Decision Making under Uncertainty

- Craig Boutilier; `cebly@cs`; 946-5714; PT398C

- Course details
  - Web page: `~cebly/2534`
  - Tuesdays: 1:00-3:00PM;   Room BA B024

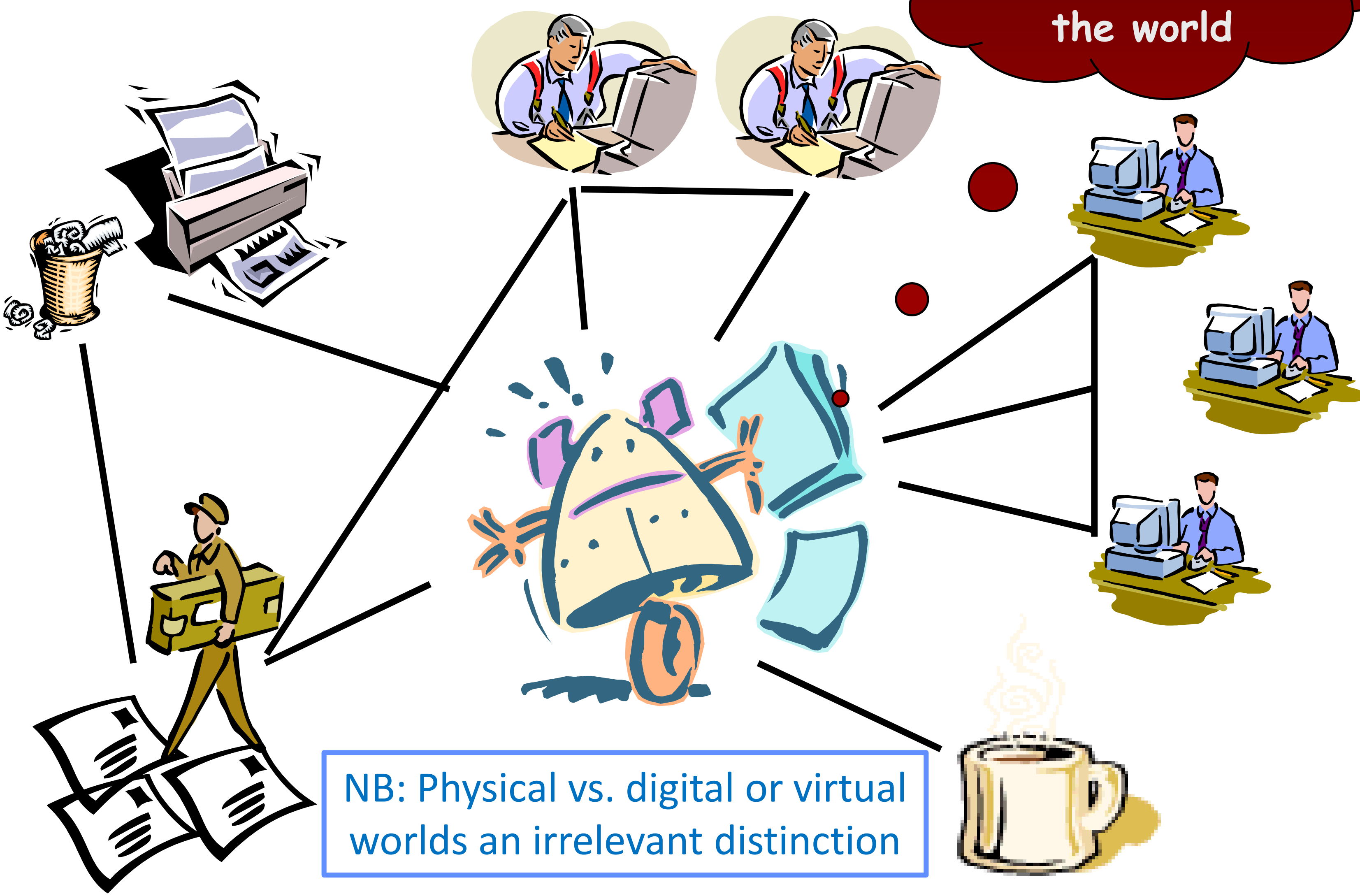
- Evaluation
  - Three assignments: 45%
  - Class participation: 10%
  - Project, incl. proposal, (possibly) presentation: 45%

Rough Overview

- Decision making under uncertainty (DMUU) of all forms
  - one-shot, sequential; single-agent, multi-agent
  - largely probabilistic models of uncertainty
- Main topics
  - *Beliefs: probabilistic inference, computation (Bayes nets)**
  - Single-agent decision making
    - preferences, utilities: foundations, representations, elicitation
    - sequential decision making: MDPs and POMDPs, maybe RL
  - Multiagent decision making
    - basics of game theory, including equilibrium concepts
    - coordination, stochastic games, mechanism design, auctions
    - social choice: voting, stable matchings
- Combination: lectures and readings
  - emphasis on perspective, discussion

A Planning Problem

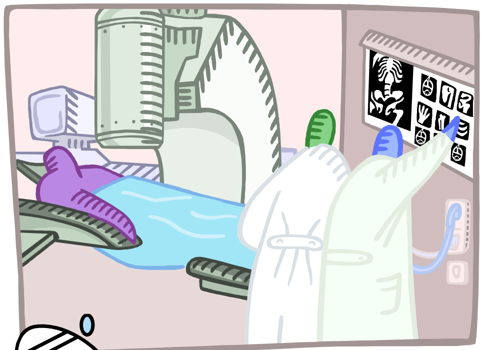
Take actions to bring about changes in the state of the world

