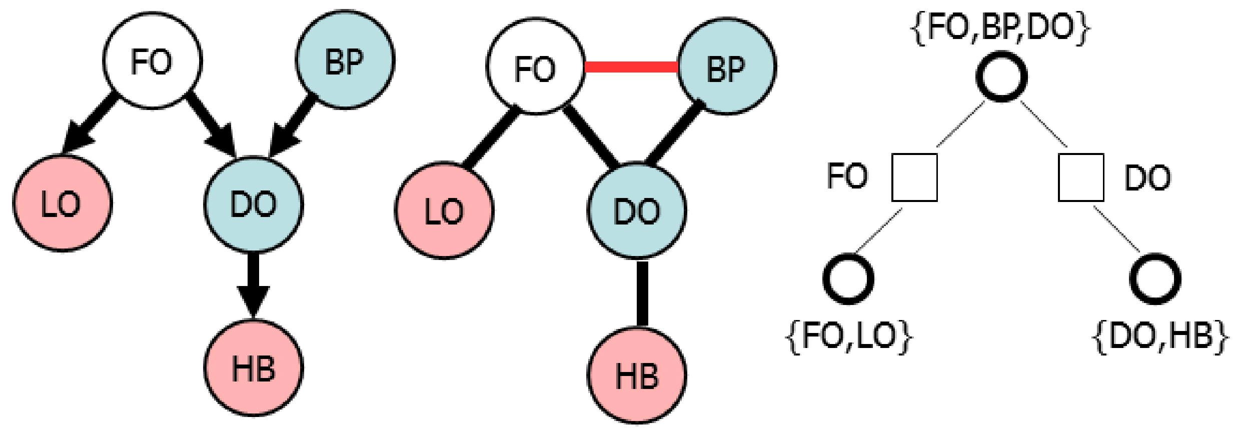
Junction tree algorithm

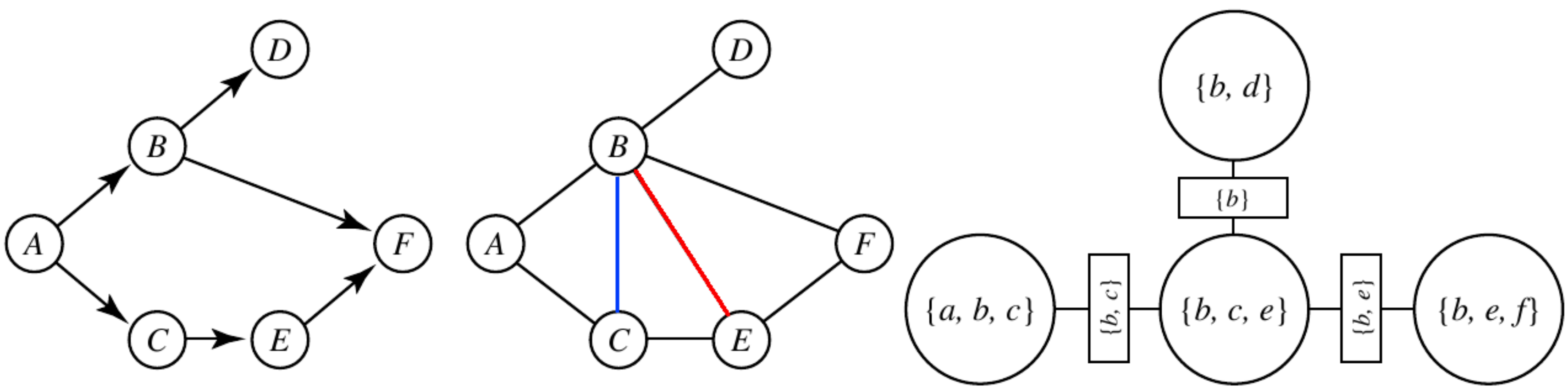
1. Moralization
   - connect nodes that have a common child with an undirected edge,
   - make all edges in the graph undirected,
2. triangulation
   - extend the existing graph to be triangulated,
   - each of its cycles of four or more nodes has a chord
     (an edge joining two nodes that are not adjacent in the cycle),
3. triangulated = decomposable (chordal) = a junction tree exists,
4. junction tree construction
   - clique nodes = cliques of triangulated graph,
   - cliques C_i in graph G can be ordered such that "running intersection property" holds

$$\forall i = 2 \ldots K \quad \exists 1 \leq j < i \quad C_i \cap \left(\bigcup_{k=1}^{i-1} C_k \right) \subseteq C_j$$

   - the connected graph has one edge less than the number of its nodes = tree,
   - tree is completed by separator nodes = intersections of adjacent cliques.

