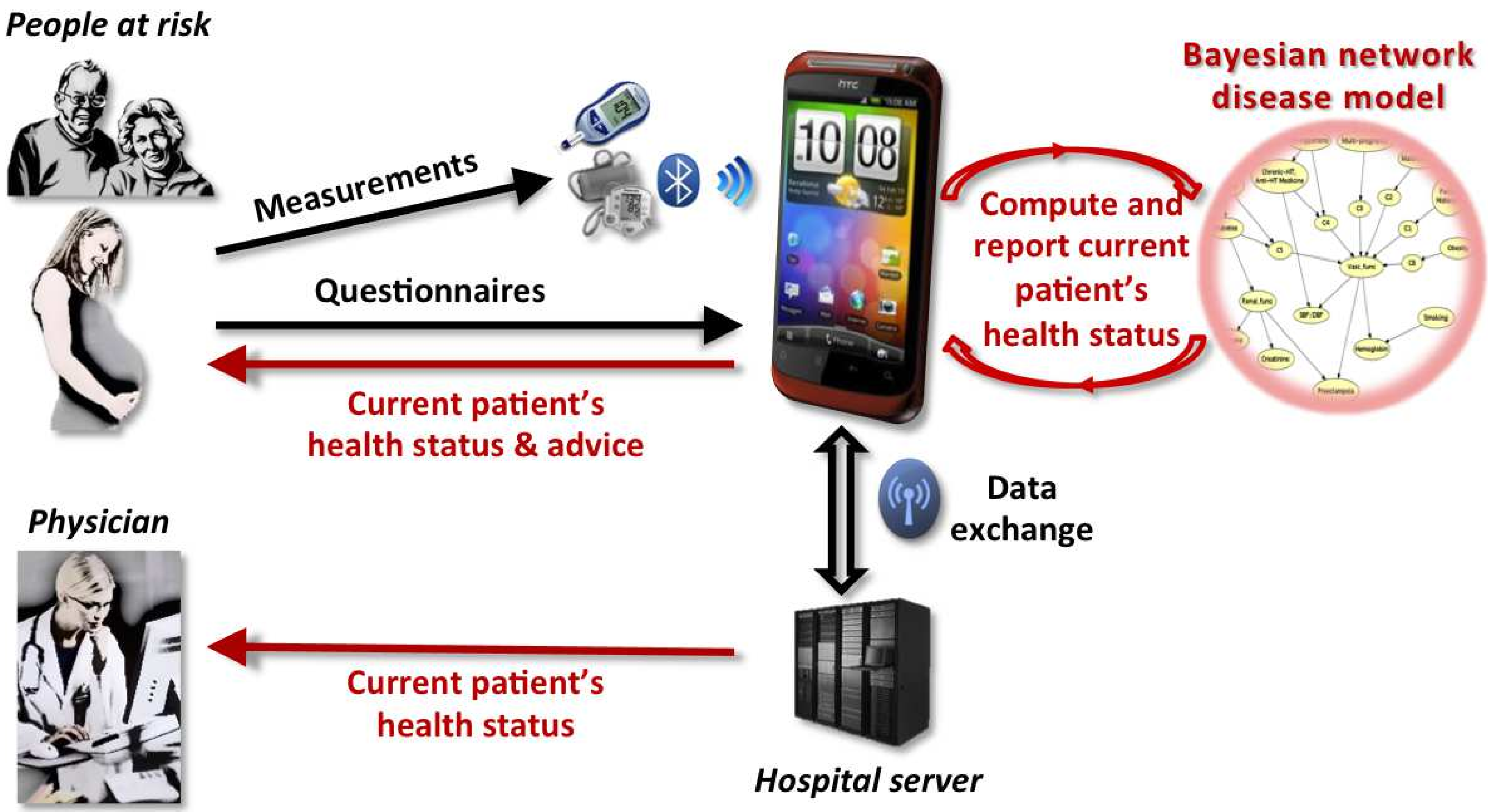
Bayesian networks

Principles and Definitions

The focus today . . .

- Probability theory
  - Joint probability
  - Marginal probability
  - Conditional probability
  - Chain rule
  - Bayes’ rule
- Bayesian networks
  - Definition
  - Conditional independence

Intelligent patient monitoring at home



Why Bayesian networks?