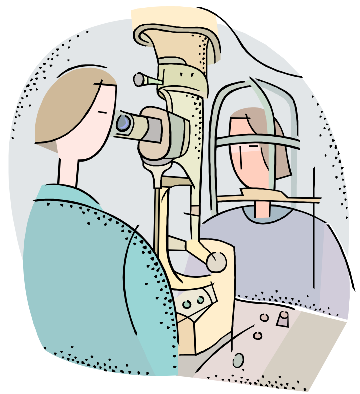


NB: Physical vs. digital or virtual worlds an irrelevant distinction

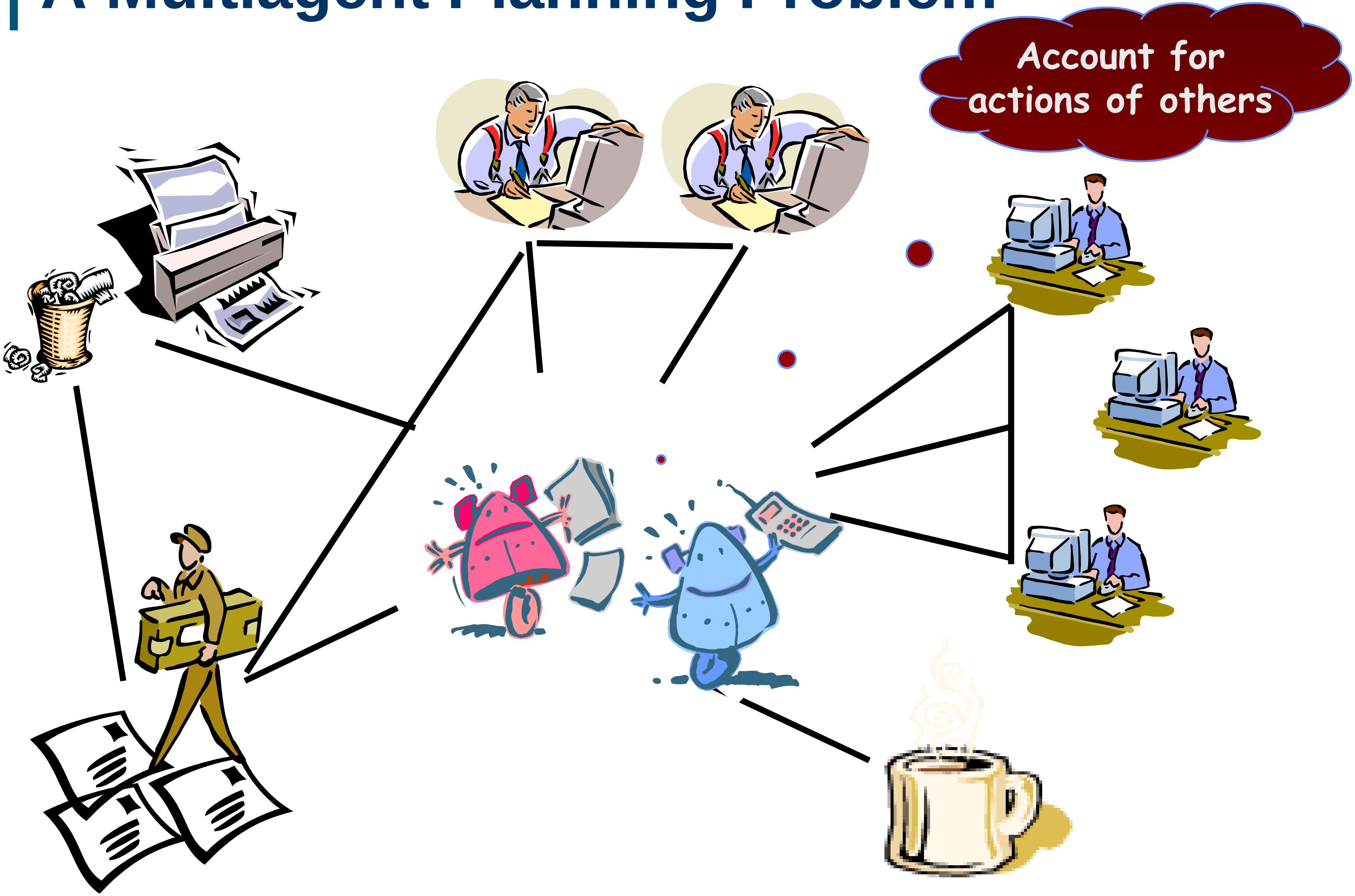
Value/Cost of Information

Take actions to discover state of the world (and make better decisions)





A Multiagent Planning Problem



Lessons of Decision Making

- Robbie’s goal: “tidy lab”
  - classical plan: `goto(lab), kickout(students), pickup(cup17),...`
  - what if I ask for coffee in middle of plan? fire alarm? broken wheel? goes to lab and finds it tidy?
- Lesson #1: appropriate courses of action *contingent* on current state of affairs
  - state can change exogenously (uncertainty)
  - effects of actions can be uncertain (endogenous uncertainty)
  - program structure should be conditional (*policy*, not plan)

Lessons of Decision Making

- Why should Robbie stop tidying when **`coffeereq`**?