Junction tree algorithm – examples





Calculations in junction trees – examples

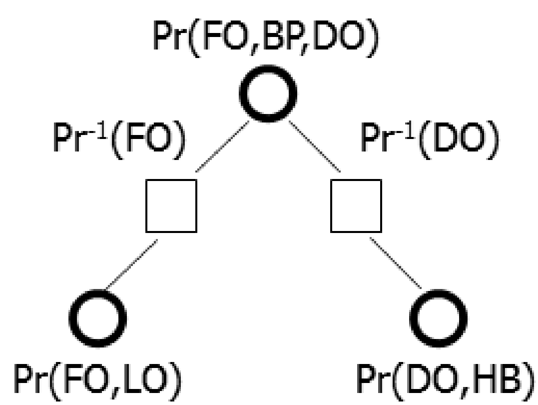
- stems from joint probability factorization along the triangulated graph

$$Pr^G = Pr^{C_1} \frac{Pr^{C_2}}{Pr^{C_2 \cap C_1}} \cdots \frac{Pr^{C_K}}{Pr^{C_K \cap (C_1 \cup C_2 \cup \ldots C_{K-1})}}$$

- product of all the junction tree nodes is at any moment and constantly equal to Pr^G,
- FAMILY example

$$Pr(FO, LO, BP, DO, HB) = \frac{Pr(FO, BP, DO) \times Pr(FO, LO) \times Pr(DO, HB)}{Pr(FO) \times Pr(DO)}$$

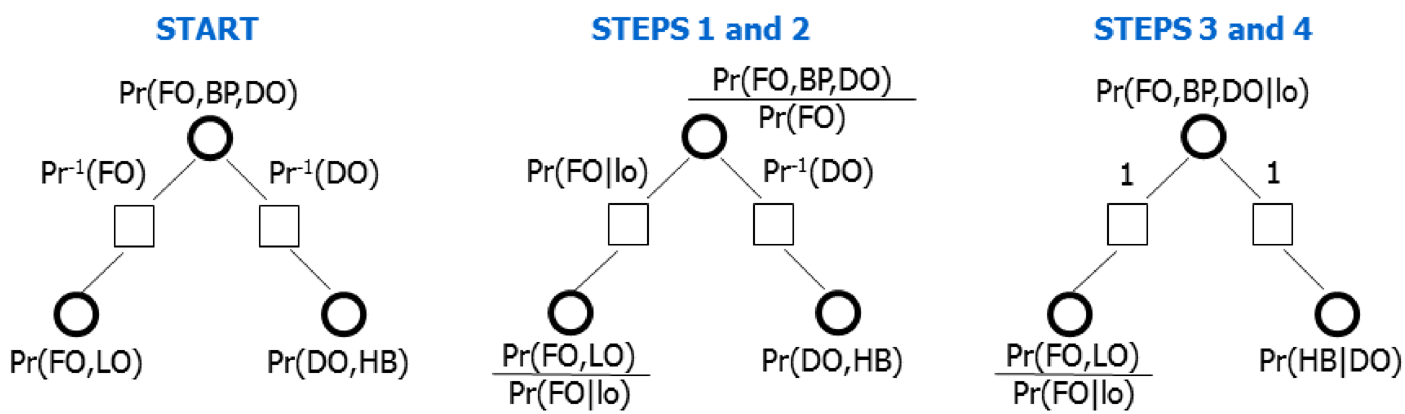
  – the node annotation probability tables are computed from the original network:
    * $Pr(FO, BP, DO) = Pr(FO) \times Pr(BP) \times Pr(DO|FO, BP)$,
    * $Pr(DO) = \sum_{FO,BP} Pr(FO, BP, DO)$,
    * $Pr(DO, HB) = Pr(DO) \times Pr(HB|DO)$, ...