Probabilistic graphical models, such as Bayesian networks, are now the most popular uncertainty formalisms because:

- Handle noise, missing information and probabilistic relations
- Learn from data and can incorporate domain knowledge
- Offer flexible reasoning
- Have compact graphical representation (interface)

- Founded principles: probability theory

- Engineering principles: knowledge acquisition, machine learning and statistics

General notation

- Stochastic (= statistical = random) variable: upper-case letter, e.g. X, or upper-case string, e.g. FEVER
- Values: variables can take on values, e.g. $X = x$, FEVER = *yes*
- Binary variables: take one of *two* values, e.g. X = *true* and X = *false*
- Discrete variables: take only one of a finite set of possible values, e.g. $\mathrm{TEMP} \in \{$*low, medium, high*$\}$
- Continuous variables: take any value in some interval or intervals of real numbers $\mathcal{R}$, e.g. $\mathrm{TEMP} \in [-50, 50]$

Abbreviated notation

- *Binary* variables: X = *true* as x, and X = *false* as $\neg\mathrm{x}$
- *Non-binary* variables: $X = x$ as x or CITY = *tokyo* as *tokyo*
- *Sets* of variables: analogous to variables
  - Example:

$$
\begin{array}{c}
X_1 = x_1 \\
X_2 = x_2 \\
\cdot \\
\cdot \\
X_n = x_n
\end{array}
\quad \Longrightarrow \quad
\begin{array}{ccl}
X & = & (X_1, X_2, \ldots X_n) \\
x & = & (x_1, x_2, \ldots x_n) \\
X & = & x
\end{array}
$$

Abreviated notation (cont.)

- Conjunctions: $(X = x) \wedge (Y = y)$ as $(X = x, Y = y)$
- Templates: (X, Y) means $(X = x, Y = y)$, for *any* value x, y, i.e. the choice of the values x and y does not really matter
- Examples:
  - $P(X = x, Y = y) \Leftrightarrow P(X = x \wedge Y = y)$
  - $P(X, Y) \Leftrightarrow P(X = x, Y = y)$, for *any* value x, y
  - $P(X \mid Y) \Leftrightarrow P(X = x \mid Y = y)$, for *any* value x, y
- $\sum_X P(X) = P(\mathrm{x}) + P(\neg \mathrm{x})$, where X is binary

Probability theory

- Probability distribution P: attaches a number in (closed) interval $[0,1]$ to *Boolean expressions*
- Boolean algebra $\mathbb{B}$ (for two variables RAIN and HAPPY):

  $\top$ (*true*),
  rain, $\neg$*rain*,
  happy, $\neg$*happy*,
  rain $\wedge$ *happy*,$\ldots$, *rain* $\wedge$ *happy* $\wedge$ $\neg$*happy*$, \ldots$,
  $\neg$*rain* $\wedge$ *happy*,$\ldots$, *rain* $\vee$ *happy*,
  $\perp$ (*false*)

  such that:

  - $\perp \leq$ *rain*, *rain* $\leq$ (*rain* $\vee$ *happy*)$, \ldots$ (in general $\perp \leq x$ for each Boolean expression $x \in \mathbb{B}$);
  - $x \leq \top$ for each Boolean expression $x \in \mathbb{B}$

Probability distribution

- A probability distribution P is defined as a function $P : \mathbb{B} \to [0, 1]$, such that:
  - $P(\bot) = 0$
  - $P(\top) = 1$
  - $P(x \vee y) = P(x) + P(y)$, if $x \wedge y = \bot$ with $x, y \in \mathbb{B}$
- Examples:
  - $P(\textit{rain} \vee \textit{happy}) = P(\textit{rain}) + P(\textit{happy})$, as $\textit{rain} \wedge \textit{happy} = \bot$ (why? Because I define it that way)
  - $P(\textit{rain} \wedge \textit{happy}) = P(\bot) = 0$
  - $P(\neg\textit{rain} \vee \textit{rain}) = P(\neg\textit{rain}) + P(\textit{rain}) = P(\top) = 1 \Rightarrow P(\neg\textit{rain}) = 1 - P(\textit{rain})$
  - $0 \leq P(\textit{rain}) \leq 1$

Probability distribution (cont.)

