OPPA European Social Fund
Prague & EU: We invest in your future.

Graphical probabilistic models – introduction

Jiří Kléma

Department of Cybernetics,
FEE, CTU at Prague

http://ida.felk.cvut.cz

GPM lectures – an overview

- L1: introduction
  - Bayesian networks – motivation and definitions,
  - how graphs can help – conditional independence,
- L2: inference
  - network applications in predictive tasks,
  - inference engine – fundamental algorithms,
- L3: learning networks from data
  - using networks for modelling,
  - networks as tools for understanding of relations among variables,
- L4: extensions
  - time, continuous variables, undirected graphs,
- L5: simple (restricted) graph models
  - feasible models in expert systems,
  - final exam – form, questions.

Agenda

- Motivation for graphical models
  - general probabilistic model and its curse of dimensionality,
  - general probabilistic model and knowledge?
- conditional independence
  - definition, examples,
  - graph equivalent – d-separation,
  - graph equivalence wrt conditional independence,
- essential types of graphical probabilistic models
  - brief categorization,
- Bayesian networks
  - basic idea behind,
  - example – family house with a dog,
  - fundamental tasks and their complexity.

Notation (binary random variables):

$A\ldots$ random variable, $a \ldots A = True$, $\neg a \ldots A = False$, $Pr(A, B)$ …joint probability distribution (a table), $Pr(a, b) = Pr(A = True, B = True)$ …prob of a particular event (a single item in table $Pr(A, B)$).

Why not a general probabilistic model?

- Ex.: 3 statements about world (people), each statement valid or invalid for a person
  - the world can be captured by joint probability,
- H: The person is higher then 180cm. M: The person is a man. Z: The person is a jockey.
  - women and men are equally frequent, men tend to be tall, a jockey is mostly a short man,

T	M	J	Pr(T,M,J)	$j \Rightarrow \neg t$
F	F	F	0.298	F
F	F	T	0.002	T
F	T	F	0.245	F
F	T	T	0.005	T
T	F	F	0.199	T
T	F	T	0.001	T
T	T	F	0.248	T
T	T	T	0.002	T
			1	

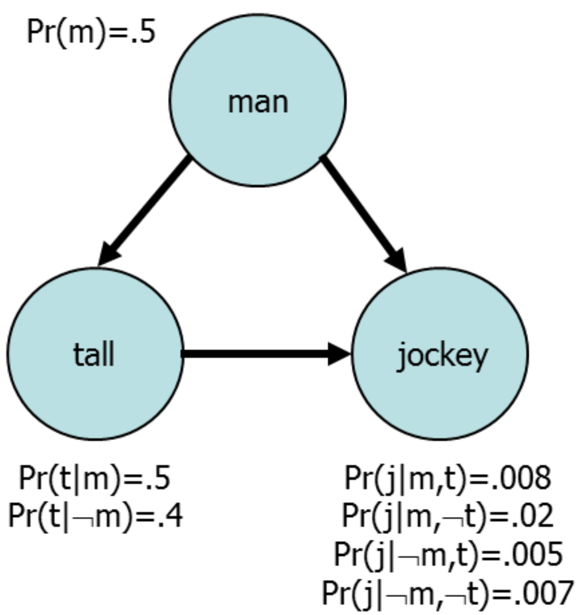
- probability of a formula equals the sum of probability of interpretations that satisfy it
  - Pr(t)= 0.199 + 0.001 + 0.248 + 0.002 = 0.45, 45% of population is tall,
  - Pr($j \Rightarrow \neg t$)= 1 - 0.001 - 0.002 = 0.997, 99.7% of population is not tall or not a jockey,
- arbitrary probabilistic operations can also be applied
  - $Pr(\neg t|j) = \frac{Pr(\neg t,j)}{Pr(j)} = \frac{0.007}{0.01} = 0.7$, 70% of jockeys are not tall
  - $Pr(m|j) = \frac{Pr(m,j)}{Pr(j)} = \frac{0.007}{0.01} = 0.7$, 70% of jockeys are men knowing a person is a jockey, in 70% cases it is a man as well

Why not a general probabilistic model?

- universality makes an asset of this model
  - identical and trivial model structure for all problems,
  - for a **sufficient** sample size its learning converges
    - model learning means to estimate (joint) probabilities,
- intractable for real problems
  - $2^n - 1$ probabilities when dealing with n propositions (for discrete variables a different base, for continuous parametric models),
  - infeasible for experts, the same holds for empirical settings based on data,
  - even if probs were known, still exponential in memory and inference time
    - obvious for a joint continuous distribution function,
    - **curse of dimensionality** - the number of observations needed grows exp with the number of variables,
- impenetrable for real tasks
  - model gives no explicit **knowledge** about the domain,
  - relations among objects remain hidden in a flood of numbers.

The ways to simplify and better organize the model?

- utilize the domain knowledge:
  - is there any relationship between all the random variables?
  - the example: gender influences both height and occupation, height influences occupation.
- let us consider the graph probabilistic representation
  - can relations be posed in terms of graphs?
  - in which way to interpret graphs in probabilistic context?
  - still 7 probability values needed, no simplification, only reformulation,
  - why? edges among all the nodes, no use of **(conditional) independence**.