Calculations in junction trees

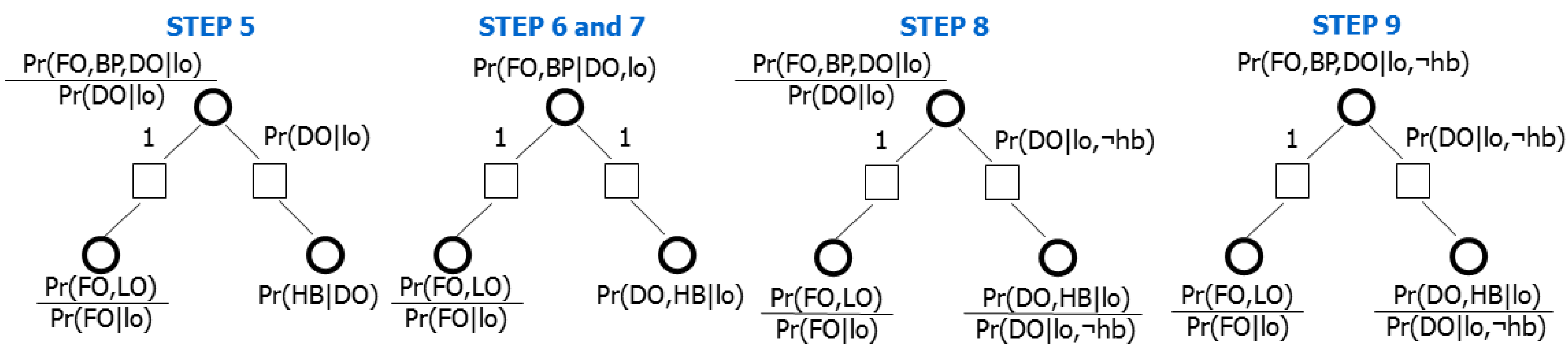
- the JT algorithm uses belief propagation to pass messages through the graph,
- enumerate $Pr(fo|lo,\neg hb)$:

1. {FO} node annotation is moved into {FO,BP,DO} node,
2. $Pr^*(lo) = 1 \rightarrow$ compute $Pr(FO|lo)$ from $Pr(FO,LO)$ and propagate it into {FO},
3. multiply probs in {FO,BP,DO} and {FO}, utilize $BP, DO \perp\!\!\!\perp LO|FO$ relationship $Pr(BP,DO|FO) \times Pr(FO|lo) = Pr(FO,BP,DO|lo)$,
4. multiply probs in {DO} and {DO,HB} nodes: $\frac{Pr(DO,HB)}{Pr(DO)} = Pr(HB|DO)$,



Calculations in junction trees

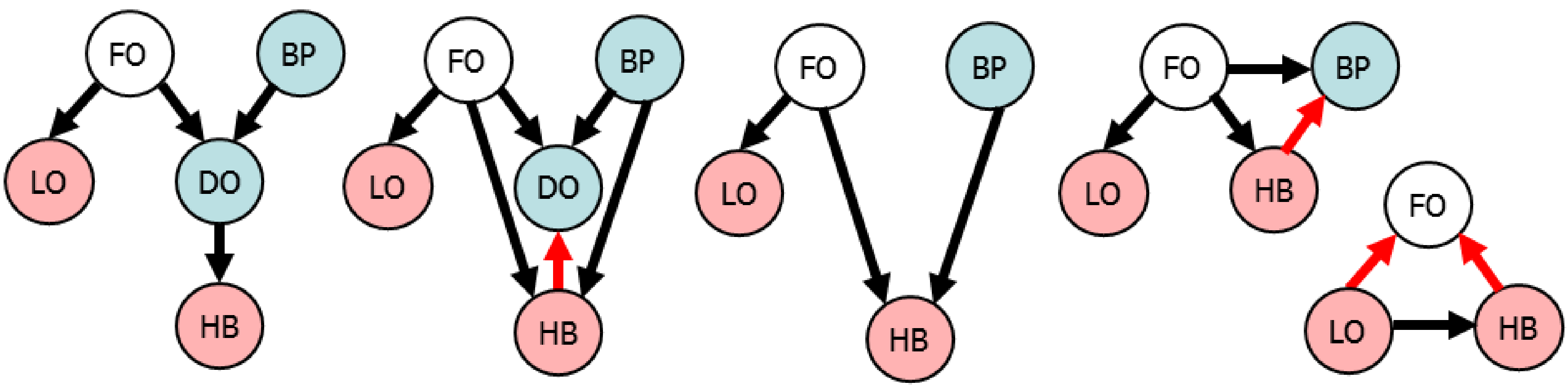
5. knowing $Pr(FO, BP, DO|lo)$, $Pr(DO|lo)$ is computed and passed into {DO} node,
6. {FO,BP,DO} annotation is updated: $\frac{Pr(FO,BP,DO|lo)}{Pr(DO|lo)} = Pr(FO, BP|DO, lo)$,
7. multiply probs in {DO} and {DO,HB} nodes, make use of $LO \perp\!\!\!\perp HB|DO$ property $Pr(DO|lo) \times Pr(HB|DO) = Pr(DO, HB|lo)$,
8. $Pr^*(lo) = 1 \rightarrow$ from $Pr(DO, HB|lo)$ compute $Pr(DO|lo, \neg hb)$ and pass it into {DO},
9. multiply probs in {FO,BP,DO} and {DO} nodes, make use of $FO, BP \perp\!\!\!\perp HB|DO$ property $Pr(FO, BP|DO, lo) \times Pr(DO|lo, \neg hb) = Pr(FO, BP, DO|lo, \neg hb)$,
10. through marginalization of {FO,BP,DO} node we obtain $Pr(FO|lo, \neg hb)$.