- Boolean algebras $\Leftrightarrow$ sets:
  - $\top \Leftrightarrow \Omega$
  - $\bot \Leftrightarrow \varnothing$
  - $x \Leftrightarrow X$
  - $\neg x \Leftrightarrow \bar{X}$
  - $(x \vee y) \Leftrightarrow (X \cup Y)$
  - $(x \wedge y) \Leftrightarrow (X \cap Y)$
  - $x \leq (x \vee y) \Leftrightarrow X \subseteq (X \cup Y)$

with $\Leftrightarrow$ 1-1 correspondence, e.g.

$$P(\overline{\text{Rain}}) = 1 - P(\text{Rain})$$

Joint probability distribution

Let X and Y be random variables with domains

$dom(X) = \{x_1, x_2, \ldots, x_n\}$ and $dom(Y) = \{y_1, y_2, \ldots, y_m\}$.

The product set

$$dom(X) \times dom(Y) = \{x_1, x_2, \ldots, x_n\} \times \{y_1, y_2, \ldots, y_m\}$$

is made into a probability space by defining

$$P(X = x_i \wedge Y = y_j) = P(x_i, y_j)$$

where P is a joint probability function of X and Y.

Marginalisation

Suppose the joint probability distribution of two variables X and Y is given; then

$$\begin{aligned} P(x) = P(X = x) &= P(x \wedge \top) \\ &= P(x \wedge (y \vee \neg y)) \\ &= P((x \wedge y) \vee (x \wedge \neg y)) \\ &= P(x \wedge y) + P(x \wedge \neg y) \end{aligned}$$

since $P(a \vee b) = P(a) + P(b)$, if $a \wedge b = \bot$

$$\Longrightarrow P(x) = \sum_{Y} P(x, Y)$$

also known as marginal probability function of X.

Example

- Assume that X_1, X_2, X_3 and X_4 are binary variables. Then $P(X1, X_2, X_3, X_4)$:

$$
\begin{aligned}
P(x_1, x_2, x_3, x_4) &= 0.1 & P(x_1, \neg x_2, \neg x_3, x_4) &= 0.015 \\
P(x_1, \neg x_2, x_3, x_4) &= 0.04 & P(x_1, \neg x_2, x_3, \neg x_4) &= 0.1 \\
P(x_1, x_2, \neg x_3, x_4) &= 0.03 & P(x_1, x_2, \neg x_3, \neg x_4) &= 0.004 \\
P(x_1, x_2, x_3, \neg x_4) &= 0.1 & P(\neg x_1, \neg x_2, \neg x_3, x_4) &= 0.005 \\
P(\neg x_1, x_2, x_3, x_4) &= 0.0 & P(\neg x_1, \neg x_2, x_3, \neg x_4) &= 0.01 \\
P(\neg x_1, \neg x_2, x_3, x_4) &= 0.2 & P(\neg x_1, x_2, \neg x_3, \neg x_4) &= 0.01 \\
P(\neg x_1, x_2, \neg x_3, x_4) &= 0.08 & P(x_1, \neg x_2, \neg x_3, \neg x_4) &= 0.006 \\
P(\neg x_1, x_2, x_3, \neg x_4) &= 0.1 & P(\neg x_1, \neg x_2, \neg x_3, \neg x_4) &= 0.2
\end{aligned}
$$

- $\sum_{X1,X2,X3,X4} P(X_1, X_2, X_3, X_4) = 1$
- Marginalisation:

$$P(x_2, \neg x_3) =?$$

Conditional probability

- $P(X \mid Y)$: Chance that X will occur knowing that Y has occured
- Definition:

$$P(X \mid Y) \quad = \quad \frac{P(X \cap Y)}{P(Y)}$$

$\nearrow$

normalize, so

that uncertainty in Y

is removed

Example: *flu* and *fever*

- $P(\textit{flu} \wedge \textit{fever})$: chance of *flu* and *fever* at the same time
- $P(\textit{flu} \mid \textit{fever})$: chance of *flu* knowing that the person already has *fever* (conditional probability)
- Definition:

$$P(\textit{flu} \mid \textit{fever}) = \frac{P(\textit{flu} \wedge \textit{fever})}{P(\textit{fever})}$$

$\nearrow$

adjust $P(\textit{flu} \wedge \textit{fever})$, so
that uncertainty in ‘fever’
is removed

- Recall: $P(\text{Flu} \cap \text{Fever})$ is different notation, with same meaning as $P(\textit{flu} \wedge \textit{fever})$

Reversal of chances

- $P(\textit{flu} \mid \textit{fever})$ is usually unknown:

flu ← $P(\textit{flu} \mid \textit{fever})$ — fever

h (hypothesis) e (evidence)