- Lesson #2: decisions depend on relative importance of conflicting/competing objectives: preferences
  - **`<coffee@10AM, tidy@11AM>`** preferred to **`<coffee@11AM, tidy@10AM>`**

Lessons of Decision Making

- Whose preferences?
- Lesson #3: decisions should reflect *preferences of user* on whose behalf agent is acting
  - agents act on behalf of users; so “preprogramming” impossible (e.g., shopping agent, medical decision aid (or doctor!), scheduler, bargaining/bidding agent, etc…)
  - *preference elicitation/assessment* needed if agent decides itself
- Consider:
  - price of coffee skyrockets, you like tea almost as much ???
  - Treatment1: faster cure, more expensive/painful;

    Treatment2: much slower, but cheaper/more tolerable

Lessons of Decision Making

- Robbie hears rumor of surprise NSERC Site visit, but Craig unaware: keep tidying or coffee? (If untidy at visit, funding will be cut!)

- Lesson #4: decisions reflect *tradeoffs* between likelihood of outcomes and preferences over them
- Consider:
  - Robbie has \$2: coffee or lottery ticket? highodds lott? coffee \$50?
  - Prob successful coffee delivery: 0.3? 0.1? 0.0001? 0.7? 0.9999?
  - Trtmt1: 0.99 odds of success, \$100,000 vs. Trtmt2: 0.95 and \$5000

Lessons of Decision Making

- I prefer more money to less: so Robbie goes to Starbucks, punches a guy, takes $100, and brings me a coffee and $98!
- Lesson #5: decisions reflect both immediate and *long-term consequences* of actions (and long-term objectives)
- Consider:
  - smoking if prob of lung cancer was 0.17 in *six months* (not 30yrs)?
  - why write an NSERC Grant proposal: some actions enable others

Lessons of Decision Making

- Two robots, I need coffee and Amazon package, each robot equidistant from coffee, mailroom: who does what?
  - one slightly closer to the coffee? one slightly better at coffee delivery? red robot got the coffee yesterday?
- One robot yours, one robot mine: one cup left (lots of tea)
  - both of us like tea (how much)? I hate tea?
- Lesson #6: decisions reflect (anticipated) *behavior of other agents*
  - coordination, cooperation, inherent competition
  - equilibrium (multiple, mixed), side payments/transferable utility
  - elicitation and incentives: mechanism design and social choice

Summary of Key Issues

- Actions change state of the world, enable other actions
- Forms of uncertainty:
  - action effects, exogenous events
  - knowledge of world state
  - behavior of others (different from exogenous events)
- Actions change your state of knowledge:
  - provide info, but not certainty: value of information
- Action effects, preferences not known in advance
  - preference elicitation (more generally preference assessment)
  - learning (especially reinforcement learning)
- Other actors in the world pursuing their own interests
  - cooperative settings: key is coordination of activities
  - competitive (fully, partially) settings: key is strategic/equilibrium effects

In more depth

- **The components rational action.**
- ***Probabilistic semantics for belief.****
  - *Representation of probabilities, Bayesian networks (briefly)*
  - *Inference in Bayes nets (briefly)*
- **Preferences and utilities.**
  - Rational decision-making.
  - Foundations of utility theory.
  - Multi-attribute utility theory.
  - Preference elicitation.
- **Multi-stage decision making.**
  - Markov decision processes.
  - Structured computation for MDPs.
  - Function approximation
  - Partially-observable MDPs.
  - Reinforcement learning (if time/interest)
- **Multiagent DM: Game theory**
  - Basics of game theory.
  - Refinements of Nash equilibria
  - Stochastic and Markov games.
  - Cooperation.
  - Games of incomplete information (Bayesian games)
  - Mechanism design (and computational approaches to MD).
  - Auction theory.
- **Multiagent DM: Social choice**
  - Elements of social choice and voting.
  - Voting rules.
  - Computational considerations.
  - Manipulation and control.
  - Voting with partial information.
  - Matching problems.
  - Other forms of “MD without money".

Probabilistic Inference: Very Brief Review

- As discussed, beliefs about world form a critical component in decision making. And these beliefs should must *quantify* our degree of uncertainty, so appropriate tradeoffs can be made.
- We'll quantify our beliefs using *probabilities*
  - denotes probability that you believe is true
  - we take *subjectivist* viewpoint (cf. *frequentist*)
- Note: statistics/data *influence* degrees of belief
- Let's formalize things just so we're on the same page
  - *This particular perspective will be valuable for decision making, MDPs and POMDPs, Bayesian games, etc.*

Random Variables

- Assume set ***V*** of *random variables*: *X, Y*, etc.
  - Each RV *X* has a *domain* of values *Dom(X)*
  - *X* can take on any value from *Dom(X)*
  - Assume **V** and *Dom(X)* finite
- Examples (finite)
  - $Dom(X) = \{x_1, x_2, x_3\}$
  - *Dom(Weather) = {sunny, cloudy, rainy}*
  - *Dom(Stdnts) = {tyler, joanna, xin, amirali, joel, victoria, andrew}*
  - *Dom(CraigHasCoffee) = {T,F}* (boolean var)

