Limitations of First-Order Logic

- FOL is very expressive, but...consider how to translate these:
  - "*most* students graduate in 4 years"
    - $\forall x$ student(x) $\rightarrow$ duration(undergrad(x))$\leq$years(4) (all???)
  - "*only a few* students switch majors"
    - $\exists$s,m1,m2,t1,t2 student(s)^major(s,m1,t1)$\wedge$major(s,m2,t2) $\wedge$m1$\neq$m2 $\wedge$ t1$\neq$t2 (exists???)
  - "all birds can fly, *except* penguins, stuffed birds, plastic birds, birds with broken wings..."

- The problem(s) with FOL involve expressing:
  - default rules & exceptions
  - degrees of truth
  - strength of rules

Practical needs for modeling Uncertainty in KBS

- What happens if you do not know whether an antecedent is T or F?
- neither T nor F, but 'unknown' (not allowed in Boolean logic)
- FOL treats 'unasserted' facts as "could be either T or F" when determining entailment, e.g. { $A \wedge B \rightarrow C$, A } *does not entail* C
- what we often want to do is assume the most likely state (B or $\neg$B) by default
- examples:
  - what if a doctor has to make a diagnosis before white blood cell count is available?
  - or treat a patient even if history of seizures is unknown (because they are unconscious)?

- We will show how to:
  - make default assumptions that are most likely, and derive inferences from them
  - utilize prior and conditional probabilities
  - marginalize over unknowns (which is like weighted averaging by conditional probability of T or F)

Limitation of First-Order Logic

- FOL is not good at handling exceptions
  - universal quantifier means ALL; can't say "most" birds fly
  - $\forall x\ bird(x) \rightarrow flies(x)$
  - asserting $bird(opus) \wedge \neg flies(opus)$ in the KB would cause it to be inconsistent
  - FOL is *monotonic*: if $\alpha \models \beta$, then $\alpha \wedge \omega \models \beta$
    - adding new facts does not undo conclusions
- we could say: $\forall x\ bird(x) \wedge \neg penguin(x) \rightarrow flies(x)$
- but we can't enumerate all possible exceptions
  - what about a robin with a broken wing?
  - what about birds that are made out of plastic?
  - what about Big Bird?




- Uncertainty in reasoning about <u>actions</u>:
  - If a gun is loaded and you pull the trigger, the gun will fire, right?
  - ...unless it is a toy gun
  - ...unless it is defective
  - ...unless it is underwater
  - ...unless the barrel is filled with concrete

Possible Solutions

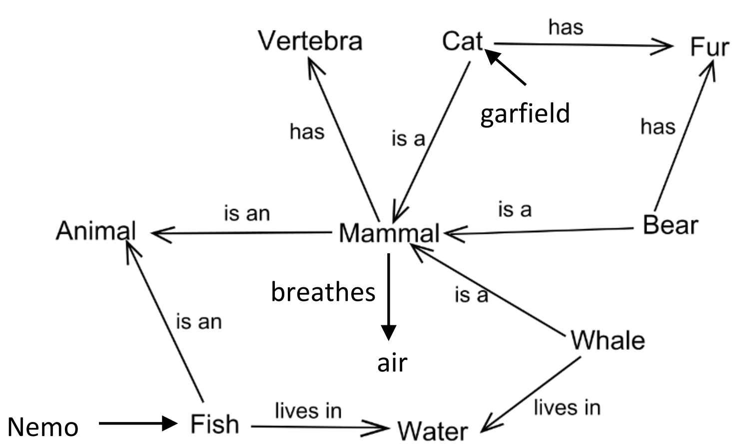
- Add rule *strengths* or *priorities*
  - common in early Expert Systems
  - ...an old ad-hoc approach (with unclear semantics)
  - penguin(x) $\rightarrow_{0.9} \neg$flies(x)
  - bird(x) $\rightarrow_{0.5}$ flies(x)

