Adapted from AIMA slides

Bayesian networks

Peter Antal
antal@mit.bme.hu

Outline

- Reminder: inference in the joint distribution
- Reminder: properties of independencies, independence models
- Using independences: representation & inference
  - Example: Naive Bayesian networks
  - Example: Hidden Markov models
- Bayesian networks
- Beyond independence models: causal diagrams
- Inference in causal models
- Special local models
  - Noisy-OR
  - Decision tree CPDs
  - Decision graph CPDs
- ➔Learning Bayesian networks

Inference by enumeration

Every question about a domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** – **Y** – **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Properties of independence

a Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ iff } I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

b Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

c Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

d Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

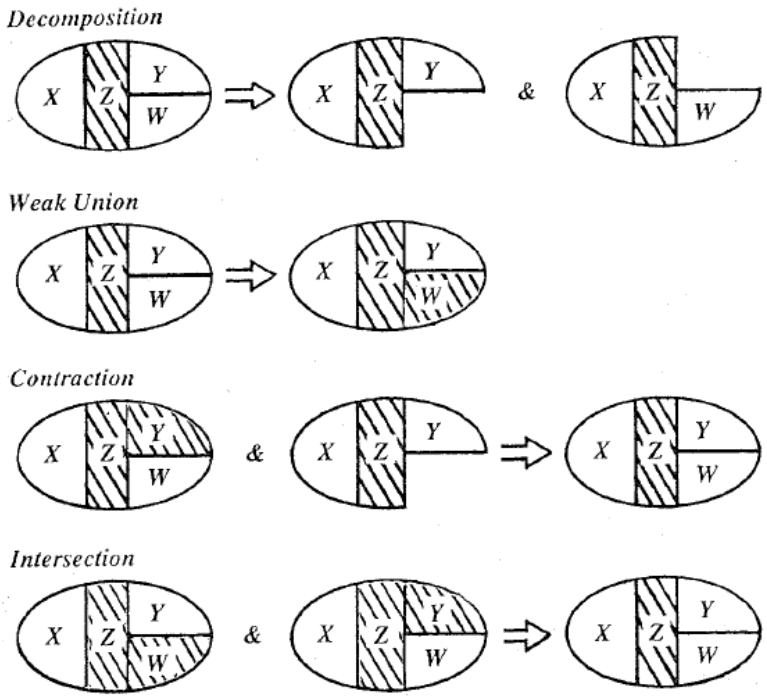
e Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

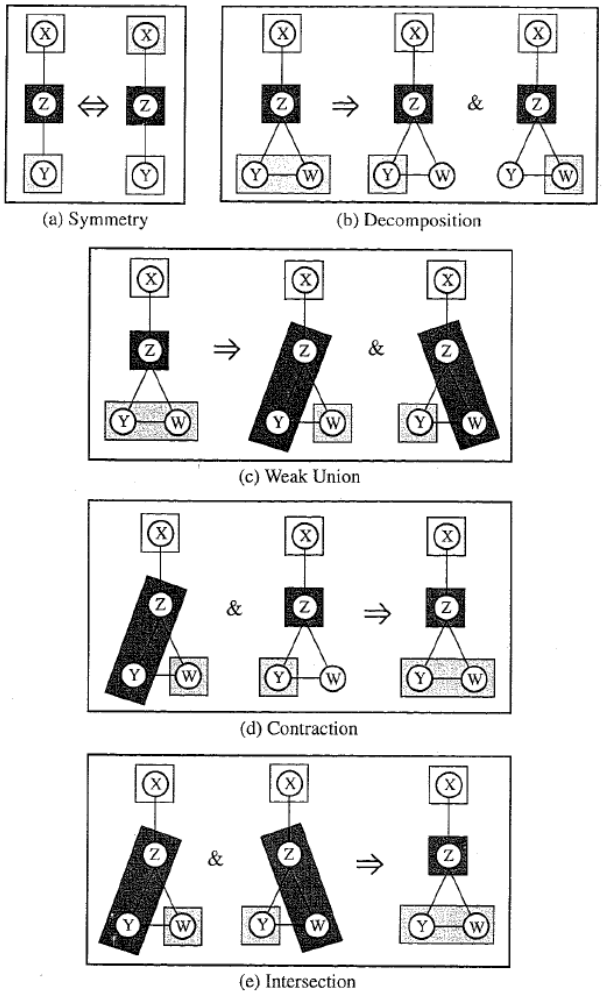
$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W}) \textit{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

Semi-graphoids, graphoids

Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution). SG is sound, but incomplete inference.

Graphoids: Semi-graphoids+Intersection
(holds only in strictly positive distribution)




(a) Symmetry
(b) Decomposition
(c) Weak Union
(d) Contraction
(e) Intersection

J.Pearl: Probabilistic Reasoning in intelligent systems, 1998

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$

If P(X,Y,Z) is a Markov chain, then

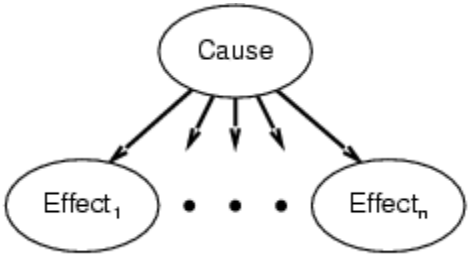
$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: D(X;Z)

Exceptionally: I(X;Z)

Naive Bayesian network

Assumptions:



1, Two types of nodes: a cause and effects.

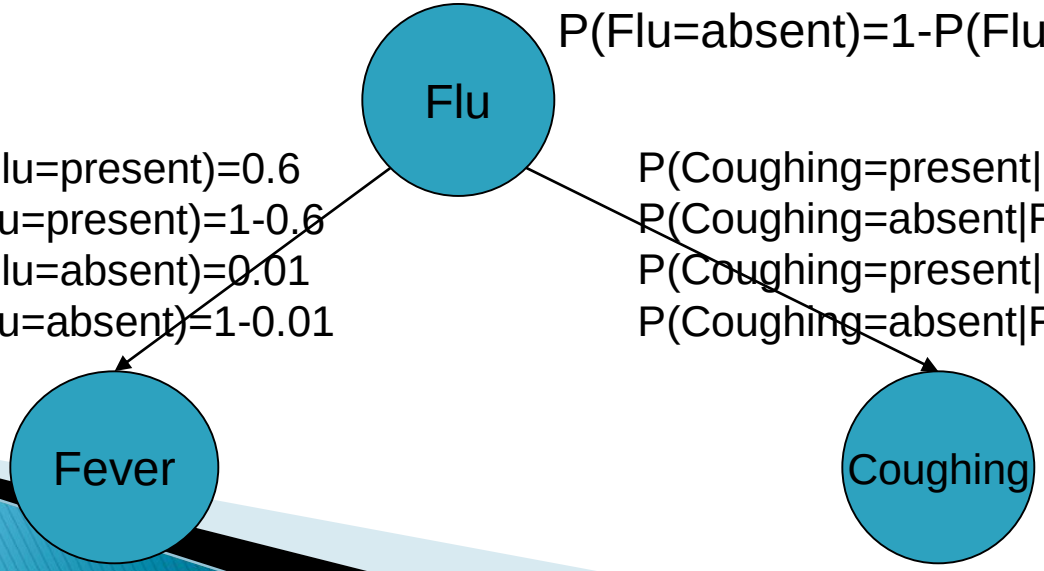
2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model