Arc reversal algorithm

- transform the original Bayesian network into a different one,
- the represented joint distribution either does not change or it is marginal wrt original one,
- the transformed network must include the query and evidence nodes (Q and E),
- the target marginal distribution $Pr(Q|\mathbf{E})$ is available directly in the final transformed network,
- algorithm has 2 steps

1. node elimination – a node makes a tail (initial vertex) of no edge (its output degree is 0),
2. arc reversal – if there is an edge from a parent to its child and there is no alternative directed path between them, a transformation that does not change the joint distribution can be made – the arc is reversed, its incident nodes mutually inherit their parents.



Arc reversal algorithm

- each arc reversal from $P_k \rightarrow P_l$ to $P_l \rightarrow P_k$ is accompanied by CPT recomputations,
- let us start with CPT of the new parent node
  (the old$^-$ and new$^+$ graph need to be used concurrently):

$$Pr(P_l|parents^+(P_l)) = \sum_{\forall p \in P_k} Pr(P_k = p|parents^-(P_k)) \times Pr(P_l|parents^-(P_l) \setminus P_k, P_k = p)$$

  – the paths leading through the former parent P_k replaced by an edge,
  – the new edge sums the information flows for all the possible P_k values,
- next, let us derive CPT for the new child:

$$Pr(P_k|parents^+(P_k)) = \frac{Pr(P_k|parents^-(P_k)) \times Pr(P_l|parents^-(P_l))}{Pr(P_l|parents^+(P_l))}$$

  – in the trivial case $parents^-(P_k) = parents^+(P_l) = \emptyset$ the recomputation formula reduces on Bayes theorem.

Arc reversal algorithm – example

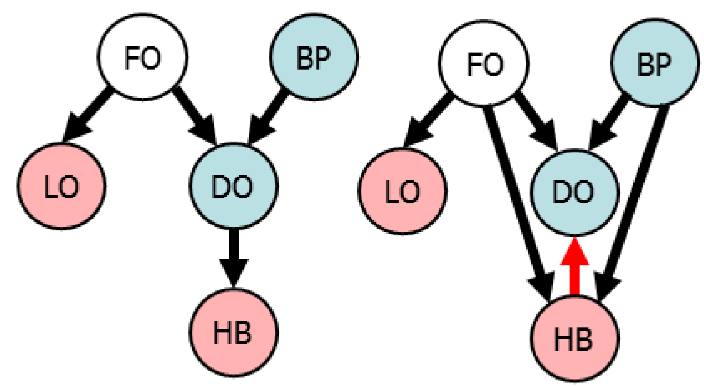
- let us consider a particular arc reversal from $DO \rightarrow HB$ to $HB \rightarrow DO$, it holds:

$$Pr(HB|FO,BP) = \sum_{p \in \{do, \neg do\}} Pr(DO=p|FO,BP) \times Pr(HB|DO=p)$$

$$Pr(DO|FO,DO,HB) = \frac{Pr(DO|FO,BP) \times Pr(HB|DO)}{Pr(HB|FO,BP)}$$

FO	BP	HB	Pr(HB|FO,BP)
T	T	T	$.99 \times .7 + .01 \times .01 = .6931$
T	F	T	$.9 \times .7 + .1 \times .01 = .631$
F	T	T	$.97 \times .7 + .03 \times .01 = .6793$
F	F	T	$.3 \times .7 + .7 \times .01 = .217$

FO	BP	HB	DO	Pr(DO|FO,BP,HB)
T	T	T	T	$.99 \times .7/.6931 = .9999$
T	F	T	T	$.9 \times .7/.631 = .9984$
				$\ldots$
F	T	F	T	$.97 \times .3/.3207 = .9074$
F	F	F	T	$.3 \times .3/.783 = .1149$