Random Variables/Possible Worlds

- A *formula* is a logical combination of variable assignments:
  - $X = x_1$; $(X = x_2 \vee X = x_3) \wedge Y = y_2$; $(x_2 \vee x_3) \wedge y_2$
  - $chc \wedge \sim cm$, etc…
  - let $\mathcal{L}$ denote the set of formulae (our language)
- A *possible world* (or a *state*) is an assignment of values to each variable
  - these are analogous to truth assts (models)
  - Let W be the set of worlds

Probability Distributions

- A probability distribution $Pr : \mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq Pr(\alpha) \leq 1$
  - $Pr(\alpha) = Pr(\beta)$ if α is logically equivalent to β
  - $Pr(\alpha) = 1$ if α is a tautology
  - $Pr(\alpha \vee \beta) = Pr(\alpha) + Pr(\beta) - Pr(\alpha \wedge \beta)$

- *Pr(α)* denotes our *degree of belief* in *α*; e.g.
  - $Pr(X = x_1) = Pr(x_1) = 0.9$
  - $Pr((x_2 \wedge x_3) \vee y_2) = 0.9$
  - $Pr(loc(craig) = off) = 0.6$
  - $Pr(\ loc(craig) = off \vee loc(craig) = lab\ ) = 1.0$
  - $Pr(loc(craig) = lounge) = 0.0$

Semantics of Prob. Distributions

- A probability measure *μ: W* → [0,1] s.t.

$$\sum_{w \in W} \mu(w) = 1$$

- Intuitively, *μ(w)* measures the probability that the actual world is *w* (your *belief* in *w*). Thus, the relative likelihood of any world you consider possible is specified. If *w* has measure 0, you consider it to be impossible!

- *Our focus is on discrete joint distributions, but analogous concepts apply to continuous (and mixed): use density functions (reflecting "relative" likelihood), CDFs, integrals over measurable sets, etc.*

Semantics of Distributions

- Given measure μ, we can readily determine degree of belief in formula *Pr(α)*
  - simply sum the measures of all worlds satisfying the formula of interest

$$\Pr(\alpha) = \sum_{w \in W} \{\mu(w) : w \models \alpha\}$$

Toy Example Distribution

T – Fedex truck outside
P – purchase from Amazon waiting
C – craig wants coffee
A – craig is angry

$t\ c\ p\ a$	0.162	$\bar{t}\ c\ p\ a$	0.0
$t\ c\ p\ \bar{a}$	0.018	$\bar{t}\ c\ p\ \bar{a}$	0.0
$t\ c\ \bar{p}\ a$	0.016	$\bar{t}\ c\ \bar{p}\ a$	0.0
$t\ c\ \bar{p}\ \bar{a}$	0.004	$\bar{t}\ c\ \bar{p}\ \bar{a}$	0.0
$t\ \bar{c}\ p\ a$	0.432	$\bar{t}\ \bar{c}\ p\ a$	0.0
$t\ \bar{c}\ p\ \bar{a}$	0.288	$\bar{t}\ \bar{c}\ p\ \bar{a}$	0.0
$t\ \bar{c}\ \bar{p}\ a$	0.008	$\bar{t}\ \bar{c}\ \bar{p}\ a$	0.0
$t\ \bar{c}\ \bar{p}\ \bar{a}$	0.072	$\bar{t}\ \bar{c}\ \bar{p}\ \bar{a}$	0.0

Pr(t) =1
Pr(-t) = 0
Pr(c) = .2
Pr(-c) = .8
Pr(p) = .9
Pr(a) = .618
Pr(c & p) = .18
Pr(c v p) = .92
Pr(a -> p)
= Pr(-a v p)
= 1 – Pr(a & -p)
= .976

Exercise: figure out the graphical model/Bayes net used to generate this joint distribution (*can't construct *all* terms)

Relationship

- For any measure μ the induced mapping *Pr* is a distribution.
- For any distribution *Pr* there is a corresponding measure μ that induces *Pr*.
- Thus, the syntactic and semantic restrictions correspond (soundness and completeness)

Some Important Properties

- $Pr(\alpha) = 1 - Pr(-\alpha)$, α can be a "generalized" formula
- $$\sum \{\Pr(x) : x \in Dom(X)\} = 1$$
  - "marginal over X" : $<P(x_1), P(x_2), \ldots, P(x_n)>$
- $Pr(\alpha \vee \beta) = 1$ if $\alpha \supset -\beta$
- $$\Pr(x) = \sum_{y \in Dom(Y)} \Pr(x \wedge y)$$
  - this is called the *summing out* property: holds for sets *Y* as well
  - e.g., $Pr(a) = Pr(a \;\&\; p) + Pr(a \;\&\; -p)$

Conditional Probability

- Conditional probability critical in inference

$$\Pr(b \mid a) = \frac{\Pr(b \wedge a)}{\Pr(a)}$$

  - if *Pr(a) = 0*, we often treat *Pr(b|a)=1* by convention

Semantics of Conditional Probability

- Semantics of *Pr(b|a)*:
  - denotes relative weight/measure of *b*-worlds among *a*-worlds
  - ~*a*-worlds play no role

$$\Pr(b \mid a) = \frac{\sum \{\mu(w) : w \models a \wedge b\}}{\sum \{\mu(w) : w \models a\}}$$

Intuitive Meaning of Conditional Prob.