- any joint probability can be calculated (and thus any other probability)
- $Pr(t,m,j) = Pr(m) \times Pr(t|m) \times Pr(j|t,m) =$
  $= 0.5 \times 0.5 \times 0.008 = 0.002$
- $Pr(m,j) = Pr(t,m,j) + Pr(\neg t,m,j) =$
  $= 0.002 + 0.005 = 0.007$
- $Pr(m|j) = \frac{Pr(m,j)}{Pr(j)} = \frac{0.007}{0.01} = 0.7$

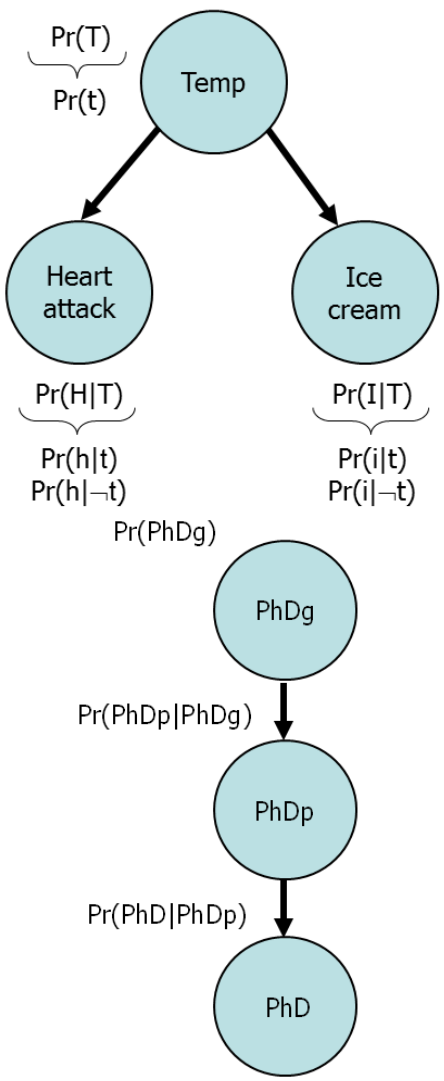
(Conditional) independence

- **definition**: A and B are conditionally independent given C if:
  - $Pr(A, B|C) = Pr(A|C) \times Pr(B|C)$, $\forall A, B, C, Pr(C) \neq 0$
  - denoted as $A \perp\!\!\!\perp B|C$ (conditional dependence $A \not\perp\!\!\!\perp B|C$)
  - (classical independence between A and B: $Pr(A, B) = Pr(A) \times Pr(B)$)
- some observations make other observations uninteresting
  - under assumption of conditional independence it holds:
    $Pr(B|C) = Pr(B|A, C)$ a $Pr(A|C) = Pr(A|B, C)$,
  - observing C, knowledge of A becomes redundant for knowing B,
  - observing C, knowledge of B becomes redundant for knowing A.

(Conditional) independence

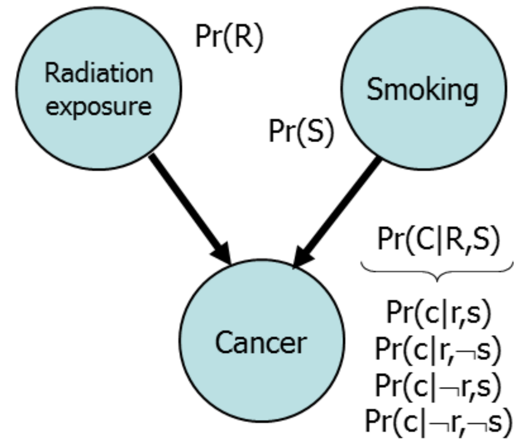
- Example 1:
  - heart attack rate (H) grows with ice cream sales (I),
  - variables H and I are dependent: $Pr(h|i) > Pr(h)$,
  - both grow only because of temperature (T),
  - when conditioned by T, H and I become independent: $Pr(H|I,T) = Pr(H|T)$.

- Example 2:
  - educated grandparents (PhDg) often have educated grandchildren (PhD): $Pr(phd|phdg) > Pr(phd)$
  - knowledge of the parents' education level (PhDp) makes grandparents unimportant: $Pr(PhD|PhDp, PhDg) = Pr(PhD|PhDp)$



(Conditional) independence

- Example 3:
  - both radiation (R) and smoking (S) can cause cancer (C)
  - R and S are obviously independent variables:
    $Pr(R, S) = Pr(R) \times Pr(S)$
  - concerning C, R and S become seemingly dependent!!!
    $Pr(r|s, c) < Pr(r|c)$ or $Pr(r|s, \not c) < Pr(r| \not c)$



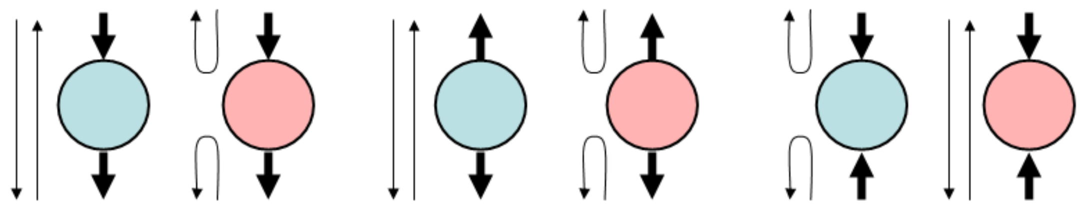
- Summary
  - Ad 1 and 2) conditional independence
    the intermediate variable explains dependency between the ultimate ones,
  - Ad 3) independence
    the intermediate variable introduces spurious dependency.