P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)



P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1-0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1-0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

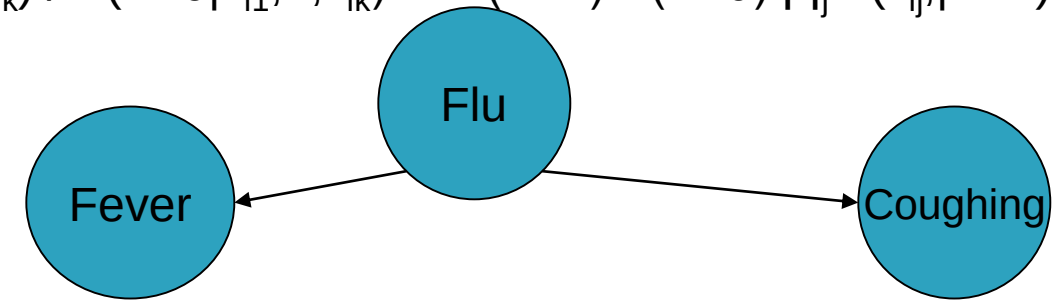
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

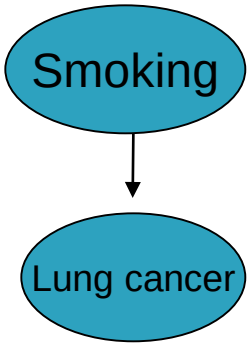
$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \mid Fever = absent, Coughing = present)$$

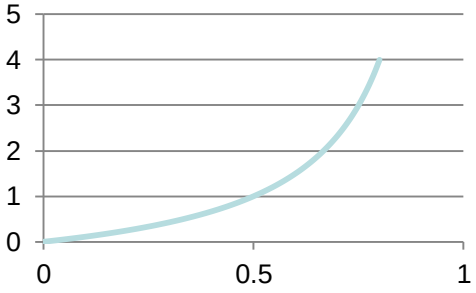
$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Conditional probabilities, odds, odds ratios

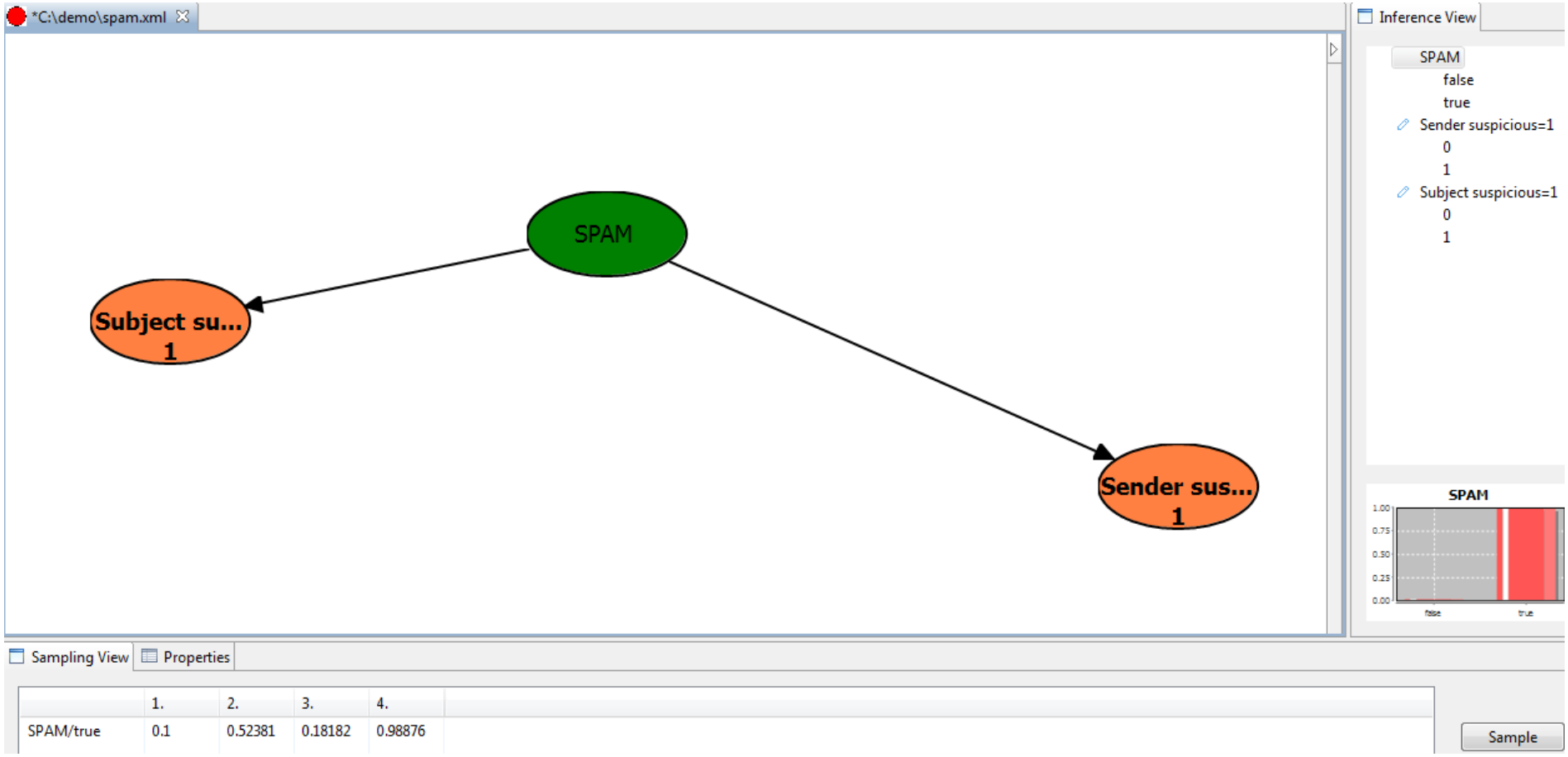


	¬S	S	
¬LC	P(¬S, ¬LC)	P(S, ¬LC)	P(¬LC)
LC	P(¬S, LC)	P(S, LC)	P(LC)
	P(¬S)	P(S)	

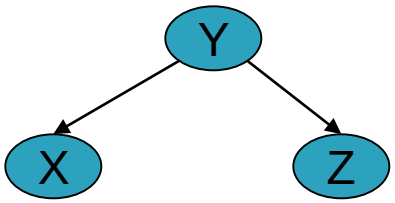
Probability:
P(LC)
Conditional probabilities (e.g., probability of LC given S)**:**
P(LC| ¬S)= ??? P(LC| S)= ??? P(LC)
Odds:
[0,1] →[0,∞]: Odds(p)=p/(1-p)
O(LC| ¬S)= ??? O(LC| S)
Odds Ratio (OR) Independent of prevalence!
OR(LC,S)=O(LC| S)/O(LC| ¬S)



Example: Construct a spam filter



The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then

M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}

Normally/almost always: D(X;Z)

Exceptionally: I(X;Z)

Learning of N-BNs?