- Intuitively, if you learned *a*, you would change your degree of belief in *b* from *Pr(b)* to *Pr(b|a)*
- In our example:
  - *Pr(p|c) = 0.9*
  - *Pr(p| ~c) = 0.9*
  - *Pr(a) = 0.618*
  - *Pr(a|~p) = 0.27*
  - *Pr(a|~p & c) = 0.8*
- Notice the *nonmonotonicity* in the last three cases when additional evidence is added
  - contrast this with logical inference

Some Important Properties

- **Product Rule:** *Pr(ab) = Pr(a|b)Pr(b)*

- **Summing Out Rule:**

$$\Pr(a) = \sum_{b \in Dom(B)} \Pr(a \mid b)\Pr(b)$$

- **Chain Rule:**

  Pr(abcd) = Pr(a|bcd)Pr(b|cd)Pr(c|d)Pr(d)

  - holds for any number of variables

Bayes Rule

- **Bayes Rule:**

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\Pr(a)}{\Pr(b)}$$

or $$\Pr(a \mid b) \propto \Pr(b \mid a)\Pr(a)$$

- Bayes rule follows by simple algebraic manipulation of the definition of conditional probability
  - why is it so important? why significant?
  - usually, one “direction” easier to assess than other

Example of Use of Bayes Rule

- *Disease* $\in$ *{malaria, cold, flu}; Symptom = fever*
  - Must compute *Pr(D|fever)* to prescribe treatment
- Why not assess this quantity directly?
  - *Pr(mal | fever)* is not natural to assess; *Pr(fever | mal)* reflects the underlying "causal" mechanism
  - *Pr(mal | fever)* is not "stable": a malaria epidemic changes this quantity (for example)
- So we use Bayes rule:
  - *Pr(mal | fever) = Pr(fever | mal) Pr(mal) / Pr(fever)*
  - note that *Pr(fe) = Pr(fe|m)Pr(m)* + *Pr(fe|c)Pr(c)* + *Pr(fe|fl)Pr(fl)*
  - so if we compute *Pr* of each disease given fever using Bayes rule, normalizing constant is "free"

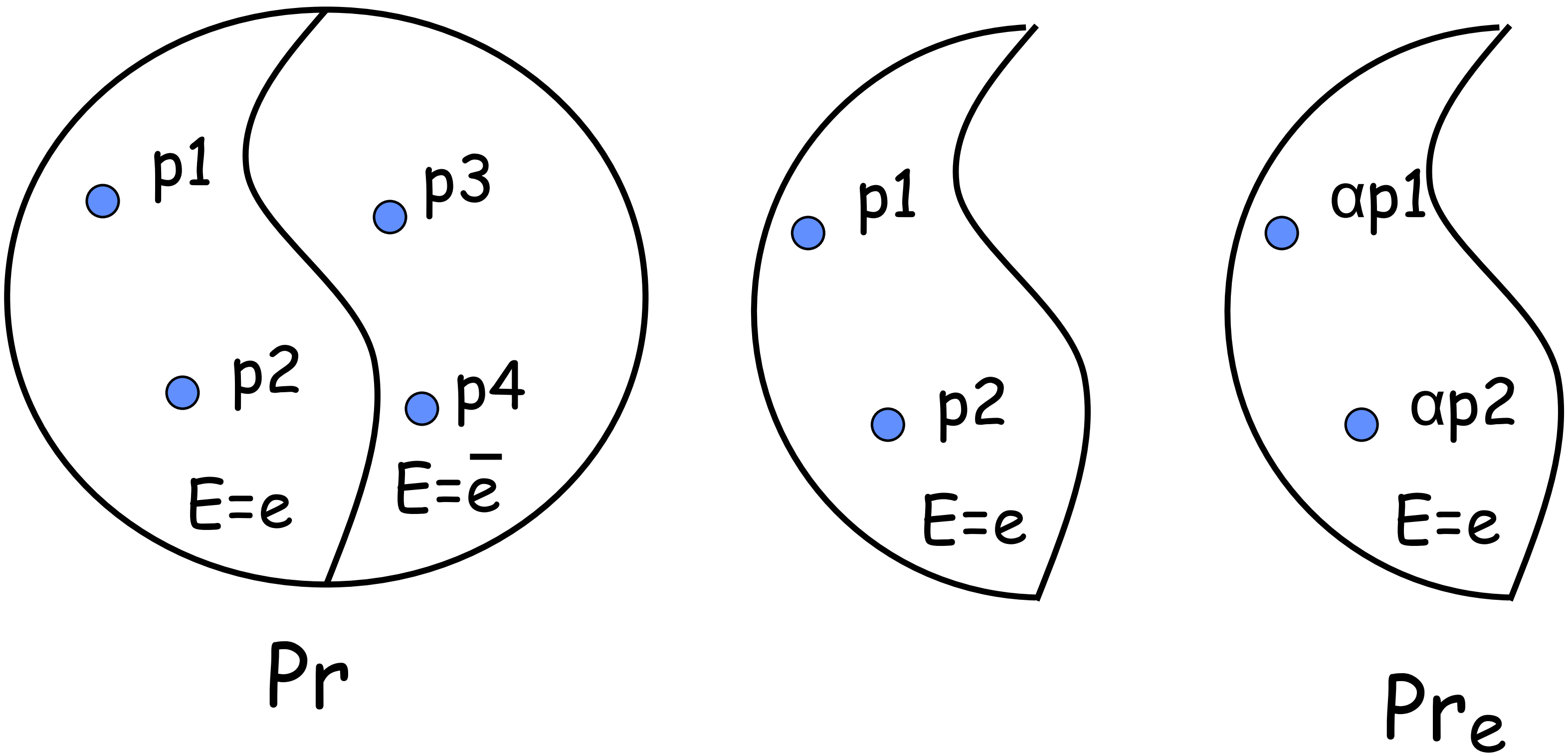
Probabilistic Inference

- By probabilistic inference, we mean
  - given a *prior* distribution *Pr* over variables of interest, representing degrees of belief
  - and given new evidence $E=e$ for some var E
  - Revise your degrees of belief: *posterior* Pr_e
  - *(Many other forms of "inference"/types of queries)*
- How do your degrees of belief change as a result of learning $E=e$ (or more generally $\boldsymbol{E}=\mathbf{e}$, for set $\boldsymbol{E}$)