Connection types

- Nomenclature
  - parent p and child/son c – a directed edge from p to c,
  - ancestor a and descendant d – a directed path from a to d,
- three connection types
  - diverging
    * terminal nodes dependent, dependence disappears when (surely) knowing middle node,
    * intermediate variable (daytime) explains dependence,
    * crime-rate $\leftarrow$ daytime $\rightarrow$ energy consumption (and Ex. 1 – heart attacks).
  - linear (serial)
    * terminal nodes dependent, dependence disappears when (surely) knowing middle node,
    * intermediate variable (branch of study) explains dependence,
    * Simpson's paradox: gender $\rightarrow$ branch of study $\rightarrow$ admission (and Ex. 2 – PhD),
  - converging
    * terminal nodes indep., spurious dependence introduced with knowledge of middle node,
    * temperature $\rightarrow$ ice cream sales $\leftarrow$ salesperson skills (and Ex. 3 – radiation exposure),
- analogy e.g. with partial correlations.

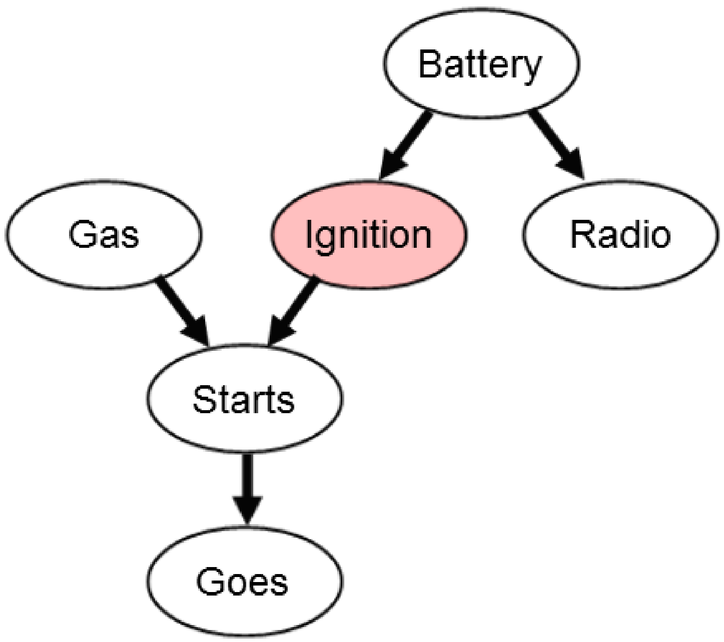
D-separation

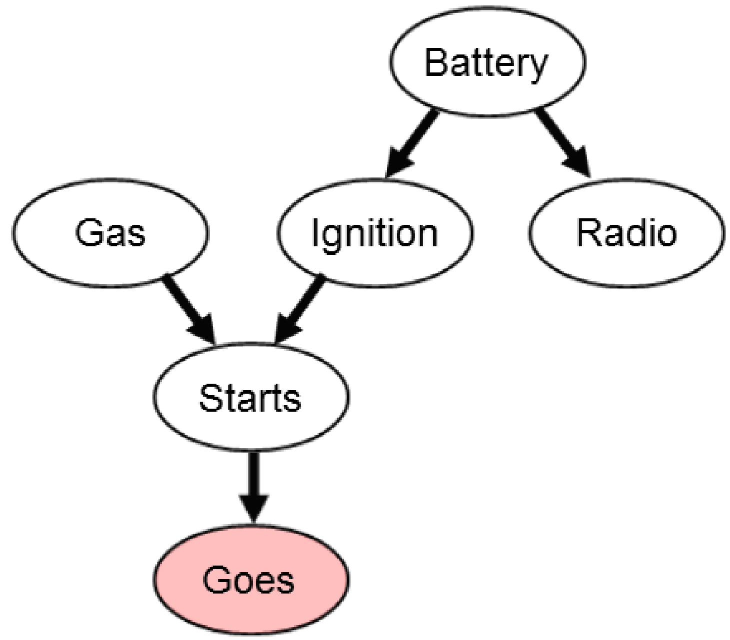
- uses connections to determine conditional independence between sets of nodes
  - linear and diverging connection transmit information **not given** middle node,
  - converging connection transmits information **given** middle node or its descendant,



- two node sets X and Y are d-separated by a node set Z iff
  - all undirected paths between arbitrary node pairs $x \in X$ and $y \in Y$ are blocked
    * there is a linear or diverging node $z \in Z$ on the path, or
    * there is a converging node $w \notin Z$ (none of its descendants w must not be in Z),
- d-separation is equivalent of conditional independence between X and Y given Z,
- a tool of abstraction
  - generalizes from 3 to multiple nodes when studying information flow through a network.