Pr(do|fo,bp)=.99
Pr(do|fo,¬bp)=.9
Pr(do|¬fo,bp)=.97
Pr(do|¬fo,¬bp)=.3

Pr(hb|do)=.7
Pr(hb|¬do)=.01

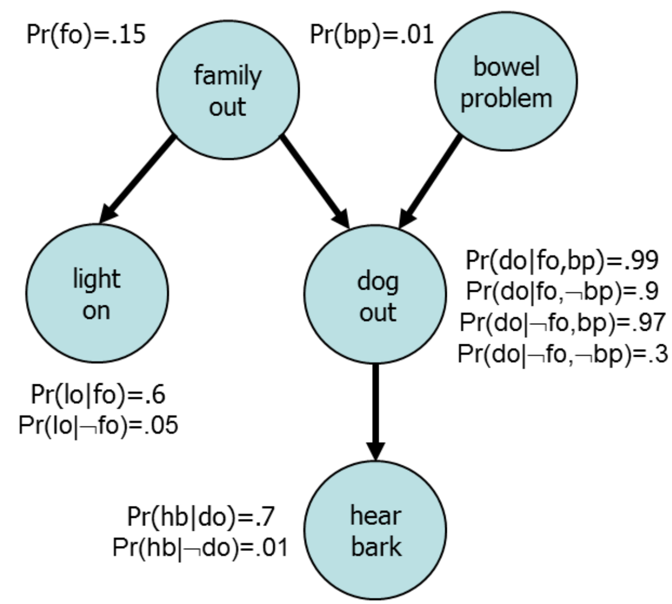
Approximate inference by stochastic sampling

- a general **Monte-Carlo** method, samples from the joint prob distribution,
- estimates the target conditional probability (query) from a sample set,
- the joint prob distribution is not explicitly given, its factorization is available only (network),
- the most straightforward is direct **forward sampling**
  1. topologically sort the network nodes
     - for every edge it holds that parent comes before its children in the ordering,
  2. instantiate variables along the topological ordering
     - take $Pr(P_j|parents(P_j))$, randomly sample P_j,
  3. repeat step 2 for all the samples (the sample size M is given a priori),
- from samples to probabilities?
  - $Pr(q|\mathbf{e}) \approx \frac{N(q,\mathbf{e})}{N(\mathbf{e})}$
  - samples that contradict evidence not used,
  - forward sampling gets inefficient if $Pr(\mathbf{e})$ is small,
- **rejection sampling** brings a slight improvement
  - rejects partially generated samples as soon as they violate the evidence event $\mathbf{e}$,
  - sample generation often stops early.

Rejection sampling – example

- FAMILY example, estimate $Pr(fo|lo, \neg hb)$
  1. topologically sort the network nodes
     – e.g., $\langle FO, BP, LO, DO, HB \rangle$ (or $\langle BP, FO, DO, HB, LO \rangle$, etc.)
  2. instantiate variables along the topological ordering
     – $Pr(FO) \rightarrow \neg fo$, $Pr(BP) \rightarrow \neg bp$,
       $Pr(LO|\neg fo) \rightarrow lo$, $Pr(DO|\neg fo, \neg bp) \rightarrow \neg do$, $Pr(HB|\neg do) \rightarrow \neg hb$
     – sample agrees with the evidence $\mathbf{e} = lo \wedge \neg hb$, no rejection needed,
  3. generate 1000 samples, repeat step 2,

- let $N(fo, lo, \neg hb)$ is 491 (the number of samples with the given values of three variables under consideration),
- in rejection sampling $N(\mathbf{e})$ necessarily equals M,
  – $Pr(fo|lo, \neg hb) \approx \frac{N(q,\mathbf{e})}{N(\mathbf{e})} = \frac{491}{1000} = 0.491$



Likelihood weighting

- **Likelihood weighting** is a sampling method that avoids necessity to reject samples
  - the values of $\mathbf{E}$ are fixed, the rest of variables is sampled only,
  - however, not all events are equally probable, samples need to be weighted,
  - the weight equals the likelihood of the event given the evidence,
- $\forall$ samples $p^m = \{P_1 = p_1^m, \ldots, P_n = p_n^m\}$, $m \in \{1, \ldots, M\}$
  1. $w^m \leftarrow 1$ (initialize the sample weight)
  2. $\forall j \in \{1, \ldots, n\}$ (instantiate variables along the topological ordering)
     - if $P_j \in \mathbf{E}$ then take p_j^m from $\mathbf{e}$ and $w^m \leftarrow w^m \times Pr(P_j|parents(P_j))$,
     - otherwise randomly sample p_j^m from $Pr(P_j|parents(P_j))$,
- from samples to probabilities?
  - evidence holds in all samples (by definition),
  - weighted averaging is applied to find $Pr(Q = P_i|\mathbf{e})$

$$Pr(p_i|\mathbf{e}) \approx \frac{\sum_{m=1}^{M} w^m \delta(p_i^m, p_i)}{\sum_{m=1}^{M} w^m} \quad \delta(i,j) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

- nevertheless, samples may have very low weights
  - it can also turn out inefficient in large networks with evidences occuring late in the ordering.