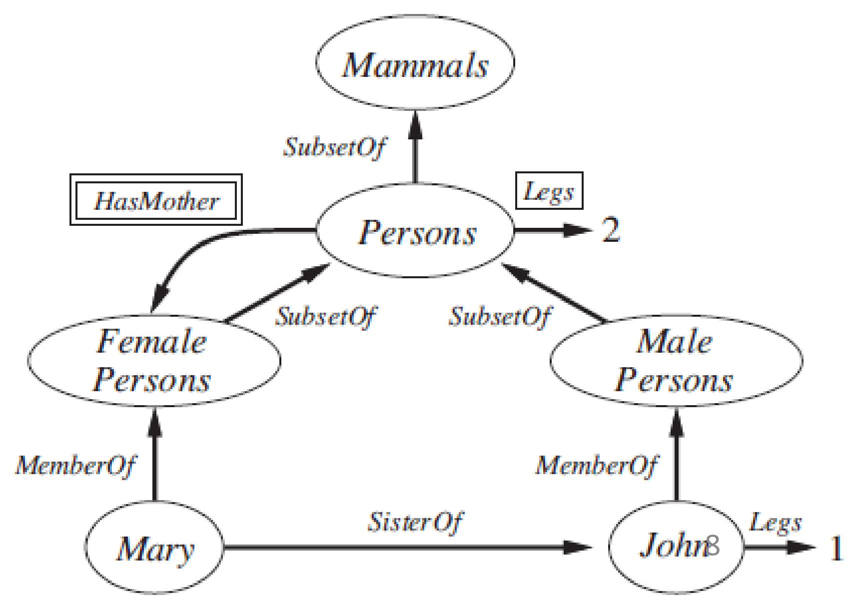
Solutions

- Semantic Networks
- Default Logic/Non-monotonic logics
- Closed-World Assumption and Negation-as-failure in PROLOG
- Fuzzy Logic
- Bayesian Probability

Semantic Networks

- graphical representation of knowledge
- nodes, slots, edges, "isa" links
- procedural mechanism for answering queries
  - follow links to get answers
  - different than formal definition of "entailment"
- <u>inheritance</u>
  - can override defaults





Semantic Networks

- semantic nets are a nice, graphical way of representing information
- an advantage is how the handle default info
- but there are different variations on the graphical symbology and how to express different things (like negation, universal, existence info)
    - difference between thin, thick, and dashed arcs?
    - how to express "safeDrivers are drivers who haven't been in an accident" graphically?
- what does a particular Sem Net formalism *mean???* (semantics of edges, etc)
- try to translate it into a logic. (need more than FOL)

Non-monotonic Logics

- allow retractions later (popular for truth-maintenance systems)
- "birds fly", "penguins are birds that don't fly"
  - $\forall$x bird(x)$\rightarrow$fly(x)
  - $\forall$x penguin(x)$\rightarrow$bird(x), $\forall$x penguin(x)$\rightarrow\neg$fly(x)
  - {bird(tweety), bird(opus)} |= fly(opus)
  - later, add that opus is a penguin, change inference
  - penguin(opus) |= $\neg$fly(opus)
- Definition: A logic is *monotonic* if everything that is entailed by a set of sentences α is entailed by any superset of sentences $\alpha \wedge \beta$
  - opus example is *non-monotonic*

Default Logic

- example syntax of a default rule:
  - bird(x): fly(x) / fly(x)   or   bird(x) ≻ fly(x)
  - analogous to $\forall$x bird(x) → fly(x) , but allows exceptions
  - meaning: "if PREMISE is satisfied and it is not inconsistent to believe CONSEQUENT, then CONSEQUENT"
  - {bird(tweety),bird(opus),¬fly(opus), bird(x): fly(x) / fly(x) } |={fly(tweet),¬fly(opus)}
- requires *fixed-point semantics* (different model theory and inference procedures)

Circumscription

- an alternative approach to default logic
- add *abnormal* predicates to rules
  - $\forall x\ bird(x) \wedge \neg abnormal_1(x) \rightarrow fly(x)$
  - $\forall x\ penguin(x) \wedge \neg abnormal_2(x) \rightarrow bird(x)$
  - $\forall x\ penguin(x) \wedge \neg abnormal_3(x) \rightarrow \neg fly(x)$
- algorithm: minimize the number of *abnormals* needed to make the KB consistent
  - {bird(tweety),fly(tweety),bird(opus),penguin(opus), ¬fly(opus)} is INCONSISTENT
  - {bird(tweety),fly(tweety),bird(opus),penguin(opus), ¬fly(opus), $abnormal_1$(opus)} is CONSISTENT

Probability (Ch. 12)