D-separation – example, car diagnosis BN [Russel: AIMA]





- $Gas, Start, Go \perp\!\!\!\perp Bat, Rad | Ign$
- $\{Gas, Start, Go\}$ and $\{Bat, Rad\}$ c.ind
- sets are d-separated
- no open path between any pair of nodes
  - Gas x Battery, Gas x Radio etc.
  - all paths blocked by the middle linear node

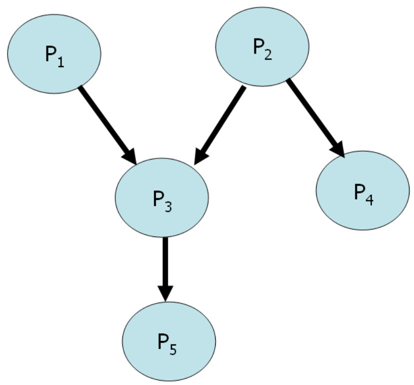
- $Gas \not\perp\!\!\!\perp Ign, Bat, Rad | Go$
- Gas and $\{Ign, Bat, Rad\}$ are c.dependent
- sets are not d-separated
- node $Goes$ opens one path at least
  - Gas connected with $Ignition$ via $Starts$
  - observed descendant of converging node

Graphical probabilistic models

- exploit both probability theory and graph theory,
- graph = qualitative part of model
  - nodes represent events / random variables,
  - edges represent dependencies between them,
  - conditional independence can be seen in graph.
- probability = quantitative part of model
  - local information about node and its neighbors,
  - the strength of dependency, way of inference,
- different graph types (directed/undirected edges, constraints), probability encoding and focus
  - Bayesian networks – causal and probabilistic processes,
  - Markov networks – images, hidden causes,
  - data flows – deterministic computations,
  - influence diagrams – decision processes.

Bayesian networks

- Bayesian or Bayes or belief or causal networks (BNs, CNs),
- What is a Bayesian network?
  - directed acyclic graph – DAG,
  - nodes represent random variables (typically discrete),
  - edges represent direct dependence,
  - nodes annotated by probabilities (tables, distributions)
    - node probability is conditioned by conjunction of all its parent nodes,
    - $Pr(P_{j+1}|P_1, \ldots, P_j) = Pr(P_{j+1}|parents(P_{j+1}))$
    - root nodes annotated by prior distributions,
    - internal nodes conditioned by their parent variables,
    - other (potential) dependencies are ignored,



- Network interpretation?
  - concised representation of probability distribution given conditional independence relations,
  - qualitative constituent = graph,
  - quantitative constituent = a set of conditional probability tables (CPTs).

Bayesian networks

- sacrifice accuracy and completeness – focus on fundamental relationships,
- reduce complexity of representation and subsequent inference,
- full probability model
  - can be deduced by the gradual decomposition (factorization):

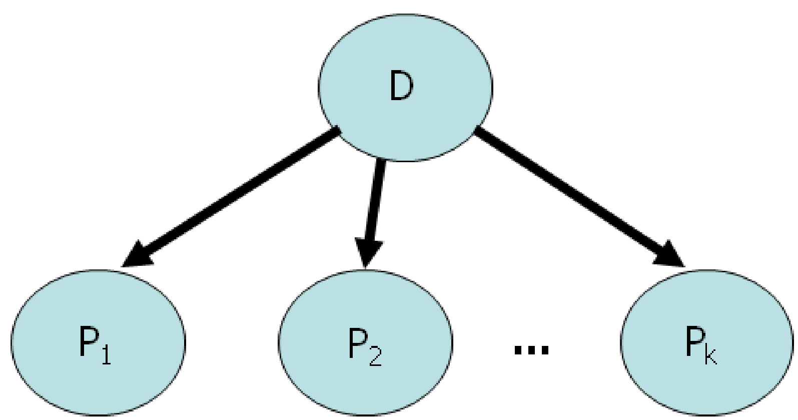
$$
\begin{aligned}
Pr(P_1, P_2, \ldots, P_n) &= Pr(P_1) \times Pr(P_2, \ldots, P_n | P_1) = \\
&= Pr(P_1) \times Pr(P_2|P_1) \times Pr(P_3, \ldots, P_n | P_1, P_2) = \cdots = \\
&= Pr(P_1) \times Pr(P_2|P_1) \times Pr(P_3|P_1, P_2) \times \cdots \times Pr(P_n|P_1, \ldots, P_{n-1})
\end{aligned}
$$

- BNs simplify the model:
  - $Pr(P_1, \ldots, P_n) = Pr(P_1|parents(P_1)) \times \cdots \times Pr(P_n|parents(P_n))$
  - ie. the other (possible) dependencies are ignored,
- ultimate case is **naïve** inference assuming variable independence
  - $Pr(P_1, P_2, \ldots, P_n) = Pr(P_1) \times Pr(P_2) \times \cdots \times Pr(P_n)$
  - uses marginal probs only – linear complexity in the number of variables,
  - used e.g. in classification.