Likelihood weighting – example

- let us approximate $Pr(fo|lo, \neg hb)$ (its exact value computed earlier is 0.5),

	p^1	p^2	p^3	...
FO	**F**	**F**	T	
BP	**F**	**F**	F	
LO	**T**	**T**	**T**	
DO	**F**	**T**	T	
HB	**F**	**F**	**F**	
w	.0495	.015	.18	

FO^1: $Pr(fo) = .15 \rightarrow \neg fo$ sampled
BP^1: $Pr(bp) = .01 \rightarrow \neg bp$ sampled
LO^1: evidence $\rightarrow lo \wedge w^1 = Pr(lo|\neg fo) = .05$
DO^1: $Pr(do|\neg fo, \neg bp) = .3 \rightarrow \neg do$ sampled
HB^1: evidence $\rightarrow \neg hb \wedge w^1 = .05 \times Pr(\neg hb|\neg do) = .0495$

FO^2: $Pr(fo) = .15 \rightarrow \neg fo$ sampled
BP^2: $Pr(bp) = .01 \rightarrow \neg bp$ sampled
LO^2: evidence $\rightarrow lo \wedge w^1 = Pr(lo|\neg fo) = .05$
DO^2: $Pr(do|\neg fo, \neg bp) = .3 \rightarrow do$ sampled
HB^2: evidence $\rightarrow \neg hb \wedge w^2 = .05 \times Pr(\neg hb|do) = .015$

- a very rough estimate having 3 samples only

$$Pr(fo|lo, \neg hb) \approx \frac{.18}{.0495 + .015 + .18} = .74$$

Gibbs sampling

- a Markov chain method – the next state depends purely on the current state
  - generates dependent samples!
  - as it is a Monte-Carlo method as well $\longrightarrow$ MCMC,
- efficient sampling method namely when some of BN variable states are known
  - it again samples nonevidence variables only, the samples never rejected,
- sampling process – samples $p^m = \{P_1 = p_1^m, \ldots, P_n = p_n^m\}$, $m \in \{1, \ldots, M\}$
  1. fix states of all observed variables from $\mathbf{E}$ (in all samples),
  2. the other variables initialized in p^0 randomly,
  3. generate p^m from p^{m-1} $(\forall P_i \notin E)$
     - $p_1^m \leftarrow Pr(P_1|p_2^{m-1}, \ldots, p_n^{m-1})$,
     - $p_2^m \leftarrow Pr(P_2|p_1^m, p_3^{m-1}, \ldots, p_n^{m-1})$,
     - $\ldots$,
     - $p_n^m \leftarrow Pr(P_n|p_1^m, \ldots, p_{n-1}^m)$,
  4. repeat step 3 for $m \in \{1, \ldots, M\}$.

Gibbs sampling

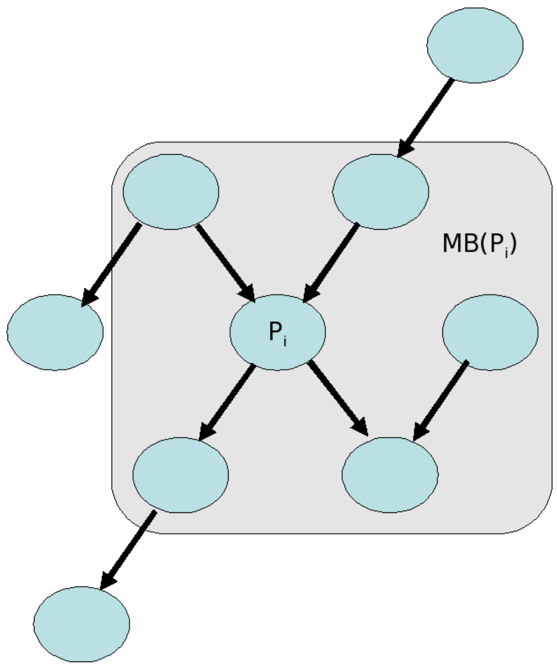
- probs $Pr(P_i|P_1,\dots,P_{i-1},P_{i+1},\dots,P_n) = Pr(P_i|P\setminus P_i)$ not explicitly given ...
  - to enumerate them, only their BN neighborhood needs to be known

$$Pr(P_i|P\setminus P_i) \propto Pr(P_i|parents(P_i)) \prod_{\forall P_j, P_i \in parents(P_j)} Pr(P_j|parents(P_j))$$

  - the neighborhood is called **Markov blanket** (MB),
  - MB covers the node, its parents, its children and their parents,
  - $MB(P_i)$ is the minimum set of nodes that d-separates P_i from the rest of the network.