Conditioning

- We define $Pr_e(\alpha) = Pr(\alpha / e)$
- That is, we produce Pr_e by *conditioning* the prior distribution on the observed evidence e
- Semantically, we take original measure μ
  - we set $\mu(w) = 0$ for any world falsifying e
  - we set $\mu(w) = \mu(w) / Pr(e)$ for any e-world
  - last step known as normalization (ensures that the new measure sums to 1)

Semantics of Conditioning



α = 1/(p1+p2)
normalizing constant

Inference: Computational Bottleneck

- Semantically/conceptually, picture is clear; but several issues must be addressed
- Issue 1: How do we specify the full joint distribution over $X_1, X_2, \ldots, X_n$?
  - *exponential* number of possible worlds
    - e.g., if the X_i are boolean, then 2^n numbers (or $2^n -1$ parameters/degrees of freedom, since they sum to 1)
  - these numbers are *not robust/stable*
  - these numbers are *not natural* to assess (what is probability that "Craig wants coffee; it's raining in Orangeville; robot charge level is low; …"?)

Inference: Computational Bottleneck

- Issue 2: Inference in this representation frightfully slow
  - Must sum over exponential number of worlds to answer query $Pr(\alpha)$ or to condition on evidence e to determine $Pr_e(\alpha)$
- How do we avoid these two problems?
  - no solution in general
  - but in practice there is *structure* we can exploit
- We'll use *conditional independence*

Independence

- Recall that *x* and *y* are *independent* iff:
  - $Pr(x) = Pr(x|y)$ iff $Pr(y) = Pr(y|x)$ iff $Pr(xy) = Pr(x)Pr(y)$
  - intuitively, learning *y* doesn't influence beliefs about *x*

- We say *x* and *y* are *conditionally independent given z* iff:
  - $Pr(x|z) = Pr(x|yz)$ iff $Pr(y|z) = Pr(y|xz)$ iff
    $Pr(xy|z) = Pr(x|z)Pr(y|z)$ iff …
  - intuitively, learning *y* doesn't influence your beliefs about *x* *if you already know z*
  - e.g., learning someone's mark on an exam can influence the probability you assign to a specific GPA; but if you already knew **final** class grade, learning the exam mark would *not* influence your GPA assessment

Variable Independence

- Two *variables* X and Y are conditionally independent given variable Z iff x, y are conditionally independent given z for all $x \in Dom(X)$, $y \in Dom(Y)$, $z \in Dom(Z)$
  - Also applies to sets of variables ***X, Y, Z***
  - Also to unconditional case (X, Y independent)

- If you know the value of Z (*whatever* it is), nothing you learn about Y will influence your beliefs about X
  - these definitions differ from earlier ones (which talk about *events*, not variables)

What does independence buys us?

- Suppose (say, boolean) variables $X_1, X_2, \dots, X_n$ are mutually independent
  - we can specify full joint distribution using only n parameters (linear) instead of $2^n - 1$ (exponential)
- How? Simply specify $Pr(X_1), \dots Pr(X_n)$
  - from this I can recover probability of any world or any (conjunctive) query easily
  - e.g. $Pr(x_1 {\sim} x_2 x_3 x_4) = Pr(x_1)\ (1-Pr(x_2))\ Pr(x_3)\ Pr(x_4)$
  - we can condition on observed value $X_k = x_k$ trivially by changing $\Pr(x_k)$ to 1, leaving $\Pr(x_i)$ untouched for $i \neq k$

The Value of Independence

- Complete independence reduces both *representation of joint* and *inference* from $O(2^n)$ to $O(n)$: pretty significant!

- Unfortunately, such complete mutual independence is very rare. Most realistic domains do not exhibit this property.

- Fortunately, most domains do exhibit a fair amount of *conditional* independence. And we can exploit conditional independence for representation and inference as well.

- **Bayesian networks** do just this

An Aside on Notation

- *Pr(X)* for variable *X* (or set of variables) refers to the *(marginal) distribution* over *X*. *Pr(X|Y)* refers to family of conditional distributions over *X*, one for each $y \in Dom(Y)$.

- Distinguish between *Pr(X)* – which is a distribution – and $Pr(x_i)$ – which is a number. Think of *Pr(X)* as a function that accepts any $x_i \in Dom(X)$ as an argument and returns $Pr(x_i)$.

- Similarly, think of *Pr(X|Y)* as a function that accepts any x_i and y_k and returns $Pr(x_i \mid y_k)$. Note that *Pr(X|Y)* is not a single distribution; rather it denotes the family of distributions (over *X*) induced by the different $y_k \in Dom(Y)$

Exploiting Conditional Independence

- Let's see what conditional independence buys us
- Consider a story:
  - If Craig woke up too early E, Craig probably needs coffee C; if C, Craig needs coffee, he's likely angry A. If A, there is an increased chance of an aneurysm (burst blood vessel) B. If B, Craig is quite likely to be hospitalized H.