- Bayesian learning?
- Identification of
  - parameters,
  - structure?

The Beta distribution

3. Definition. *A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.*

1. Example. *Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.*

$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{13}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \; where \; c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{14}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

In general a conjugate prior is updated to posterior using only an appropriate statistics of the observations to update its parametrization. It shows that the parameters frequently has an intuitive interpretation based on observations, that is in the prior specification the parameters corresponds to real or virtual past observations.

Parameter independence

For a Bayesian network structure G, the global parameter independence assumption means that

$$P(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i|G), \tag{1}$$

where $\boldsymbol{\theta}_i$ denotes the parameters corresponding to the conditional $p(X_i|Pa(X_i))$ in G. The local parameter independence assumption means that

$$p(\boldsymbol{\theta}_i|G) = \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|G), \tag{2}$$

where q_i denotes the number of parental configurations $(pa(X_i))$ for X_i in G and $\boldsymbol{\theta}_i j$ denotes the parameters corresponding to the conditional $p(X_i|pa(X_i)_j)$ in some fixed ordering of the $pa(X_i)$ configurations. The parameter independence assumption means global and local parameter independence.

Dirichlet parameter priors

Theorem

The assumption of positive densities, likelihood equivalence and parameter independence for complete structures G_c implies that $p(\boldsymbol{\theta}_U|\xi)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n}$.

The $p(\boldsymbol{\theta}_i|G_i) = J_{G_i} p(\boldsymbol{\theta}_V|\xi)$, where J_{G_i} is the Jacobian of the transformation from $\boldsymbol{\theta}_V$ to $\boldsymbol{\theta}_{G_i}$. To state the following theorem it is convenient to rewrite the hyperparameters as $N' = \sum_{x_1,\ldots,x_n} N_{x_1,\ldots,x_n}$ called prior/virtual sample size and $p^{prior} x_1, \ldots, x_n = E[\theta_{x_1,\ldots,x_n}] = N_{x_1,\ldots,x_n}/N'$. Furthermore, we need the following concept.

Definition

The parameter modularity assumption means that if $pa(X_i)$ are identical in two Bayesian network structures G_1, G_2, then

$$p(\boldsymbol{\theta}_{ij}|G_1) = p(\boldsymbol{\theta}_{ij}|G_2), \tag{34}$$

where $\boldsymbol{\theta}_i j$ denotes the parameters corresponding to the conditional $p(X_i|pa(X_i)_j)$ in some fixed ordering of the $pa(X_i)$ configurations.

Bayesian N-BNs

- Mixture of experts?
- Parameters
  - Priors for parameters
    - Discrete
    - Continuous (for nominal variables): Beta, Dirichlet
  - Super parametrization(!): BMA-equivalent (Bayesian Model Averaging, BMA)
- Structures
  - Priors for structures
  - Edge priors
  - Super parametrization(!): full-BMA-equivalent

Extensions of N-BNs

- Tree-augmented BNs
- BN-augmented BNs
- Hierarchical BNs
- Multiple parents
  - Explaining away
- „Context-sensitive" N-BNs

Hidden Markov Models (HMMs)

The world changes; we need to track and predict it

Diabetes management vs vehicle diagnosis

Basic idea: copy state and evidence variables for each time step

$\mathbf{X}_t$ = set of unobservable state variables at time t
e.g., $BloodSugar_t$, $StomachContents_t$, etc.

$\mathbf{E}_t$ = set of observable evidence variables at time t
e.g., $MeasuredBloodSugar_t$, $PulseRate_t$, $FoodEaten_t$

This assumes **discrete time**; step size depends on problem

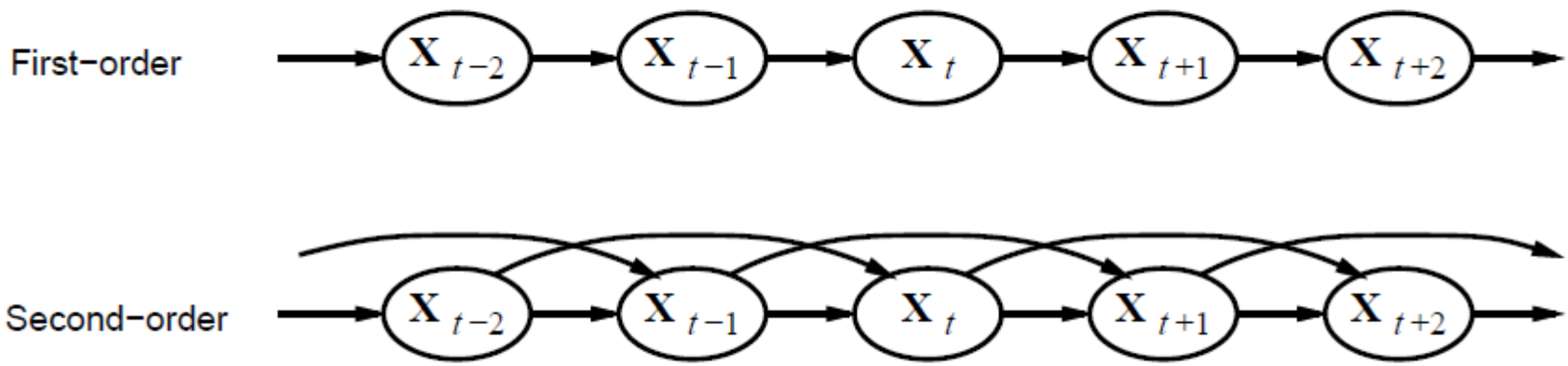
Notation: $\mathbf{X}_{a:b} = \mathbf{X}_a, \mathbf{X}_{a+1}, \ldots, \mathbf{X}_{b-1}, \mathbf{X}_b$