- from samples to probabilities?
  - evidence holds in all samples (by definition),
  - averaging $\forall m$ is applied to find $Pr(Q|\mathbf{e})$

$$Pr(p_i|\mathbf{e}) \approx \frac{\sum_{m=1}^{M} \delta(p_i^m, p_i)}{M} \quad \delta(i,j) = \begin{cases} 1 & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases}$$



Gibbs sampling – example

- let us approximate $Pr(fo|lo, \neg hb)$ (its exact value computed earlier is 0.5),

	p^0	p^1	p^2	$\dots$
FO	T	**F**	**F**	
BP	T	**F**	**F**	
LO	**T**	**T**	**T**	
DO	F	F	F	
HB	**F**	**F**	**F**	

p^0: random init of unevidenced variables

FO^1: $Pr^*(fo) \propto Pr(fo) \times Pr(lo|fo) \times Pr(\neg do|fo, bp)$

$Pr^*(\neg fo) \propto Pr(\neg fo) \times Pr(lo|\neg fo) \times Pr(\neg do|\neg fo, bp)$

$Pr^*(fo) \propto .15 \times .6 \times .01 = 9 \times 10^{-4} \rightarrow \times\alpha^1_{FO} = .41$

$Pr^*(\neg fo) \propto .85 \times .05 \times .03 = 1.275 \times 10^{-3} \rightarrow \times\alpha^1_{FO} = .59$

$\alpha^1_{FO} = \frac{1}{Pr^*(fo)+Pr^*(\neg fo)} = 460$

BP^1: $Pr^*(bp) \propto Pr(bp) \times Pr(\neg do|\neg fo, bp) = .01 \times .03 = .0003$

$Pr^*(\neg bp) \propto Pr(\neg bp) \times Pr(\neg do|\neg fo, \neg bp) = .99 \times .7 = 0.693$

$\alpha^1_{BP} = \frac{1}{Pr^*(bp)+Pr^*(\neg bp)} = 1.44 \rightarrow Pr^*(bp) = 4 \times 10^{-4}$

DO^1: by analogy, $|MB(DO)| = 5$

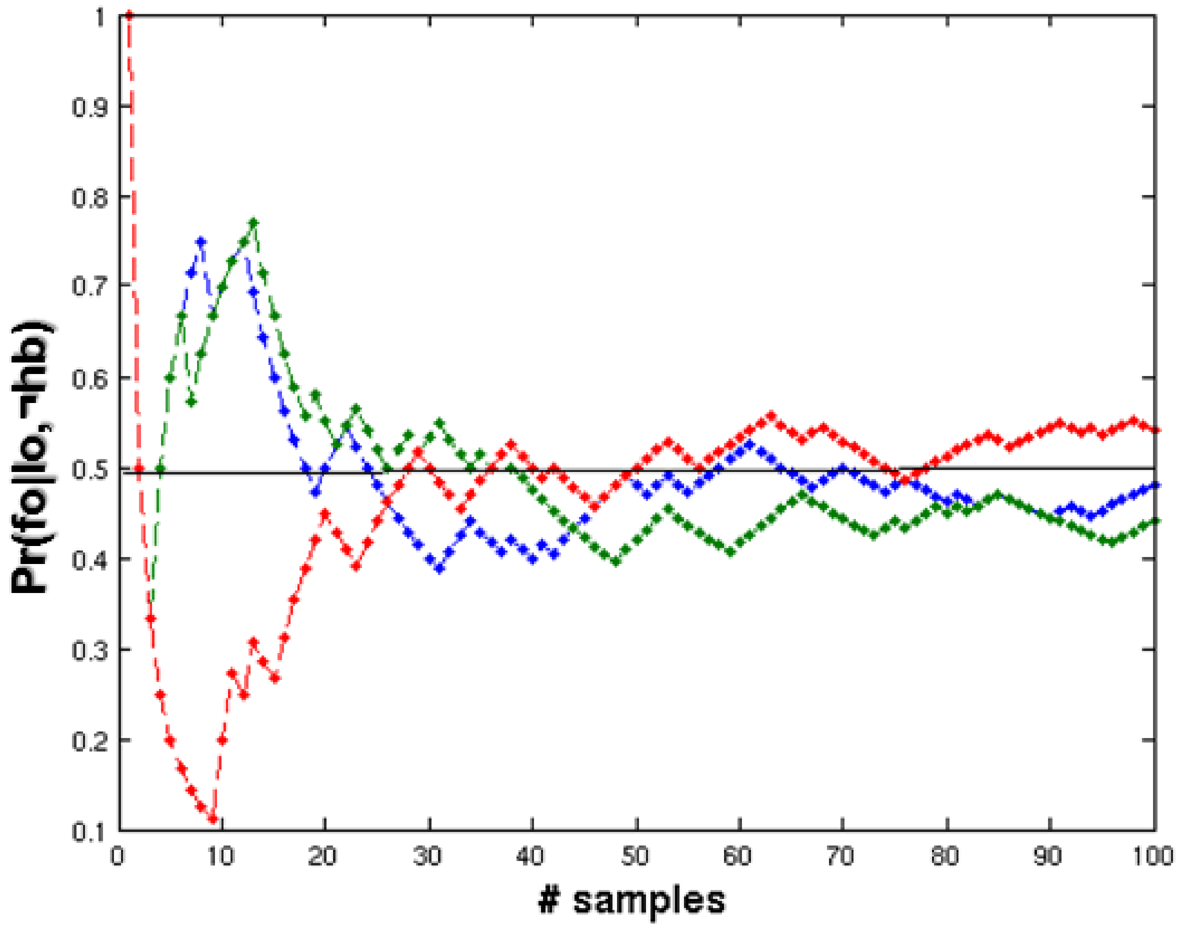
FO^2: BP value was switched, substitution is $Pr(DO|FO, \neg bp)$