- an alternative route to encoding default rules like "most birds fly" is to quantify it using probability, $p(fly|bird)=0.95$
- probabilistic reasoning has had a major impact on AI over the years
  - conferences and journals on UAI (Uncertainty in AI)
- probabilistic models has led to major algorithms like:
  - Hidden Markov Models (applications to speech, genomics...)
  - SLAM (simultaneous localization and mapping) for robotics
  - Bayesian networks/graphical models (as knowledge bases)
  - Kalman filters, ICA, POMPDs, ...
  - Reinforcement Learning

Axioms of Probability

- for event e: $0 \leq P(e) \leq 1$
- for mutually exclusive events $e_1..e_n$: $\Sigma_i P(e_i) = 1$
- negation: $P(\neg e) = 1-P(e)$
- Kolmogorov axiom for non-exclusive events:
  $P(a \vee b)=P(a)+P(b)-P(a,b)$

Prior and Conditional Probabilities

- encode knowledge in the form of *prior* probabilities and *conditional* probabilities
  - P(x speaks portugese)=0.012
  - P(x is from Brazil)=0.007
  - P(x speaks portugese|x is from Brazil)=0.9
  - P(x flies|x is a bird)=0.9 (?)

prior probs

conditional probs

- inference is done by calculating *posterior* probabilities given evidence (using Bayes' Rule)
  - compute P(cavity | toothache, flossing, dental history, recent consumption of candy...)
  - compute P(fed will raise interest rate | unemployment=5%, inflation=0.5%, GDP=2%, recent geopolitical events...)

Bayes' Rule

- product rule : *joint prob* P(A,B) = P(A|B)*P(B)
  - P(A|B) is read as "probability of A *given* B"
  - in general, P(A,B)≠P(A)*P(B) (unless A and B are independent)
- Bayes' Rule: convert between causal and diagnostic

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

H = hypothesis (cause, disease)
E = evidence (effect, symptoms)

- joint probabilities: P(E,H), priors: P(H)
- conditional probabilities play role of "rules"
  - people with a toothache are likely to have a cavity
  - p(cavity|toothache) = 0.6

Causal vs. diagnostic knowledge

- *causal*: P(x has a toothache|x has a cavity)=0.9
- *diagnostic*: P(x has a cavity|x has a toothache)=0.6

- typically it is easier to articulate knowledge in the causal direction, but we often want to use it in a diagnostic way to make inferences from observations

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

P($\neg$*cavity* | *toothache*) = ?

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\neg cavity \mid toothache) = P(\neg cavity \wedge toothache) / P(toothache)$$

$$= \frac{0.016+0.064}{(0.108 + 0.012 + 0.016 + 0.064)}$$

$$= 0.4$$

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\neg cavity \mid toothache) = P(\neg cavity \wedge toothache) / P(toothache)$$

$$= \frac{0.016+0.064}{(0.108 + 0.012 + 0.016 + 0.064)}$$

$$= 0.4$$

- useful calculations
  - **marginalization** - summing out unknown variables

$$\mathbf{P}(Cavity) = \mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch) + \mathbf{P}(Cavity, \neg toothache, catch) + \mathbf{P}(Cavity, \neg toothache, \neg catch)$$

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

P(Cav|catch,toothac)=0.108
P(¬Cav|catch,toothac)=0.016
<0.108,0.016>
a = 0.108+0.016=0.124
≅<0.87,0.13>

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

=0.2

  - **normalization** - we can use relative probabilities (and avoid computing the denominator) if we compute prob. *for* and *against* outcome <P,Q>: they must sum to one
    - α represents the proportionality constant
    - α<P,Q> means: α=(P+Q), <P',Q'>=<$\frac{P}{\alpha}$, $\frac{Q}{\alpha}$>, P'+Q'=P/α+Q/α =1

$$\mathbf{P}(Cavity|toothache) = \alpha\mathbf{P}(Cavity, toothache)$$
$$= \alpha[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$
$$= \alpha[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

Conditional Independence

- Applying Bayes' Rule in larger domains has a <u>scalability</u> problem
  - the size of the JPT grows exponentially with the number of variables (2^n for n variables)
- Solution to reduce complexity:
  - employ the <u>Independence Assumption</u>
- Most variables are not strictly independent; most variables are at least partially correlated (but which is cause and which is effect?).
- However, many variables are *conditionally* independent.