Markov chains

Markov assumption: $\mathbf{X}_t$ depends on **bounded** subset of $\mathbf{X}_{0:t-1}$

First-order Markov process: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$

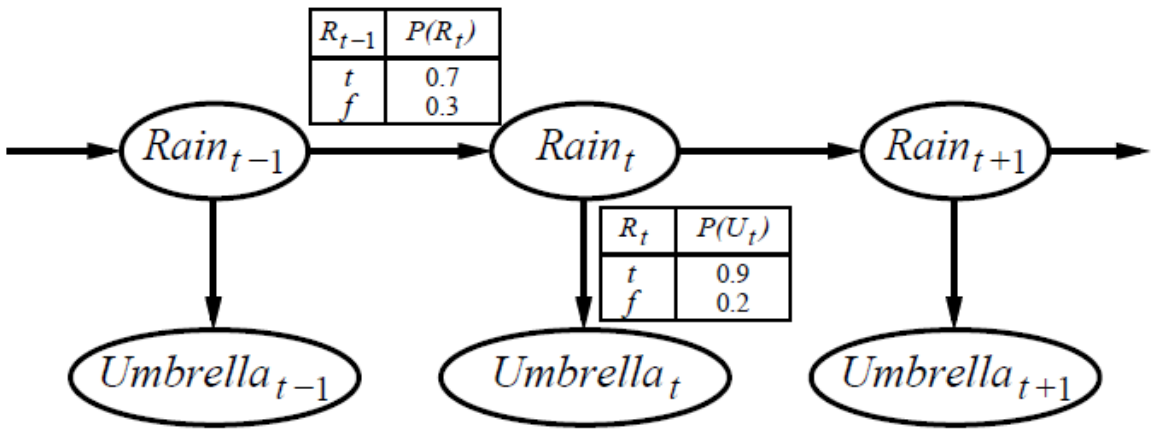
Second-order Markov process: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-2}, \mathbf{X}_{t-1})$



Sensor Markov assumption: $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_{0:t}, \mathbf{E}_{0:t-1}) = \mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$

Stationary process: transition model $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$ and sensor model $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$ fixed for all t

Example



First-order Markov assumption not exactly true in real world!

Possible fixes:

1. **Increase order** of Markov process
2. **Augment state**, e.g., add $Temp_t$, $Pressure_t$

Example: robot motion.

Augment position and velocity with $Battery_t$

Inference in HMMs

Filtering: $\mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$

belief state—input to the decision process of a rational agent

Prediction: $\mathbf{P}(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$ for $k > 0$

evaluation of possible action sequences;

like filtering without the evidence

Smoothing: $\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t})$ for $0 \leq k < t$

better estimate of past states, essential for learning

Most likely explanation: $\arg\max_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t}|\mathbf{e}_{1:t})$

speech recognition, decoding with a noisy channel

Filtering

Aim: devise a **recursive** state estimation algorithm:

$$\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) = f(\mathbf{e}_{t+1}, \mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t}))$$

$$\begin{aligned}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}, \mathbf{e}_{1:t})\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t})\end{aligned}$$

I.e., prediction + estimation. Prediction by summing out $\mathbf{X}_t$:

$$\begin{aligned}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\Sigma_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t, \mathbf{e}_{1:t})P(\mathbf{x}_t|\mathbf{e}_{1:t}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\Sigma_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{e}_{1:t})\end{aligned}$$

$\mathbf{f}_{1:t+1} = \textsc{Forward}(\mathbf{f}_{1:t}, \mathbf{e}_{t+1})$ where $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$

Time and space **constant** (independent of t)

Learning HMMs

- Parameter learning
- Structure learning
  - HMMs are equivalent with stochastic finite state automatons

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

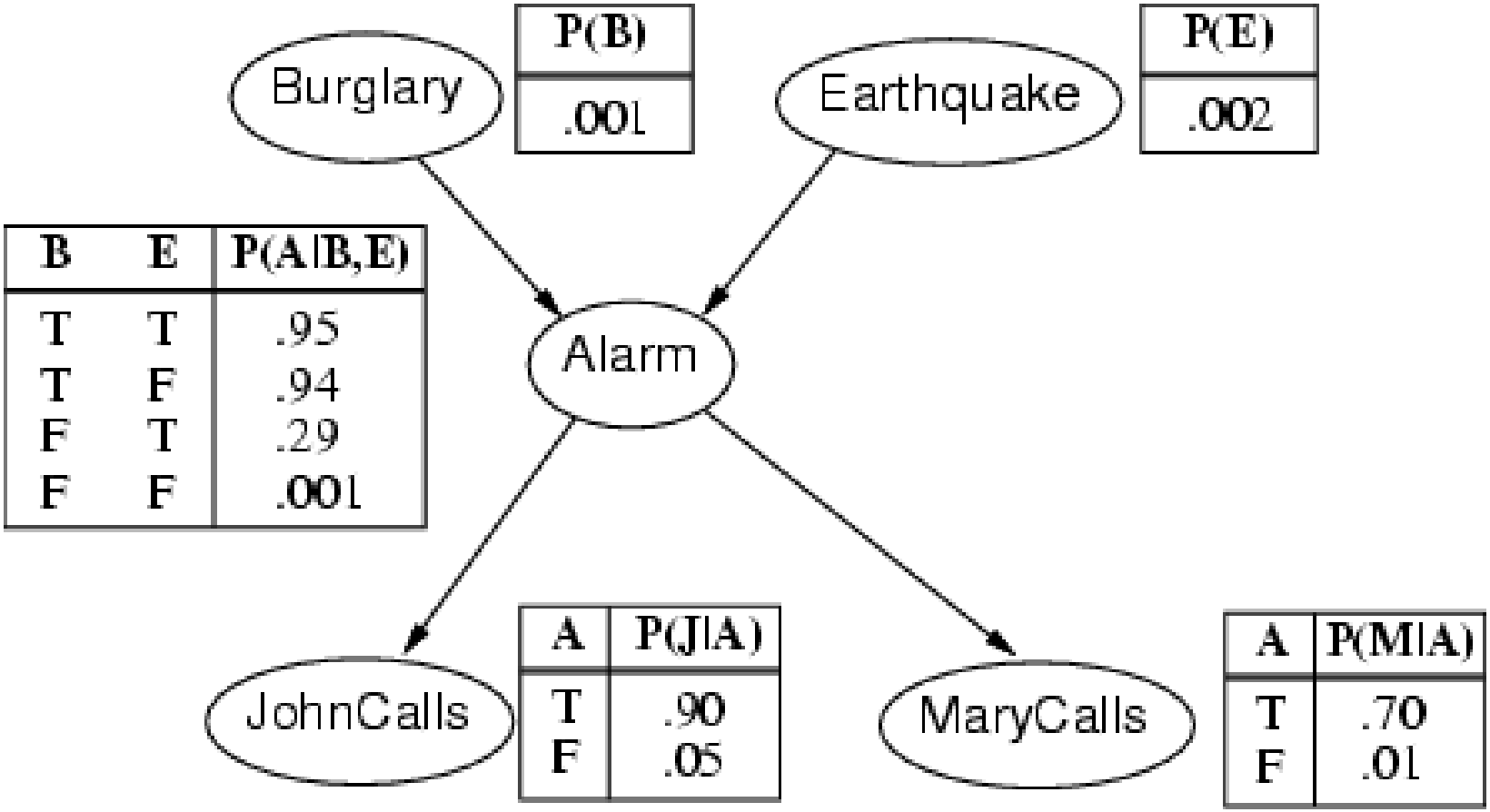
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