Naïve Bayes classifier

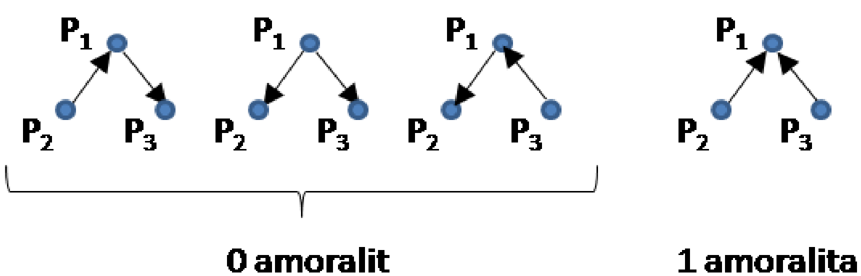
- a special case of Bayesian network
  - based on purely diagnostic reasoning,
  - assumes conditional independence among features $P_1,\ldots, P_k$ given the diagnosis D,
  - the target variable is determined in advance.

$$Pr(D|P_1,\ldots,P_k) = \frac{Pr(P_1,\ldots,P_k|D) \times Pr(D)}{Pr(P_1,\ldots,P_k)}$$
$$Pr(P_1,\ldots,P_k|D) = Pr(P_1|D) \times Pr(P_2|D) \times \cdots \times Pr(P_k|D)$$



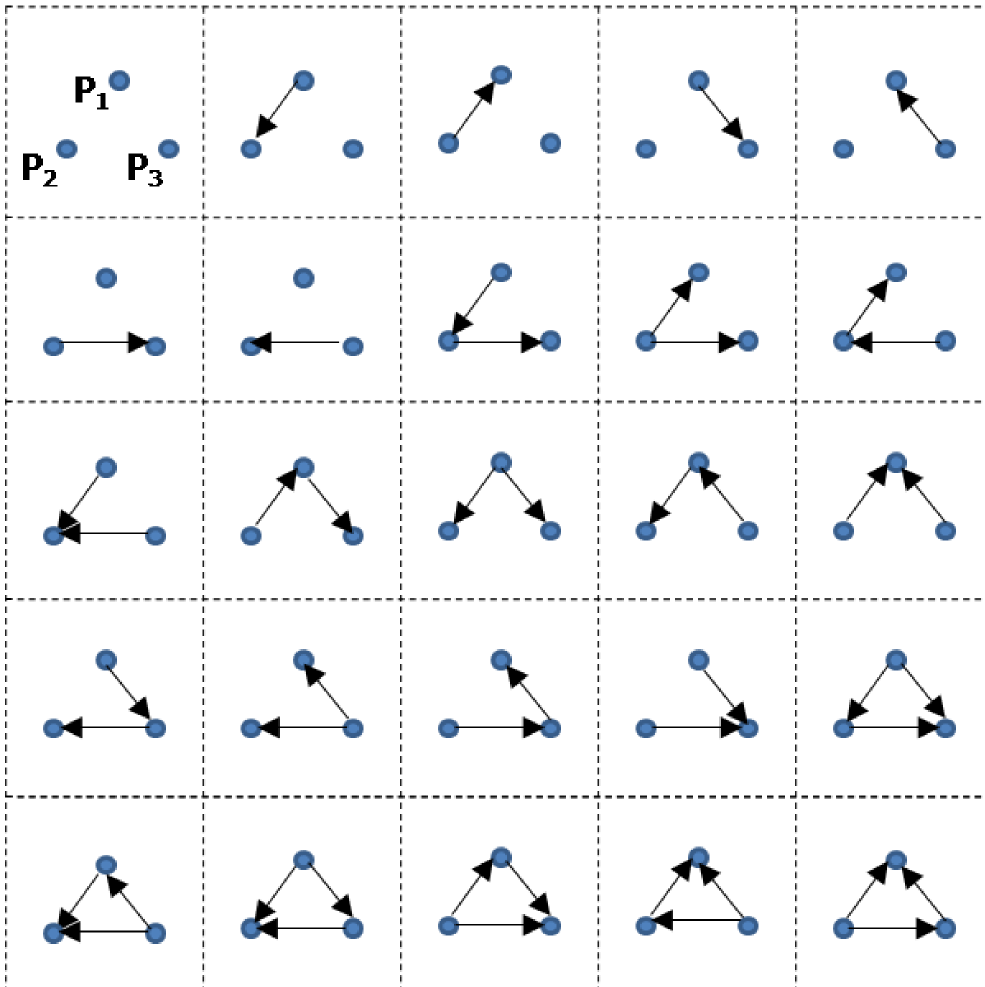
Markov equivalence classes

- DAG classes that define identical conditional independence relationships
  - represent identical joint distribution,
- **Markov equivalence** class is made by directed acyclic graphs which
  - have the identical skeleton
    * fully match when edge directions removed,
  - contain the same set of immoralities
    * immorality = 3 node subgraph such that: $X \rightarrow Z$ and $Y \rightarrow Z$, no XY arc,
    * ie. the graphs have the same sets of uncoupled parents (without an edge between them),
- when learning from data, graphs from a single class are indistinguishable,
- example: 2 distinct equivalence classes (first $P_2 \perp\!\!\!\perp P_3 | P_1$, second $P_2 \perp\!\!\!\perp P_3 | \emptyset$),