- Known is:

$P(\textit{fever} \mid \textit{flu}) = 0.9$
$P(\textit{flu}) = 0.05$
$P(\textit{fever}) = 0.09$

flu — $P(\textit{fever} \mid \textit{flu})$ → fever

$P(\textit{flu}) = 0.05$ $P(\textit{fever}) = 0.09$

Bayes' rule - Example

- Bayes' rule – reversal of chances:

  $P(e \mid h)$ $\quad P(\textit{fever} \mid \textit{flu}) = 0.9$
  $P(h)$ $\quad P(\textit{flu}) = 0.05$
  $P(e)$ $\quad P(\textit{fever}) = 0.09$

$$P(\textit{flu} \mid \textit{fever}) = \frac{P(\textit{fever} \mid \textit{flu})P(\textit{flu})}{P(\textit{fever})}$$
$$= 0.9 \cdot 0.05/0.09 = 0.5$$

- Definition of Bayes' rule (the 'chance reverter'):

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)}$$

Chain rule (derivation)

Definition of conditional probability:

$$P(X_1 \mid X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_2, \ldots, X_n)}$$

$$\Rightarrow P(X_1, X_2, \ldots, X_n) \ = P(X_1 \mid X_2, \ldots, X_n) P(X_2, \ldots, X_n)$$

Furthermore,

$$\begin{aligned} P(X_2, \ldots, X_n) &= P(X_2 \mid X_3, \ldots, X_n) P(X_3, \ldots, X_n) \\ \vdots \quad &\vdots \quad \vdots \\ P(X_{n-1}, X_n) &= P(X_{n-1} \mid X_n) P(X_n) \\ P(X_n) &= P(X_n) \end{aligned}$$

Chain rule (definition)

$$\begin{aligned}
P(X_1, X_2, \ldots, X_n) &= P(X_1 \mid X_2, \ldots, X_n) \cdot \\
& \quad P(X_2 \mid X_3, \ldots, X_n) \cdot \\
& \quad P(X_3 \mid X_4, \ldots, X_n) \cdot \\
& \quad \vdots \\
& \quad P(X_{n-1} \mid X_n) \cdot \\
& \quad P(X_n) \\
&= \prod_{i=1}^{n-1} P(X_i \mid X_{i+1}, \ldots, X_n) P(X_n)
\end{aligned}$$

Definition Bayesian network (BN)

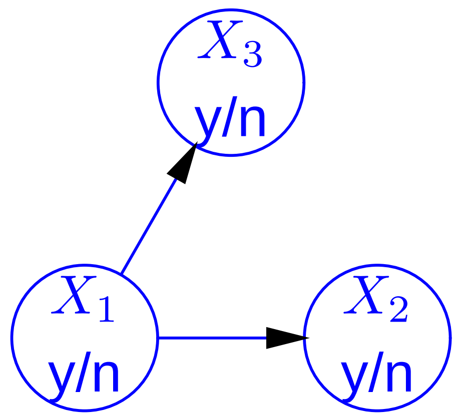
A Bayesian network $\mathcal{B}$ is a pair $\mathcal{B} = (G, P)$, where:

- $G = (V(G), A(G))$ is an acyclic directed graph, with
  - $V(G) = \{v_1, v_2, \ldots, v_n\}$, a set of vertices (nodes)
  - $A(G) \subseteq V(G) \times V(G)$ a set of arcs
- $P : \wp(V(G)) \to [0, 1]$ is a joint probability distribution, such that

$$P(V(G)) = \prod_{i=1}^{n} P(v_i \mid \pi_G(v_i))$$

  where $\pi_G(v_i)$ denotes the set of immediate ancestors (parents) of vertex v_i in G
- Notational convenience: $v_i \approx X_i$

Example of a Bayesian network



Bayesian network $\mathcal{B} = (G, P)$, where $G = (V(G), A(G))$, with

- Set of vertices: $V(G) = \{X_1, X_2, X_3\}$
- Set of arcs: $A(G) = \{(X_1, X_2), (X_1, X_3)\}$
- Joint probability distribution:

$$P(X_1, X_2, X_3) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1)$$

Example (cont.)