Learning Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\boldsymbol{P}(X_i \mid Parents(X_i)) = \boldsymbol{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

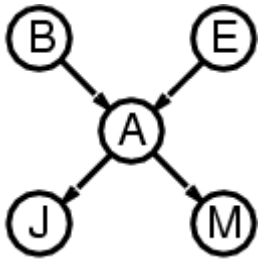
$$\boldsymbol{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \boldsymbol{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad //\text{(chain rule)}$$

$$= \pi_{i=1}^{n} \boldsymbol{P}(X_i \mid Parents(X_i)) \quad //\text{(by construction)}$$

Ordered Markov condition (OMC) is satisfied: P obeys the OMC w.r.t. G.

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

Noisy–OR

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone

$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

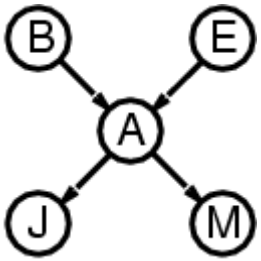
$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$



e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$$

Inference in BNs

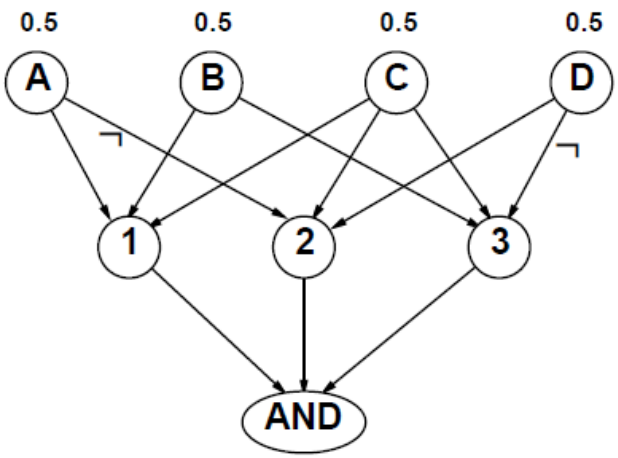
Singly connected networks (or polytrees):

- any two nodes are connected by at most one (undirected) path
- time and space cost of exact inference $O(d^k n)$

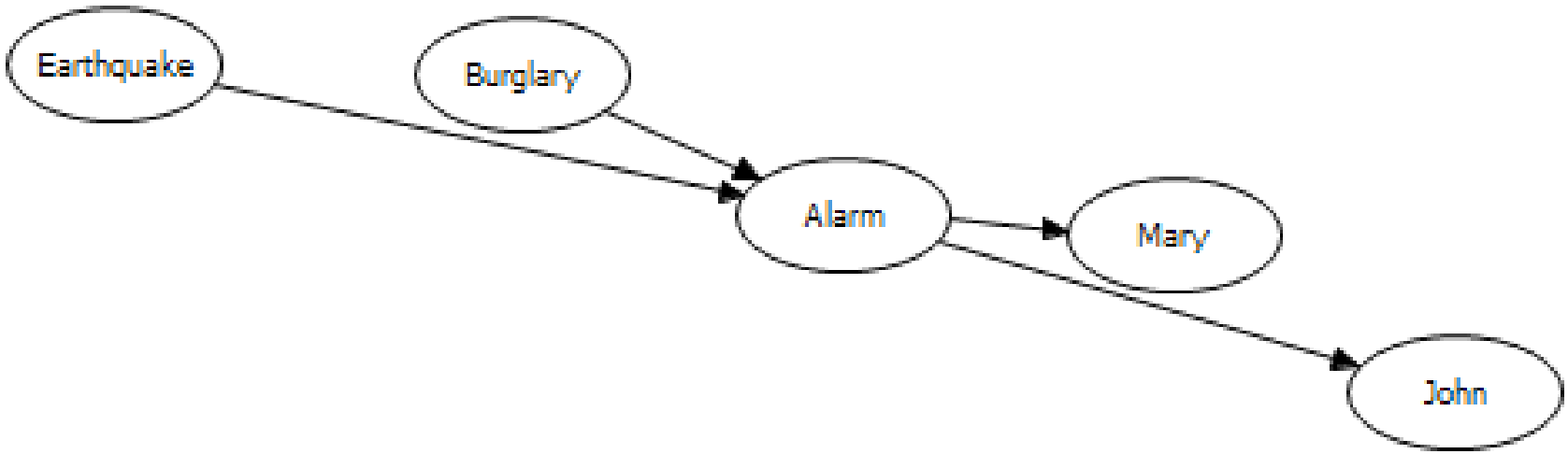
Multiply connected networks:

- can reduce 3SAT to exact inference: 0<p(AND)? $\Rightarrow$ NP-hard
- equivalent to **counting** 3SAT models $\Rightarrow$ #P-complete

1. A v B v C
2. C v D v ¬A
3. B v C v ¬D



Learning with an „acausal" ordering



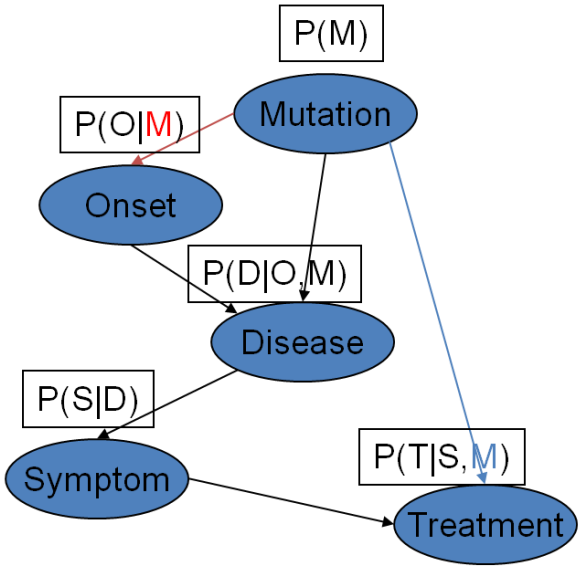
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n

   add X_i to the network

   select parents from $X_1, \ldots, X_{i-1}$ such that

$$\boldsymbol{P}(X_i \mid Parents(X_i)) = \boldsymbol{P}(X_i \mid X_1, \ldots X_{i-1})$$