$Pr^*(fo) = .21 \quad Pr^*(\neg fo) = .79$

BP^2: the same probs as is sample 1

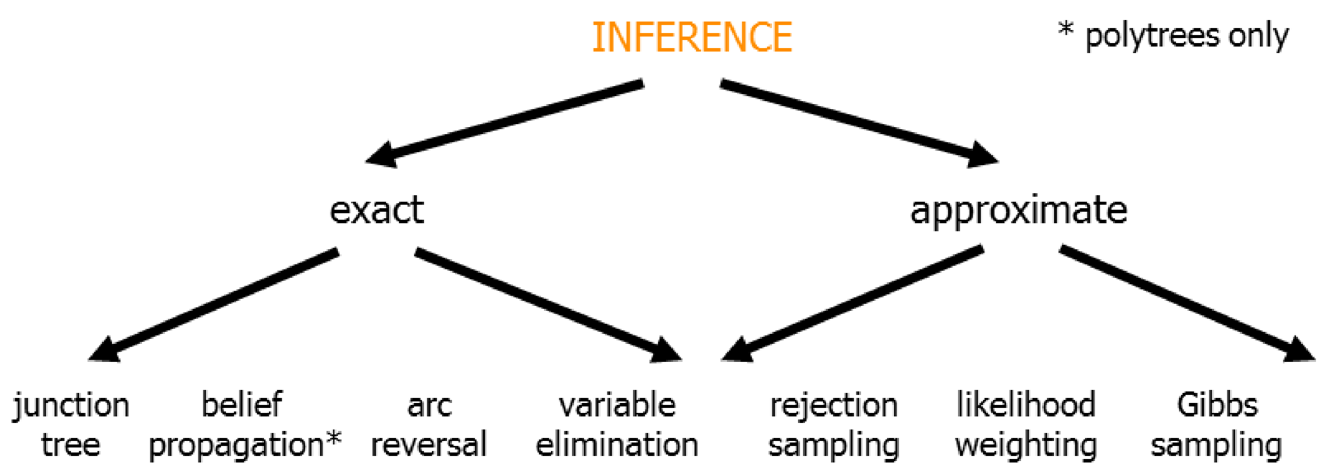
Gibbs sampling – example

- BN Matlab Toolbox, aproximation of $Pr(fo|lo, \neg hb)$,
- gibbs_sampling_inf_engine, three independent runs with 100 samples.



Summary

- independence and conditional independence ramarkably simplify prob model
  - still, BN inference remains generally **NP-hard** wrt the number of network variables,
  - inference complexity grows with the number of network edges
    - naïve Bayes model – linear complexity,
    - general complexity estimate from the size of maximal clique of triangulated graph,
  - inference complexity can be reduced by constraining model structure
    - special network types (singly connected), e.g. trees – one parent only,
  - inference time can be shorten when exact answer is not required
    - approximate inference, typically (but not only) stochastic sampling.



Recommended reading, lecture resources

- Russell, Norvig: **AI: A Modern Approach**, Uncertain Knowledge and Reasoning (Part V)
  - probabilistic reasoning (chapter 14 or 15, depends on the edition),
  - online on Google books: http://books.google.com/books?id=8jZBksh-bUMC,
- Jiroušek: **Metody reprezentace a zpracování znalostí v umělé inteligenci.**
  - bayesovské sítě (kapitola 6), metoda postupných modifikací sítě,
  - http://staff.utia.cas.cz/vomlel/r.pdf,
- Šingliar: **Pearl's algorithm.**
  - a message passing algorithm for exact inference in polytree BBNs,
  - http://www.cs.pitt.edu/ tomas/cs3750/pearl.ppt.