$$P(X_1, X_2, X_3) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1)$$

with for example:

$$\begin{aligned} P(x_1) &= 0.7 \\ P(\neg x_1) &= 0.3 = 1 - P(x_1) \end{aligned}$$

$$\begin{aligned} P(x_2 \mid x_1) &= 0.6 \\ P(\neg x_2 \mid x_1) &= 0.4 \end{aligned} \qquad \begin{aligned} P(x_3 \mid x_1) &= 0.1 \\ P(\neg x_3 \mid x_1) &= 0.9 \end{aligned}$$

$$\begin{aligned} P(x_2 \mid \neg x_1) &= 0.1 \\ P(\neg x_2 \mid \neg x_1) &= 0.9 \end{aligned} \qquad \begin{aligned} P(x_3 \mid \neg x_1) &= 0.8 \\ P(\neg x_3 \mid \neg x_1) &= 0.2 \end{aligned}$$

Conditional independence relation

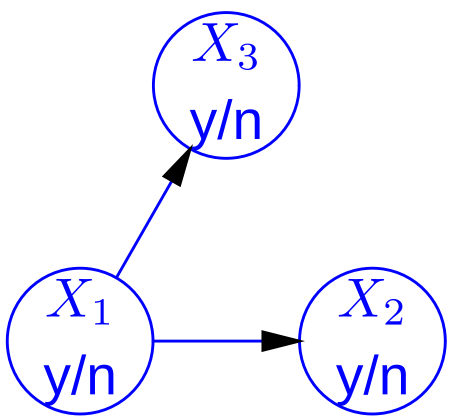
Let X, Y, Z be *sets* of variables, such that $X, Y, Z \subseteq V(G)$, then X is called conditionally independent of Y given Z, denoted as

$$X \perp\!\!\!\perp_P Y \mid Z$$

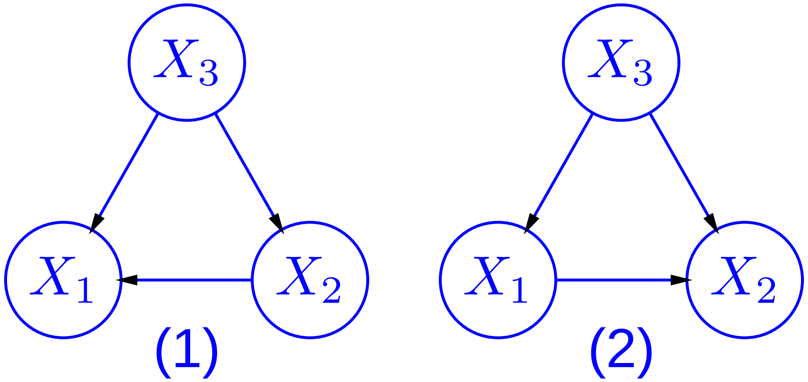
if and only if

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

Example: Representation of $X_2 \perp\!\!\!\perp_P X_3 \mid X_1$ in a directed graph



Chain rule - digraph



Factorisation (1):

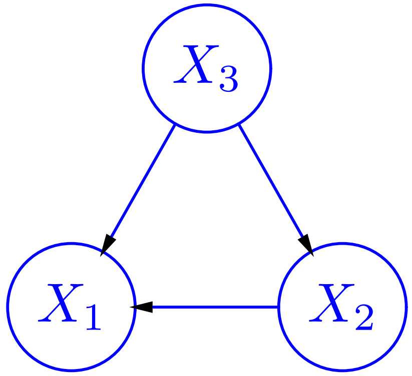
$$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3)P(X_2 \mid X_3)P(X_3)$$

Other factorisation (2):

$$P(X_1, X_2, X_3) = P(X_2 \mid X_1, X_3)P(X_1 \mid X_3)P(X_3)$$

$\Rightarrow$ different *factorisations* possible

Does the chain rule help?



$$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3)P(X_2 \mid X_3)P(X_3)$$

i.e. we need:

$$\begin{array}{rr} P(x_1 \mid x_2, x_3) & P(x_1 \mid x_2, \neg x_3) \\ P(\neg x_1 \mid x_2, x_3) & P(\neg x_1 \mid x_2, \neg x_3) \\ P(x_1 \mid \neg x_2, x_3) & P(x_1 \mid \neg x_2, \neg x_3) \\ P(\neg x_1 \mid \neg x_2, x_3) & P(\neg x_1 \mid \neg x_2, \neg x_3) \\ \vdots & \vdots \end{array}$$

Does the chain rule help?

$$\begin{array}{rr} \vdots & \vdots \\ P(x_2 \mid x_3) & P(x_3) \\ P(\neg x_2 \mid x_3) & P(\neg x_3) \\ P(x_2 \mid \neg x_3) & \\ P(\neg x_2 \mid \neg x_3) & \end{array}$$

So, 14 probabilities; however
$P(x_1 \mid X_2, X_3) = 1 - P(\neg x_1 \mid X_2, X_3)$,
$P(x_2 \mid X_3) = 1 - P(\neg x_2 \mid X_3)$, and $P(x_3) = 1 - P(\neg x_3)$

$\Rightarrow$ 7 probabilities required

How many did we have originally for $P(X_1, X_2, X_3)$?