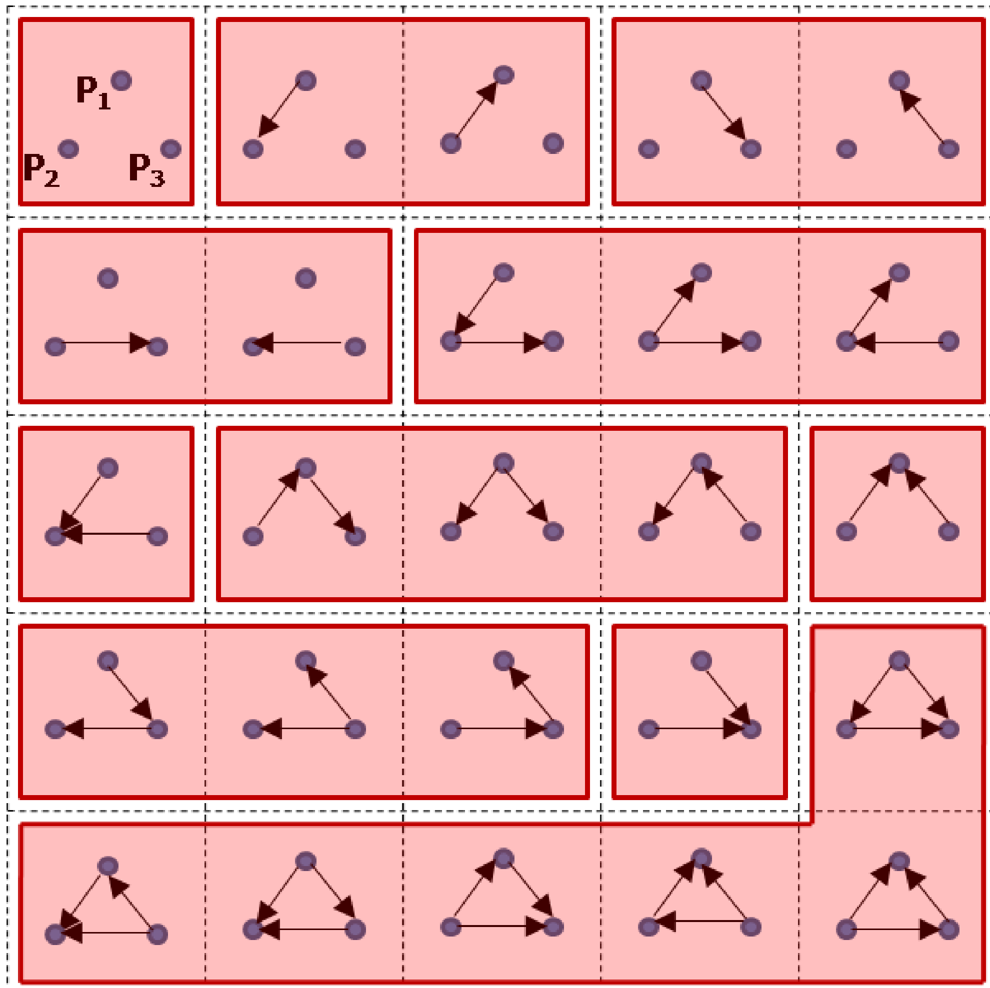
Markov equivalence classes

- let us consider all 25 directed acyclic graphs with 3 **labeled** nodes



Markov equivalence classes

- they make 11 Markov equivalence classes altogether

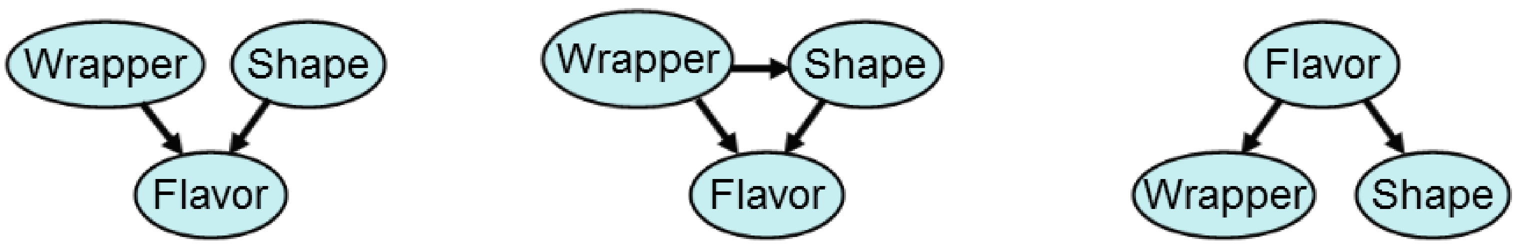


Characteristics of qualitative model

- **correctness**
  - simplification $Pr(P_{j+1}|P_1,\ldots,P_j) = Pr(P_{j+1}|rodice(P_{j+1}))$ complies with reality,
  - each network node is c.ind of its ancestor given its parents,
- **efficiency**
  - there are no redundant edges,
  - actual c.independence relations described by the minimum number of edges,
  - extra edges do not violate correctness,
  - but slow down computations and make the model difficult to read,
- **causality**
  - edge directions agree with actual cause-effect relationships,
- consequences
  - graphs lying in the same Markov equivalence class have the same correctness and efficiency,
  - complete DAG is always correct, however it is very likely inefficient.