Bayesian networks: interpretations



1. Causal model

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$

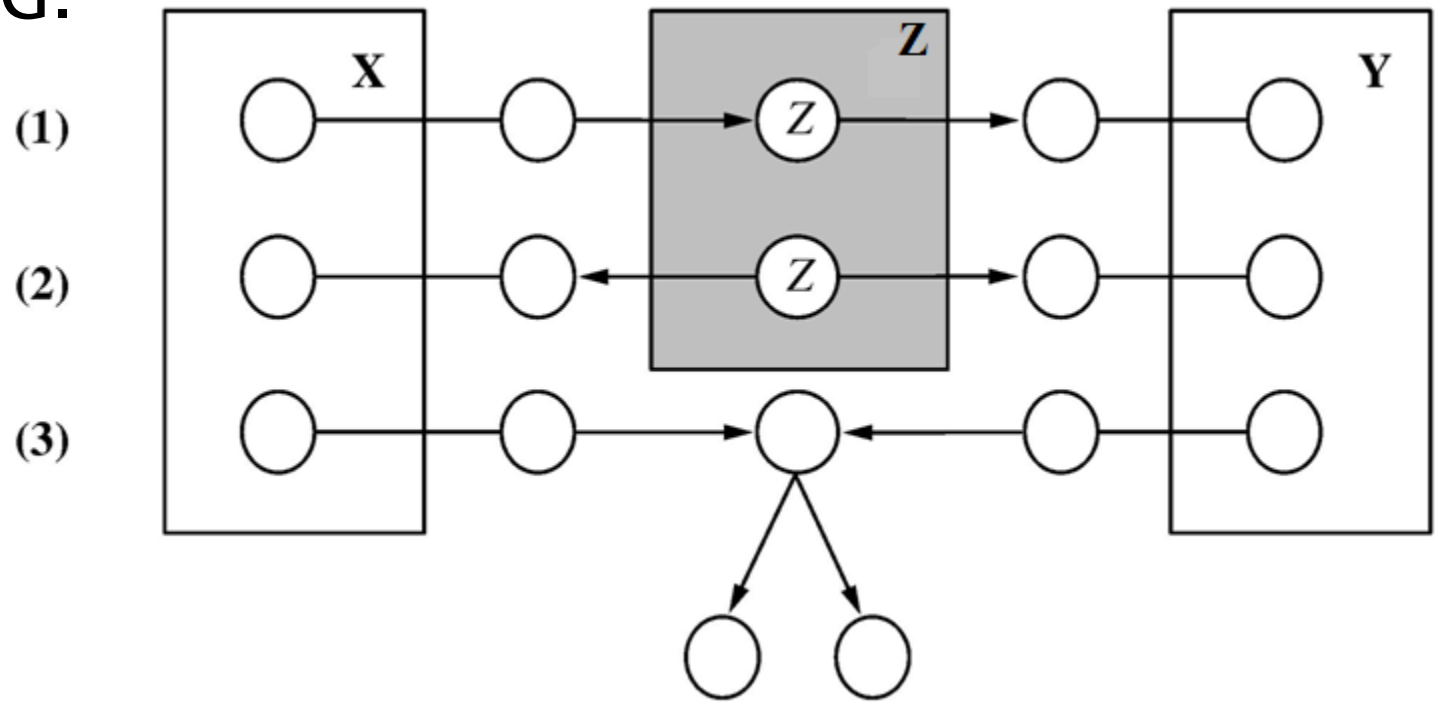
2. Graphical representation of (in)dependencies

3. Concise representation of joint distributions

$$P(M,O,D,S,T)=$$
$$P(M)P(O\mid M)P(D\mid O,M)P(S\mid D)P(T\mid S,M)$$

Inferring independencies from structure: d-separation

$I_G(X;Y|Z)$ denotes that X is d-separated (directed separated) from Y by Z in directed graph G.



d-separation and the global Markov condition

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the* global Markov *condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq U\ (X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P, \tag{9}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated *by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows*

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

Bayesian network definitions

Theorem 1 *Let $P(U)$ a probability distribution and G a DAG, then the conditions above (repeated below) are equivalent:*

- **F** P *is Markov relative* G *or* P *factorizes w.r.t* G,
- **O** P *obeys the ordered Markov condition w.r.t.* G,
- **L** P *obeys the local Markov condition w.r.t.* G,
- **G** P *obeys the global Markov condition w.r.t.* G.

Definition 8 *A directed acyclic graph (DAG)* G *is a* Bayesian network *of distribution* $P(U)$ *iff the variables are represented with nodes in* G *and* (G, P) *satisfies any of the conditions* F, O, L, G *such that* G *is minimal (i.e. no edge(s) can be omitted without violating a condition* F, O, L, G*).*

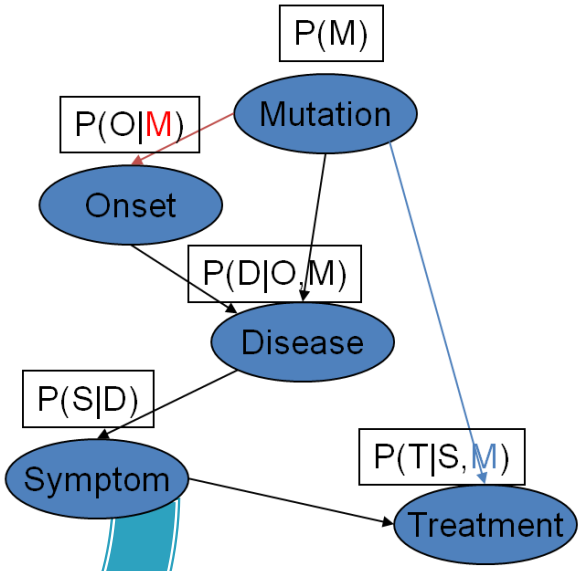
Bayesian networks

Directed acyclic graph (DAG)

- nodes – random variables/domain entities
- edges – direct probabilistic dependencies (edges– causal relations

Local models – $P(X_i|Pa(X_i))$

Three interpretations:



1. Causal model

3. Concise representation of joint distributions

$$P(M,O,D,S,T) = P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M)$$

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$

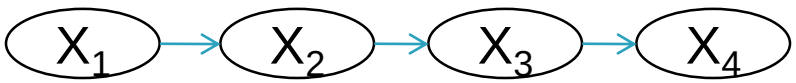
2. Graphical representation of (in)dependencies