A and B are *conditionally independent* <u>given C</u> if:

$P(A,B|C) = P(A|C)P(B|C)$, or equivalently

$P(A|B,C) = P(A|C)$

Conditional Independence

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

- conditional independence gives us an efficient way to combine evidence
  - consider P(Cav|toothache,catch)
  - using Bayes' Rule:
    - P(Cav|toothache,catch) $\propto$ P(toothache^catch|Cav)*P(Cav)
    - this requires a mini JPT for all combinations of evidence
  - assuming toothache is conditionally independent of catch given Cavity:
    - P(toothache^catch|Cav) = P(toothache|Cav)*P(catch|Cav)
    - therefore...

    P(Cav|toothache,catch) $\propto$ P(toothache|Cav)*P(catch|Cav)*P(Cav)

Naive Bayes algorithm

- suppose you have a phenomenon that causes several different effects that could be observed
- Cause $\rightarrow$ Effect_1, Effect_2,..., Effect_n
- each effect is probabilistic, but *assume they are all conditionally independent of each other*
- Then an efficient method for detecting or classifying probable causes is:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- if you have some unobserved vars (y), could marginalize them out, but it leads to same Eqn above

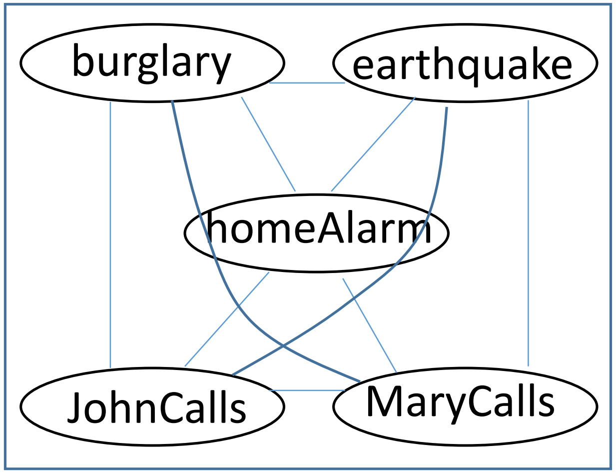
$$\mathbf{P}(Cause \mid \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause, \mathbf{e}, \mathbf{y})$$

- Example: classifying documents as Bag-of-Words
  - P(doctype=sports|words) = P(sports)*(has "score"|sports)*(has "referee"|sports)*...

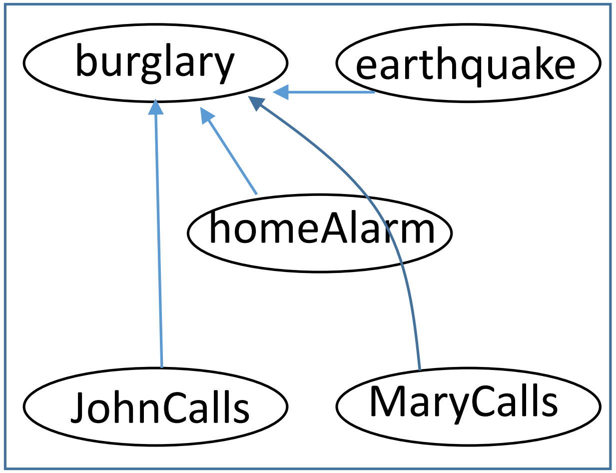
Bayesian Networks (Sec. 13.1 and the first page of Sec 13.2)

- graphical models where *edges represent conditional probabilities*
  - efficient representation because missing edges are assumed to be *conditionally independent* given the nodes in between
- popular for modern AI systems (expert systems)
  - important for handling uncertainty

all vars are correlated, $O(n^2)$ edges, requires full JPT with 2^n rows

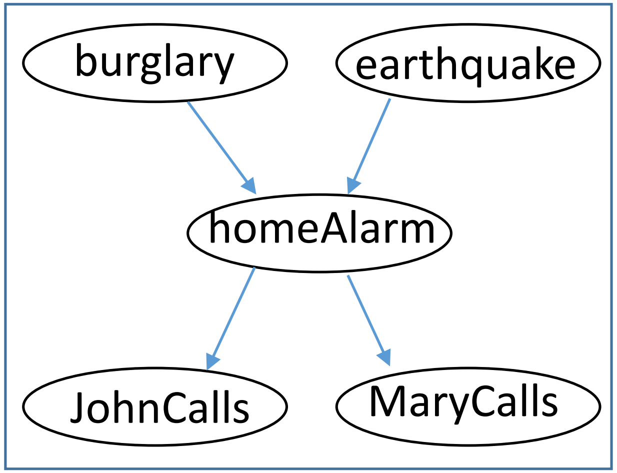


Naive Bayes: compute probability of 1 var depending on all the others (n-1)