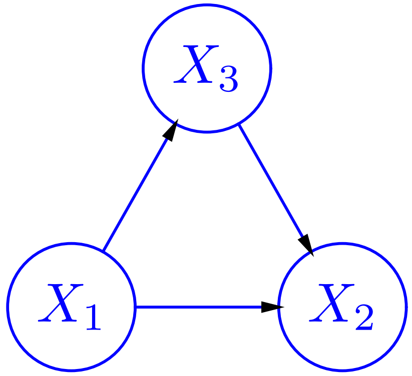
Does the chain rule help?

$$
\begin{array}{rr}
P(x_1, x_2, x_3) & P(x_1, x_2, \neg x_3) \\
P(\neg x_1, x_2, x_3) & P(\neg x_1, x_2, \neg x_3) \\
P(x_1, \neg x_2, x_3) & P(x_1, \neg x_2, \neg x_3) \\
P(\neg x_1, \neg x_2, x_3) & P(\neg x_1, \neg x_2, \neg x_3)
\end{array}
$$

8 required? No, because $\sum_{X_1,X_2,X_3} P(X_1, X_2, X_3) = 1$
Hence, e.g.

$$
\begin{aligned}
P(x_1, x_2, x_3) &= 1 - \sum_{X_2,X_3} P(\neg x_1, X_2, X_3) \\
&\quad - \sum_{X_3} P(x_1, \neg x_2, X_3) - P(x_1, x_2, \neg x_3)
\end{aligned}
$$

Let's use stochastic independence



$$P(X_1, X_2, X_3) = P(X_2 \mid X_1, X_3)P(X_3 \mid X_1)P(X_1)$$

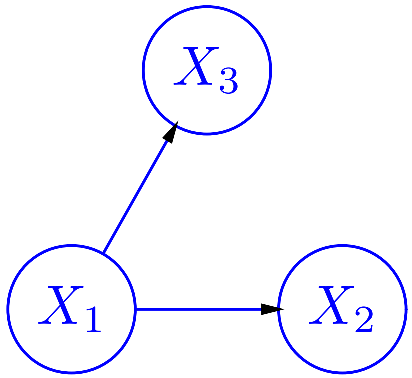
Now assume that X_2 and X_3 are conditionally independent given X_1:

$$P(X_2 \mid X_1, X_3) = P(X_2 \mid X_1)$$

and

$$P(X_3 \mid X_1, X_2) = P(X_3 \mid X_1)$$

Stochastic independence: does it help?



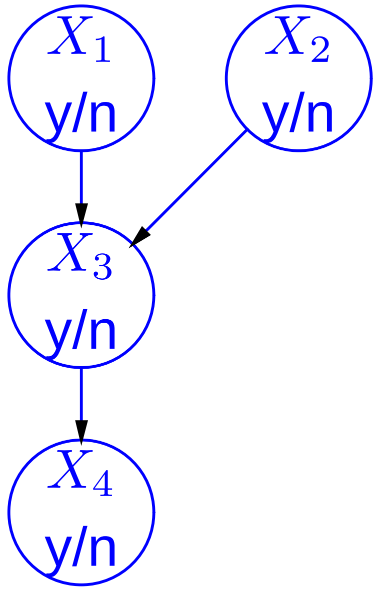
$$P(X_2 \mid X_1, X_3) = P(X_2 \mid X_1)$$

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_2 \mid X_1, X_3) P(X_3 \mid X_1) P(X_1) \\ &= P(X_2 \mid X_1) P(X_3 \mid X_1) P(X_1) \end{aligned}$$

Only $5 = 2 + 2 + 1$ probabilities required instead of 7

Probabilistic inference

Given:



$P(x_4 \mid x_3) = 0.4$
$P(x_4 \mid \neg x_3) = 0.1$
$P(x_3 \mid x_1, x_2) = 0.3$
$P(x_3 \mid \neg x_1, x_2) = 0.5$
$P(x_3 \mid x_1, \neg x_2) = 0.7$
$P(x_3 \mid \neg x_1, \neg x_2) = 0.9$
$P(x_1) = 0.6$
$P(x_2) = 0.2$

Then: $P(x_4) = P(x_4, x_3) + P(x_4, \neg x_3)$

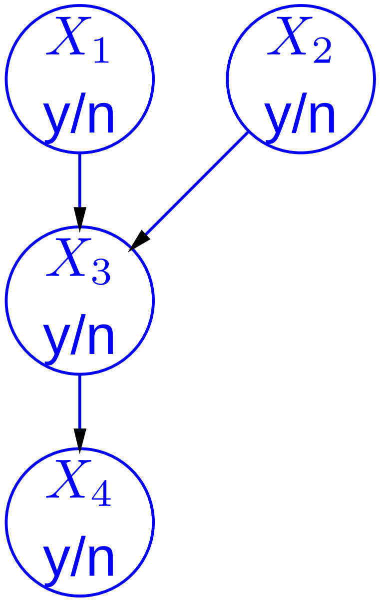
(marginalisation)

$= P(x_4 \mid x_3)P(x_3) + P(x_4 \mid \neg x_3)P(\neg x_3)$

(conditioning)

$$= \sum_{X_3} P(x_4 \mid X_3)P(X_3)$$

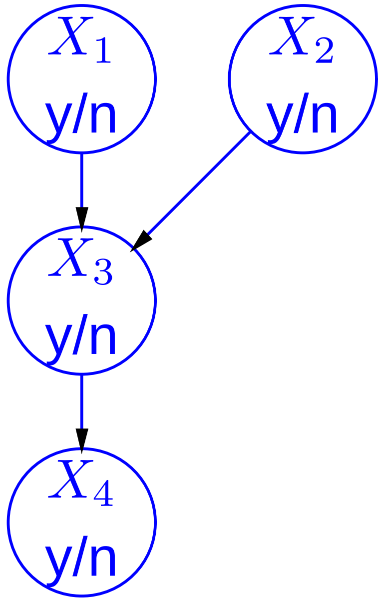
Probabilistic inference



$$P(x_4 \mid x_3) = 0.4$$
$$P(x_4 \mid \neg x_3) = 0.1$$
$$P(x_3 \mid x_1, x_2) = 0.3$$
$$P(x_3 \mid \neg x_1, x_2) = 0.5$$
$$P(x_3 \mid x_1, \neg x_2) = 0.7$$
$$P(x_3 \mid \neg x_1, \neg x_2) = 0.9$$
$$P(x_1) = 0.6$$
$$P(x_2) = 0.2$$