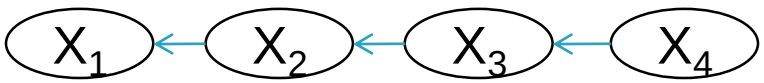
A practical definition

Definition 9 *A Bayesian network model M of domain with variables U consists of a structure G and parameters $\boldsymbol{\theta}$. The structure G is a DAG such that each node represents a variable and local probabilistic models $p(X_i|pa(X_i))$ are attached to each node w.r.t. the structure G, that is they describe the stochastic dependency of variable X_i on its parents $pa(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\boldsymbol{\theta_i}$, and $\boldsymbol{\theta}$ denotes the overall parameterezation of the model.*

Association vs. Causation: Markov chain

Causal models:





Markov chain

$P(X_1,\ldots)$

$M_P=\{I(X_{i+1};X_{i-1}|X_i)\}$

„first order Markov property"

Flow of time?

The building block of causality: v-structure (arrow of time)

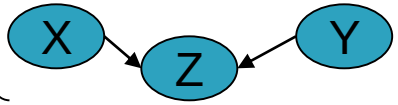
p(X),p(Z|X),p(Y|Z)

p(X|Z),p(Z|Y),p(Y)

p(X|Z),p(Z),p(Y|Z)

"transitive" M ≠ „intransitive" M

p(X),p(Z|X,Y),p(Y)



„v-structure"

M_P={D(X;Z), D(Z;Y), D(X,Y), I(X;Y|Z)}

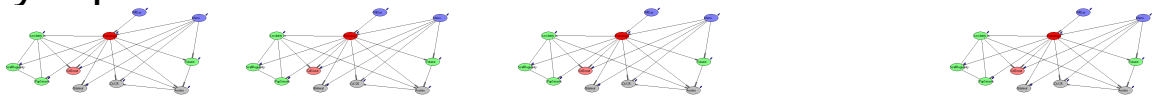
M_P={D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z) }

Often: present knowledge renders future states conditionally independent.
(confounding)

Ever(?): present knowledge renders past states conditionally independent.
(backward/atemporal confounding)

Observational equivalence of causal models

Observationally equivalent causal models:



= = = = d-separation

Independence model:

$$P(X_1,\ldots, X_n)$$

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$$

Different causal models can have the same independence map!

Typically causal models cannot be identified from passive observations, they are ***observationally equivalent***.

Observational equivalence of causal models

Definition 11 *Two DAGs* G_1, G_2 *are* *observationally equivalent* *, if they imply the same set of independence relations (i.e.* $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$*).*

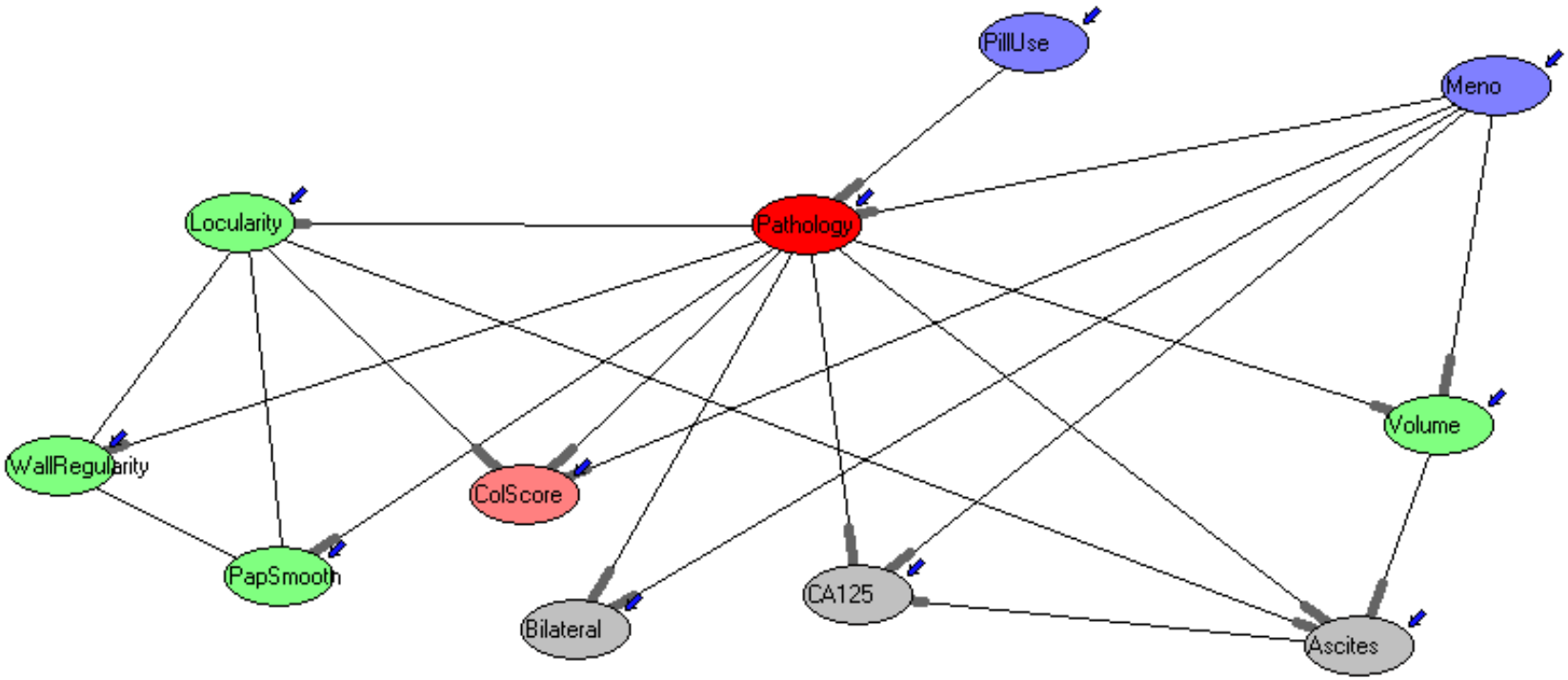
The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

Theorem 2 *Two DAGs* G_1, G_2 *are* *observationally equivalent* *, iff they have the same skeleton (i.e. the same edges without directions) and the same set of* *v-structures* *(i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The* *essential graph* *representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called* *compelled edges* *of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

A limits of learnability: compelled edges

("can we interpret edges as causal relations?"→compelled edges)