Characteristics of qualitative model – example

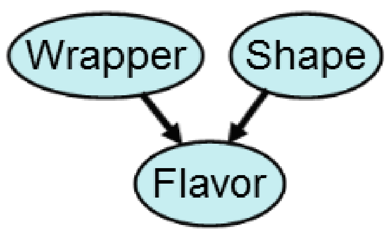
- **The Surprise Candy Company** makes candy in two flavors: 70% are strawberry flavor and 30% are anchovy flavor. Each new piece of candy starts out with a round shape; as it moves along the production line, a machine randomly selects a certain percentage to be trimmed into a square; then, each piece is wrapped in a wrapper whose color is chosen randomly to be red or brown. 80% of the strawberry candies are round and 80% have a red wrapper, while 90% of the anchovy candies are square and 90% have a brown wrapper. All candies are sold individually in sealed, identical, black boxes.



Russell, Norvig: Artificial Intelligence: A Modern Approach.

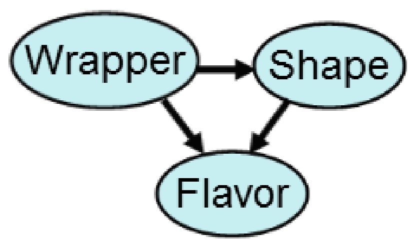
Characteristics of qualitative model – example

- **The Surprise Candy Company** makes candy in two flavors: 70% are strawberry flavor and 30% are anchovy flavor. Each new piece of candy starts out with a round shape; as it moves along the production line, a machine randomly selects a certain percentage to be trimmed into a square; then, each piece is wrapped in a wrapper whose color is chosen randomly to be red or brown. 80% of the strawberry candies are round and 80% have a red wrapper, while 90% of the anchovy candies are square and 90% have a brown wrapper. All candies are sold individually in sealed, identical, black boxes.



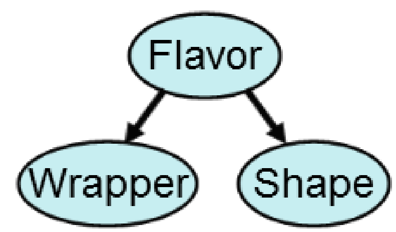
incorrect

- $Wrap \perp\!\!\!\perp Shape | \emptyset$
- contradicts reality.



correct, inefficient

- no independ. relationship,
- thus no unrealistic one.

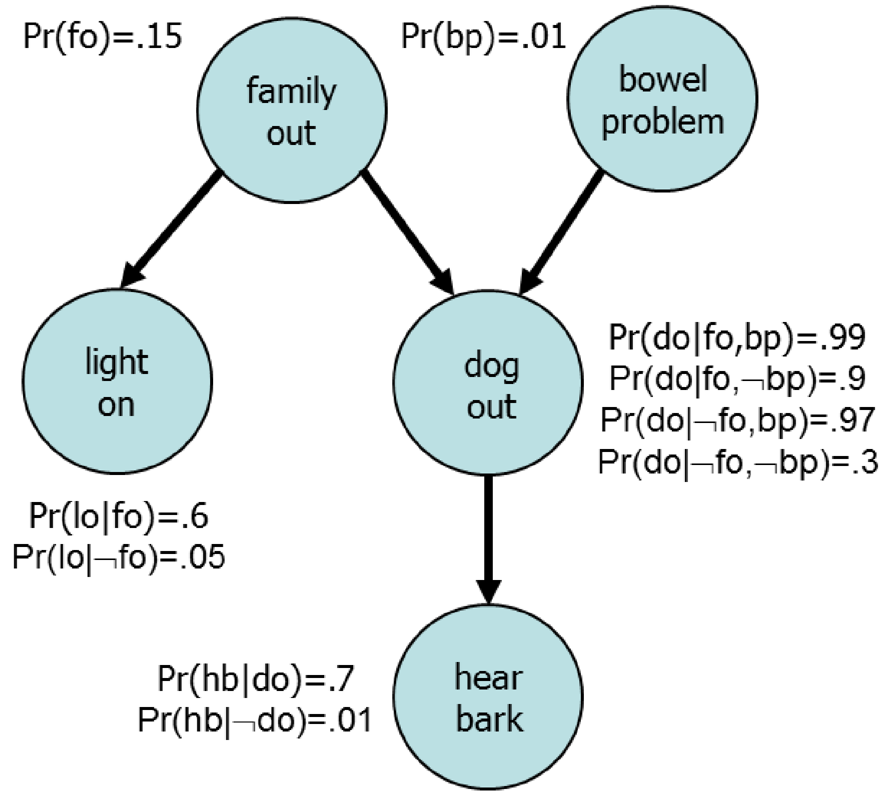


correct, efficient, causal

- $Wrap \perp\!\!\!\perp Shape | Flavor$
- complies with reality.