$P(X_3)$ = ? $\iff$ Compute $P(x_3)$ and $P(\neg x_3)$

Probabilistic inference



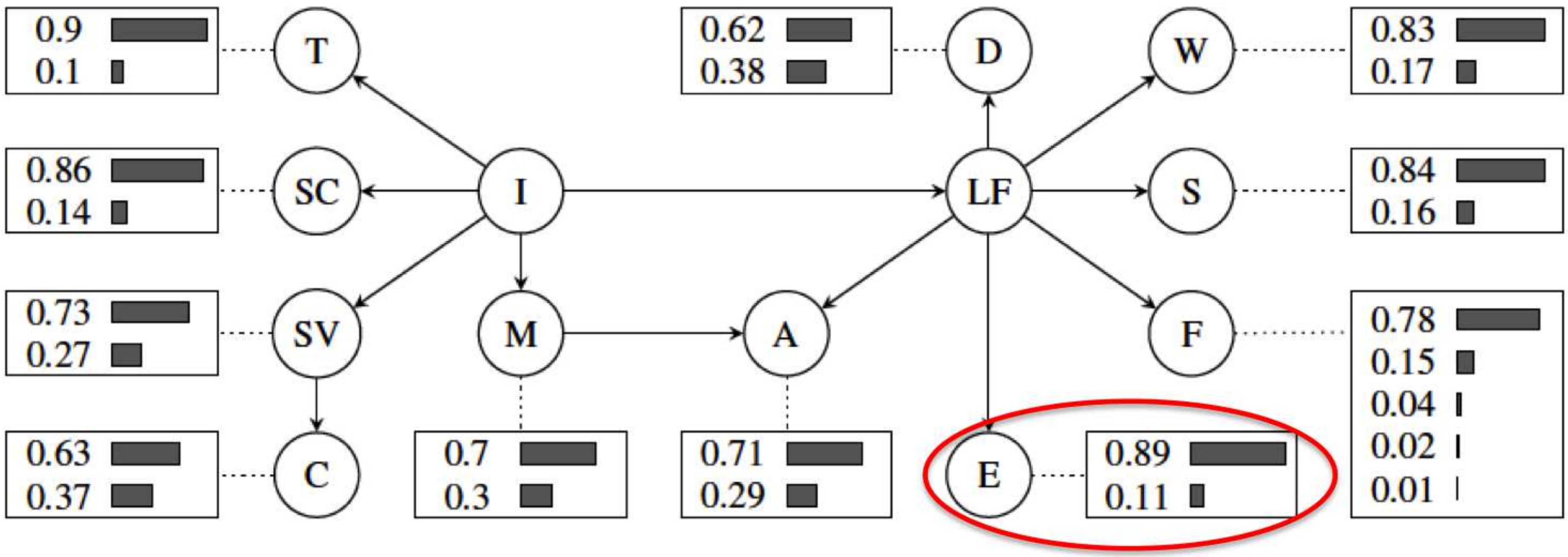
$P(x_4 \mid x_3) = 0.4$
$P(x_4 \mid \neg x_3) = 0.1$
$P(x_3 \mid x_1, x_2) = 0.3$
$P(x_3 \mid \neg x_1, x_2) = 0.5$
$P(x_3 \mid x_1, \neg x_2) = 0.7$
$P(x_3 \mid \neg x_1, \neg x_2) = 0.9$
$P(x_1) = 0.6$
$P(x_2) = 0.2$

$$\begin{aligned} P(x_3) &= \sum_{X_1, X_2} P(x_3, X_1, X_2) \\ &= \sum_{X_1, X_2} P(x_3 \mid X_1, X_2) P(X_1, X_2) \\ &= \sum_{X_1, X_2} P(x_3 \mid X_1, X_2) P(X_1) P(X_2) = 0.7 \end{aligned}$$

$\Rightarrow P(x_4) = \sum_{X_3} P(x_4 \mid X_3) P(X_3) = 0.4 \cdot 0.7 + 0.1 \cdot 0.3 = 0.31$

Back to the mobile application

Prediction of COPD (lung disease) exacerbations



Expert opinion based Bayesian network with prior probabilities shown (top probability is Normal state). A = activity, C = cough, D = dyspnea, E = exacerbation, F = FEV_1, I = infection, LF = lung function, M = malaise, S = SpO_2, SC = sputum colour, SV = sputum volume, T = temperature and W = Wheeze.

Popular applications of BNs

- Medical diagnosis and therapy selection: BNs are now the most popular paradigm for medical intelligent systems
  - Mobile healthcare applications by MBSD-Radboud University
    – COPD monitoring (`www.youtube.com/watch?v=zfqW8rX0OpM`)
    – Pregnancy monitoring (`www.youtube.com/watch?v=Ize-ydS1UiU`)
- Software/Hardware troubleshooting: Microsoft, Boeing, HP
- Biological modelling: gene expressions
- Art: orchestral music accompaniment
  `xavier.informatics.indiana.edu/~craphael/music_plus_one/`
- and more . . . see, e.g.,
  “Bayesian Networks: A Practical Guide to Applications”
  Olivier Pourret (Ed.), Patrick Naïm and Bruce Marcot, Wiley, March 2008

Bayesian networks software

- Some software companies in this area:
  - Hugin (Denmark): `www.hugin.dk`
  - Norsys (USA): `www.norsys.com`
  - Knowledge Industries (USA): `www.kic.com`
  - Bayesia (France): `www.bayesia.com`
- Some public domain software:
  - JavaBayes: `www.cs.cmu.edu/~javabayes`
  - BayesBuilder: `www.snn.ru.nl/nijmegen`
  - bnlearn package in R: `www.bnlearn.com`
  - SamIam: `reasoning.cs.ucla.edu/samiam`
  - Matlab BNT Toolbox: `code.google.com/p/bnt`
  - and many more at `www.cs.ubc.ca/~murphyk/Software/bnsoft.html`