E – Craig woke too early   A – Craig is angry   H – Craig hospitalized
C – Craig needs coffee   B – Craig burst a blood vessel

Cond'l Independence in our Story

- If you learned any of *E, C, A,* or *B*, your assessment of *Pr(H)* would change.
  - e.g., if any of these are seen to be true, you would increase *Pr(h)* and decrease *Pr(~h).*
  - So *H* is *not independent* of *E*, or *C*, or *A*, or *B*.
- But if you knew value of *B* (true or false), learning value of *E*, *C,* or *A*, would not influence *Pr(H)*. Influence these factors have on *H* is mediated by their influence on *B*.
  - Craig doesn't get sent to the hospital because he's angry, he gets sent because he's had an aneurysm.
  - So *H* is *independent* of *E*, and *C*, and *A*, *given B*

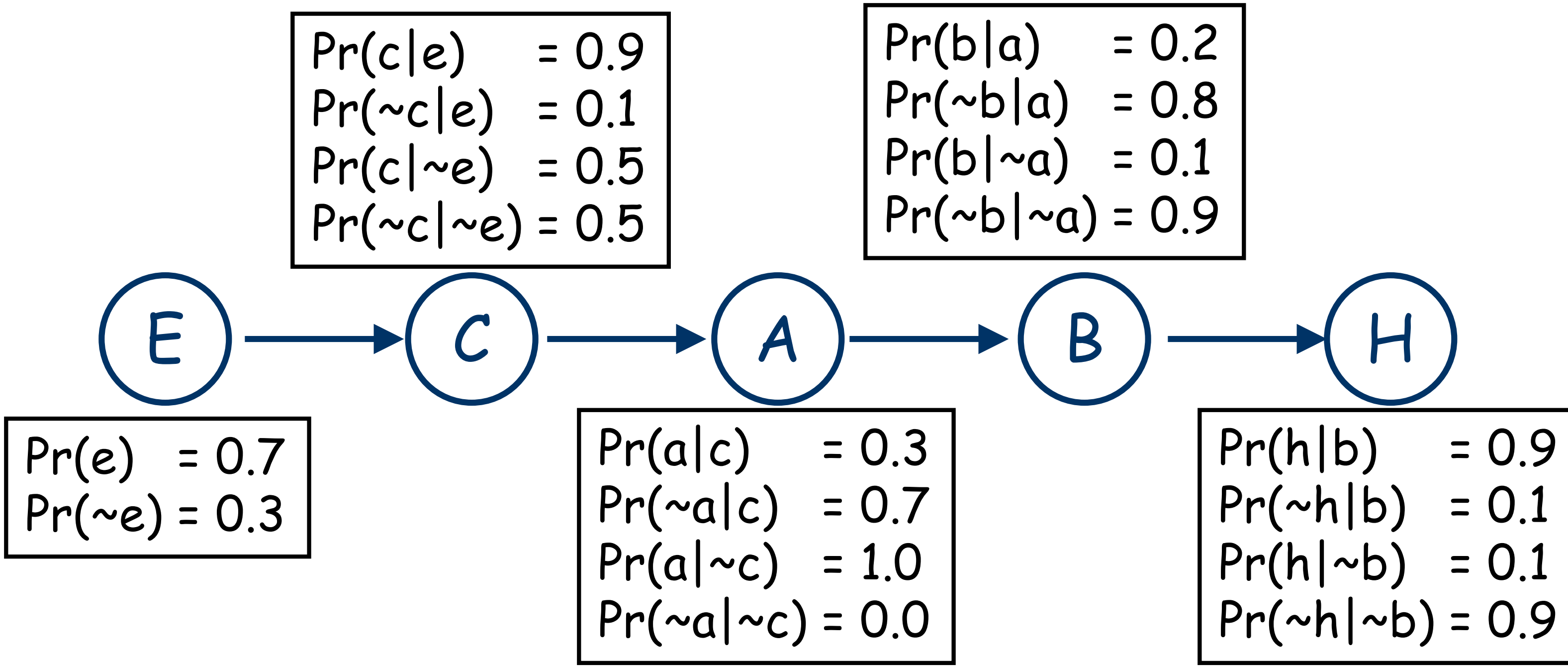
Cond'l Independence in our Story

- So *H* is *independent* of *E*, and *C*, and *A*, *given B*
- Similarly:
  - *B* is *independent* of *E*, and *C*, *given A*
  - *A* is *independent* of *E*, *given C*
- This means that:
  - *Pr(H | B, {A,C,E}) = Pr(H|B)*
    - i.e., for any subset of *{A,C,E}*, this relation holds
  - *Pr(B | A, {C,E}) = Pr(B | A)*
  - *Pr(A | C, {E}) = Pr(A | C)*
  - *Pr(C | E)* and *Pr(E)* don't "simplify"