requires *independence assumption*

Bayesian Network: selected edges represent conditional dependence

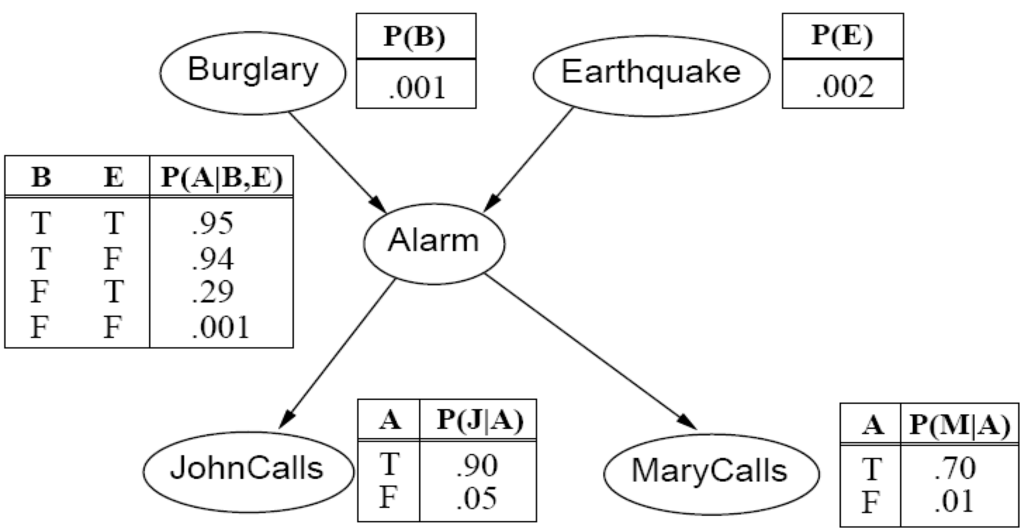


more natural: links follow causality

Bayesian Networks (Sec. 13.1-2)

- prob of each node depends on parents; specify with a mini-JPT
- full JPT has 2^5=32 entries - can answer any query from JPT
- joint prob of full state like <j,m,a,¬b,¬e> is <u>product of prob over all nodes</u>
- prob of <u>each node is conditioned on parents</u>

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

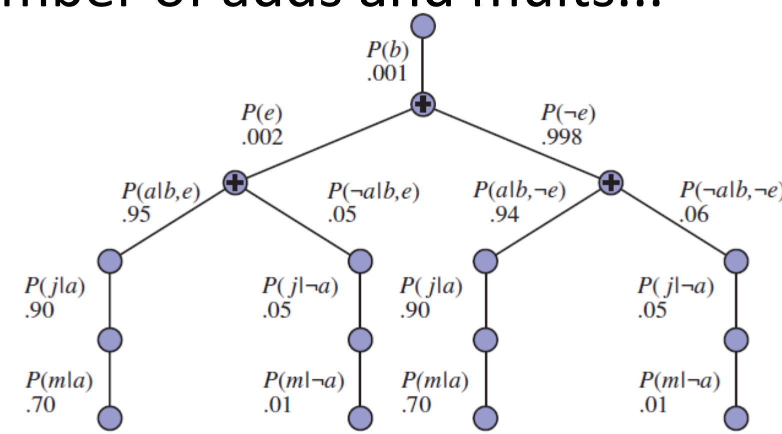
$$\begin{aligned} & P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ = \; & P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ = \; & 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ \approx \; & 0.00063 \end{aligned}$$

- Efficient algorithms for computing inferences or outcomes conditioned on observations/evidence
  - Variable elimination: factor computations into a tree of products and sums (algebraic calculation from formula)
  - rearrange to minimize number of adds and mults...

$$\mathbf{P}(Burglary \mid JohnCalls = true, MaryCalls = true)$$

$$P(b \mid j,m) = \alpha \sum_e \sum_a P(b)P(e)P(a \mid b,e)P(j \mid a)P(m \mid a)$$

$$P(b \mid j,m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b,e)P(j \mid a)P(m \mid a)$$