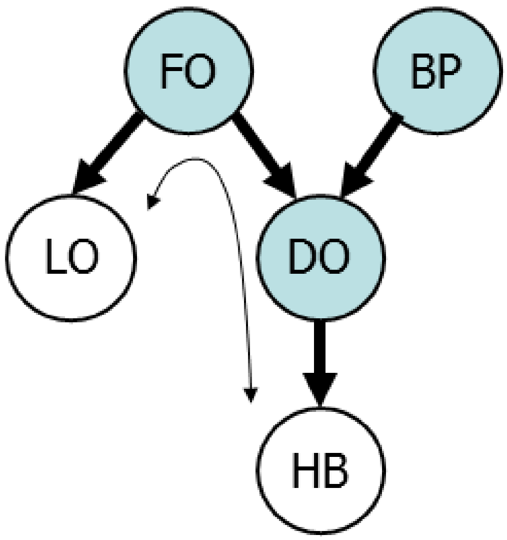
Probability networks – example FAMILY

- Family house and events in it:
  - family sometimes goes out,
  - door light can be on or off,
  - family owns a dog, rarely ill,
  - dog can stay in or out,
  - dog can bark.
- Relationships between events:
  - often switching the light on when leaving,
  - dog is rather out when leaving,
  - dog is out when ill (bowel problem),
  - dog is barking when out,
  - dog can hardly be heard when in.

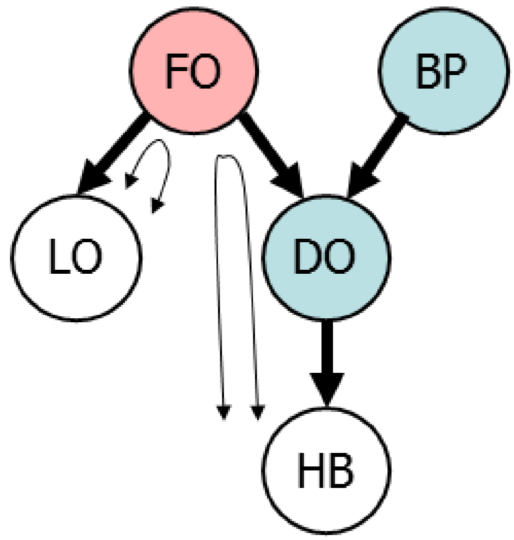


©Charniak: Bayesian Networks withou Tears.

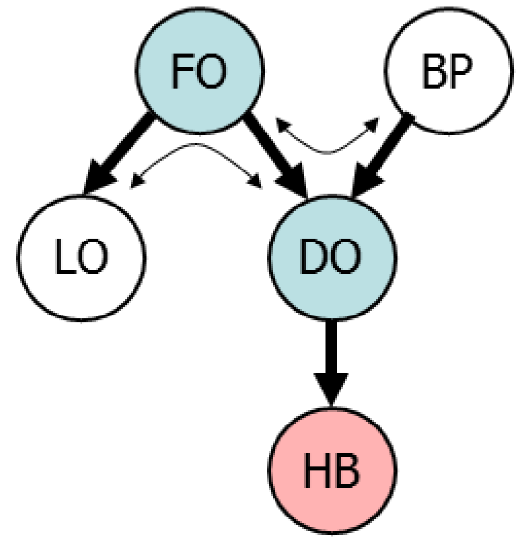
D-separation – examples



- $LO \not\perp\!\!\!\perp HB|\emptyset$
- open path from LO to HB,
- LO and HB not d-separated,
- LO and HB are dependent.



- $LO \perp\!\!\!\perp HB|FO$
- observed FO blocks path
- LO and HB c. independent,
- it also holds
  - $LO \perp\!\!\!\perp HB|DO$
  - $LO \not\perp\!\!\!\perp HB|BP$



- $LO \not\perp\!\!\!\perp BP|HB$
- observed HB opens path
- LO and BP c. dependent,
- it also holds
  - $LO \perp\!\!\!\perp BP|\emptyset$
  - $LO \not\perp\!\!\!\perp BP|DO$

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.

Summary

- probability
  - a rigorous tool for uncertainty modeling,
  - each **atomic** event is described by the joint probability distribution,
  - queries answered by enumeration of atomic events
    - (summing, sometimes with final division),
- needs to be simplified in non-trivial domains
  - reason: curse of dimensionality,
  - solution: independence and conditional independence
  - tool: GPM = graph (quality) + conditional probability tables/functions (quantity).

Recommended reading, lecture resources

- Russell, Norvig: **AI: A Modern Approach**, Uncertain Knowledge and Reasoning (Part V)
  - zejména neurčitost (kap. 14) a probabilistic usuzování (kap. 15),
  - online on Google books: http://books.google.com/books?id=8jZBksh-bUMC,
- Charniak: **Bayesian Networks without Tears**
  - http://ntu.csie.org/~piaip/docs/BayesianNetworksWithoutTears.pdf,
- Murphy: **A Brief Introduction to Graphical Models and Bayesian Networks.**
  - http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html,
- Mooney: **CS 391L: Machine Learning: Bayesian Learning: Beyond Naive Bayes.**
  - http://www.cs.utexas.edu/~mooney/cs391L/slides/bayes2.pdf,
- Bishop: **Pattern Recognition and Machine Learning.**
  - Chapter 8: Graphical models,
  - http://research.microsoft.com/%7Ecmbishop/PRML/Bishop-PRML-sample.pdf.

OPPA European Social Fund

Prague & EU: We invest in your future.