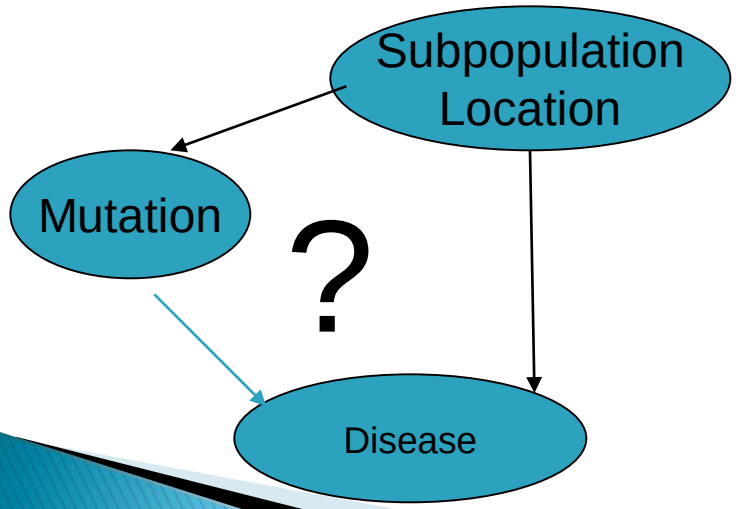
Types of inference in causal BNs

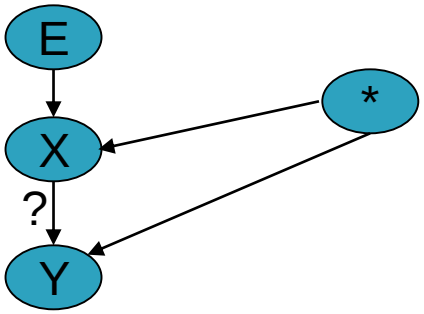
- (Passive, observational) inference
  - P(Query|Observations)
- **Interventionist inference**
  - **P(Query|Observations, Interventions)**
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Interventions and graph surgery

If G is a causal model, then compute p(Y|do(X=x)) by

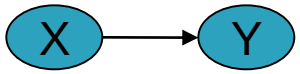
1. deleting the incoming edges to X
2. setting X=x
3. performing standard Bayesian network inference.



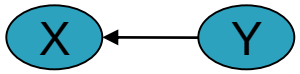


Association vs. Causation

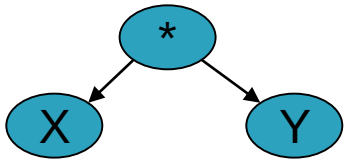
Causal models:



X causes Y

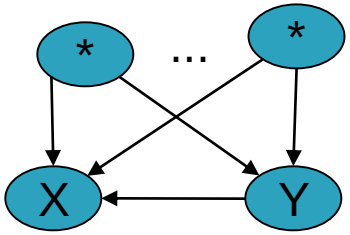


Y causes X



...

There is a common cause (pure confounding)

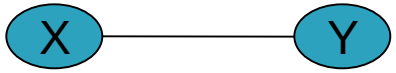


Causal effect of Y on X is confounded by many factors

From passive observations:

P(X,Y)

M_P={D(X;Y)}



„X and Y are associated”

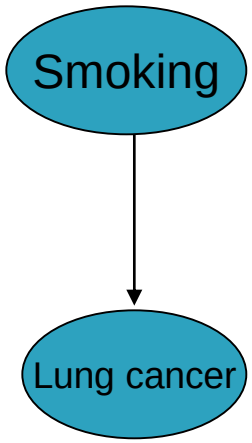
Reichenbach's Common Cause Principle:

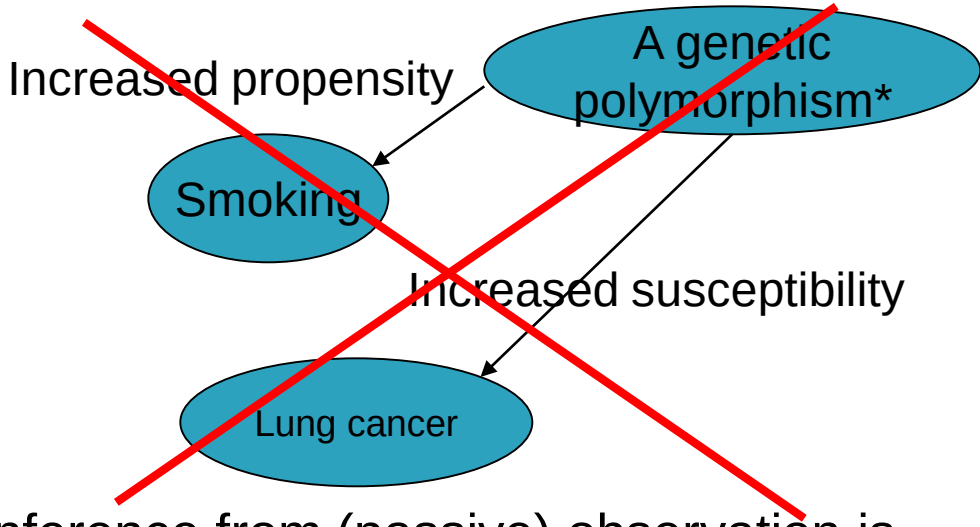
a correlation between events *X* and *Y* indicates either that *X* causes *Y*, or that *Y* causes *X*, or that *X* and *Y* have a common cause.

Local Causal Discovery

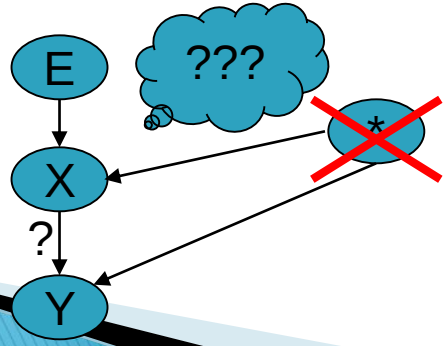
"can we interpret edges as causal relations in the presence of hidden variables?"

- Can we learn causal relations from observational data in presence of confounders???





- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



A deterministic concept of causation

- H.Simon
  - $X_i=f_i(X_1,..,X_{i-1})$ for i=1..n
  - In the linear case the sytem of equations indicates a natural causal ordering (flow of time?)

					X
				X	X
			X	X	X
		X	X	X	X
	....				

The probabilistic conceptualization is its generalization:

$$P(X_i,|X_1,..,X_{i-1}) \sim X_i=f_i(X_1,..,X_{i-1})$$

Summary

- Conditional independencies allows:
  - efficient representation of the joint probabilitity distribution,
  - efficient inference to compute conditional probabilites.
- Bayesian networks use directed acyclic graphs to represent
  - conditional independencies,
  - conditional parameters,
  - optionally, causal mechanisms.
- Design of variables and order of the variables can drastically influence structure
- Canonical conditional models can further increase efficiency.