  - Belief propagation: graph algorithm that updates probs of neighboring nodes when belief of any node changes

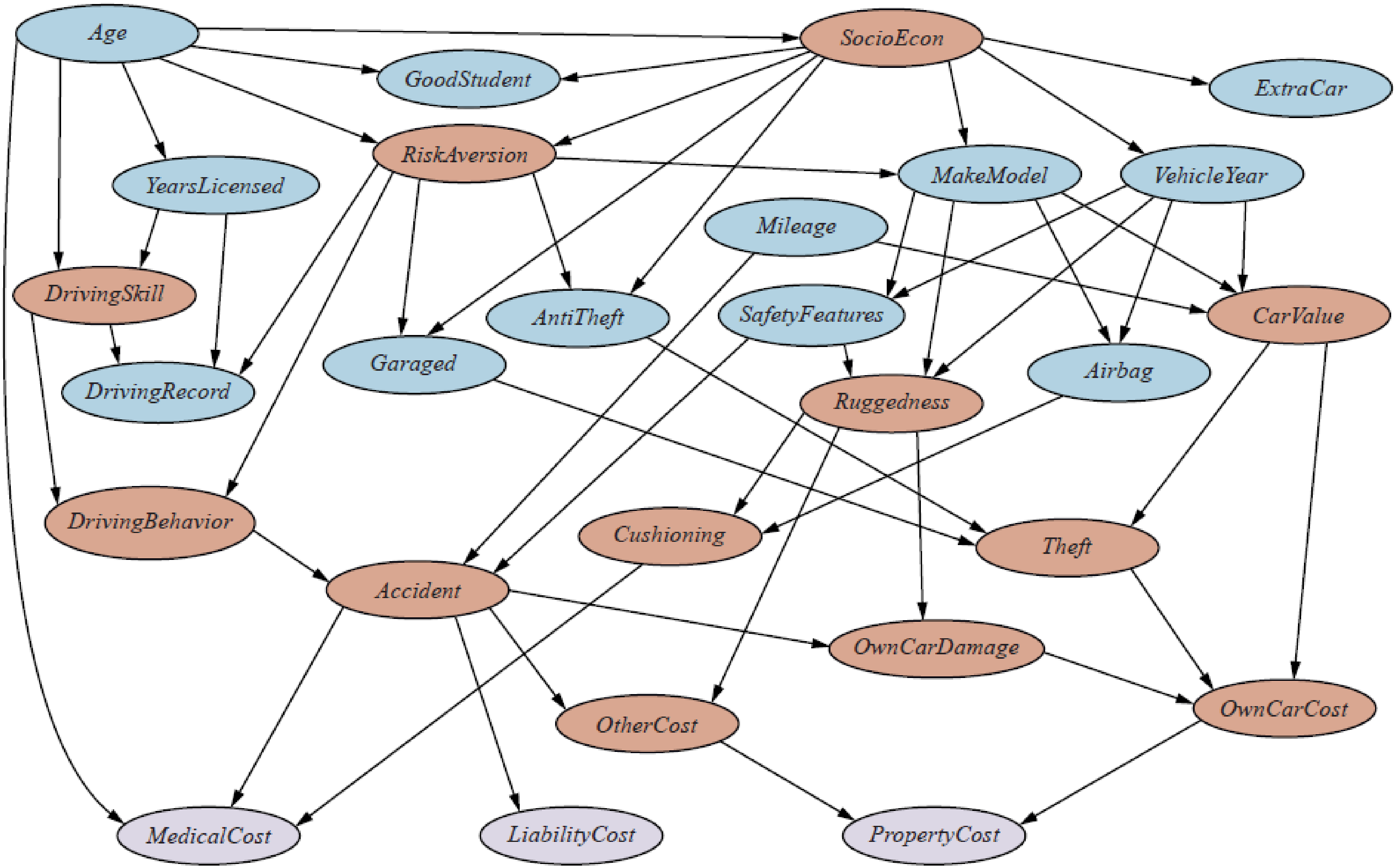
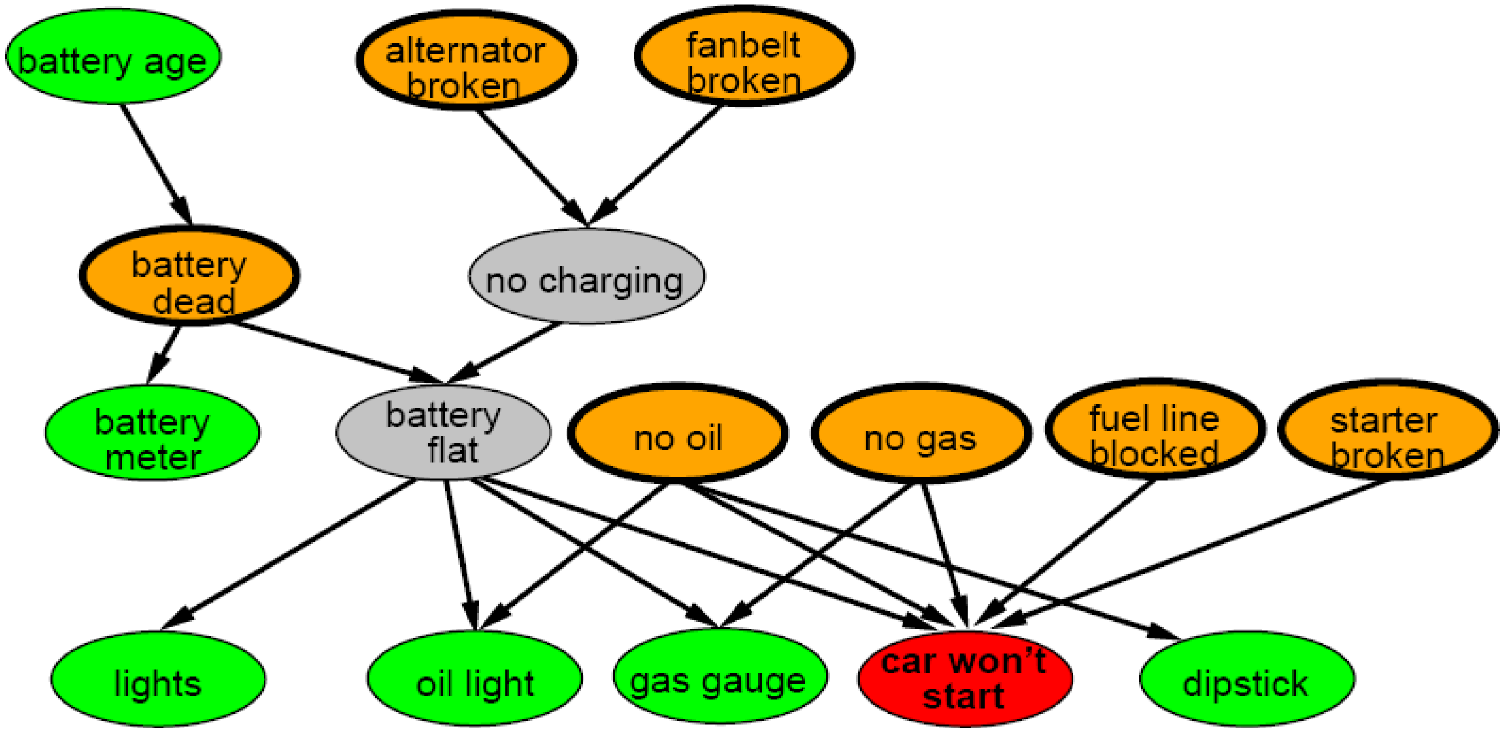


Figure 13.9 A Bayesian network for evaluating car insurance applications.

Initial evidence: car won't start
Testable variables (green), "broken, so fix it" variables (orange)
Hidden variables (gray) ensure sparse structure, reduce parameters



- Many modern knowledge-based systems are based on probabilistic inference
  - including Bayesian networks, Hidden Markov Models, (HMMs), Markov Decision Problems (MDPs)
  - example: Bayesian networks are used for inferring user goals or help needs from actions like mouse clicks in an automated software help system (think 'Clippy')
  - *Decision Theory* combines *utilities* with *probabilities* of outcomes to decide actions to take

- the challenge is capturing all the numbers needed for the prior and conditional probabilities
  - objectivists (frequentists) - probabilities represent outcomes of trials/experiments
  - subjectivists - probabilities are degrees of belief
- probability and statistics is at the core of many Machine Learning algorithms