Cond'l Independence in our Story

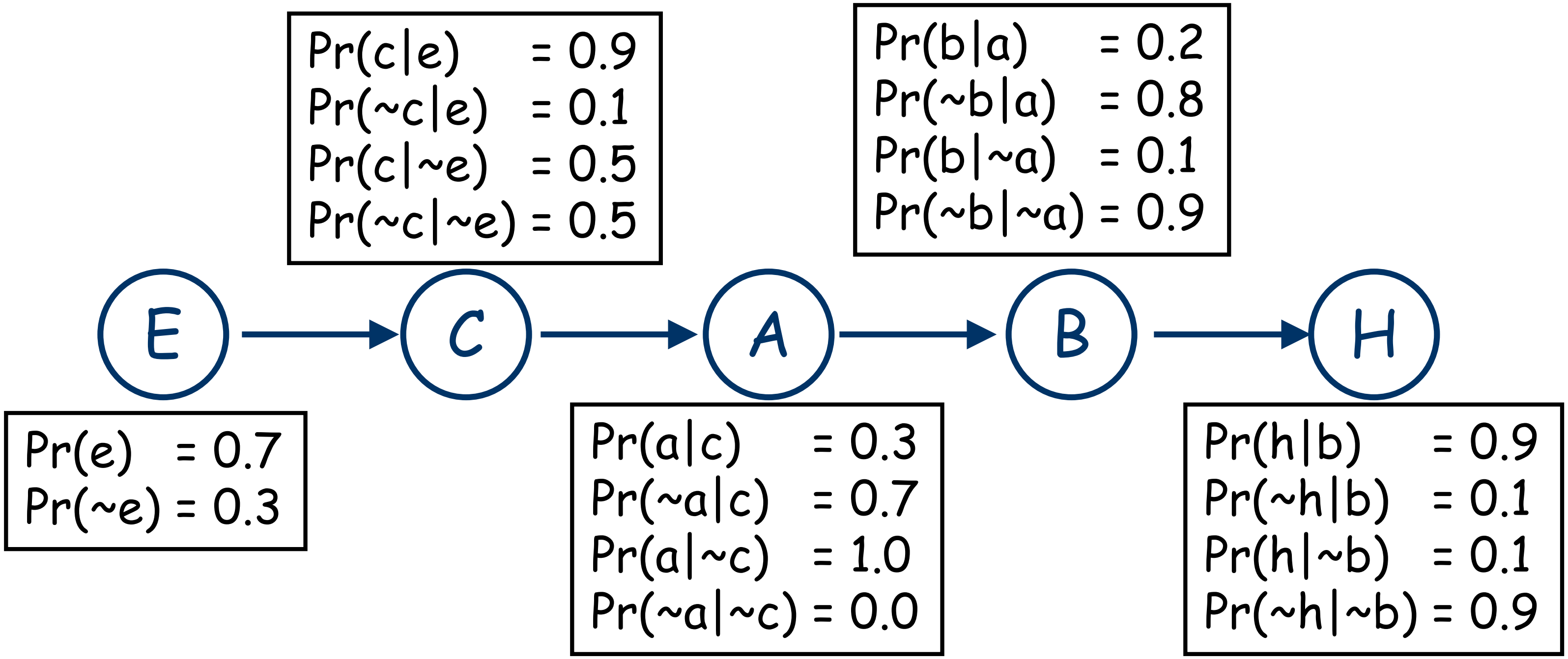
- By the chain rule (for any instantiation of *H,B,A,C,E*):
  - $Pr(H,B,A,C,E) =$
    $Pr(H|B,A,C,E)\ Pr(B|A,C,E)\ Pr(A|C,E)\ Pr(C|E)\ Pr(E)$
- By our independence assumptions:
  - $Pr(H,B,A,C,E) =$
    $Pr(H|B)\ Pr(B|A)\ Pr(A|C)\ Pr(C|E)\ Pr(E)$
- We can specify the full joint by specifying five *local conditional distributions*: $Pr(H|B)$; $Pr(B|A)$; $Pr(A|C)$; $Pr(C|E)$; and $Pr(E)$

Example Quantification



- Specifying the joint requires only 9 parameters (if we note that half of these are "1 minus" the others), instead of 31 for explicit representation
  - linear in number of variables instead of exponential!
  - linear *generally* if dependence has a chain structure

Recovering Joint is Easy



- Use chain rule and multiply parameters provided
  - *Pr(h b ~a c ~e)*

$$= Pr(h|b)P(b|{\sim}a)P({\sim}a|c)P(c|{\sim}e)P({\sim}e)$$
$$= 0.9 * 0.1 * 0.3 * 0.5 * 0.3$$
$$= 0.00405$$

Inference is Easy

- Want to know *P(a)*? Use summing out rule:

$$
\begin{aligned}
P(a) &= \sum_{c_i \in Dom(C)} \Pr(a \mid c_i)\Pr(c_i) \\
&= \sum_{c_i \in Dom(C)} \Pr(a \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)
\end{aligned}
$$

Inference is Easy

- Computing *P(a)* in more concrete terms:
  - $P(c) = P(c|e)P(e) + P(c|{\sim}e)P({\sim}e)$
    $= 0.8 * 0.7 + 0.5 * 0.3 = 0.78$
  - $P({\sim}c) = P({\sim}c|e)P(e) + P({\sim}c|{\sim}e)P({\sim}e) = 0.22$
    - $P({\sim}c) = 1 - P(c)$, as well
  - $P(a) = P(a|c)P(c) + P(a|{\sim}c)P({\sim}c)$
    $= 0.7 * 0.78 + 0.0 * 0.22 = 0.546$
  - $P({\sim}a) = 1 - P(a) = 0.454$

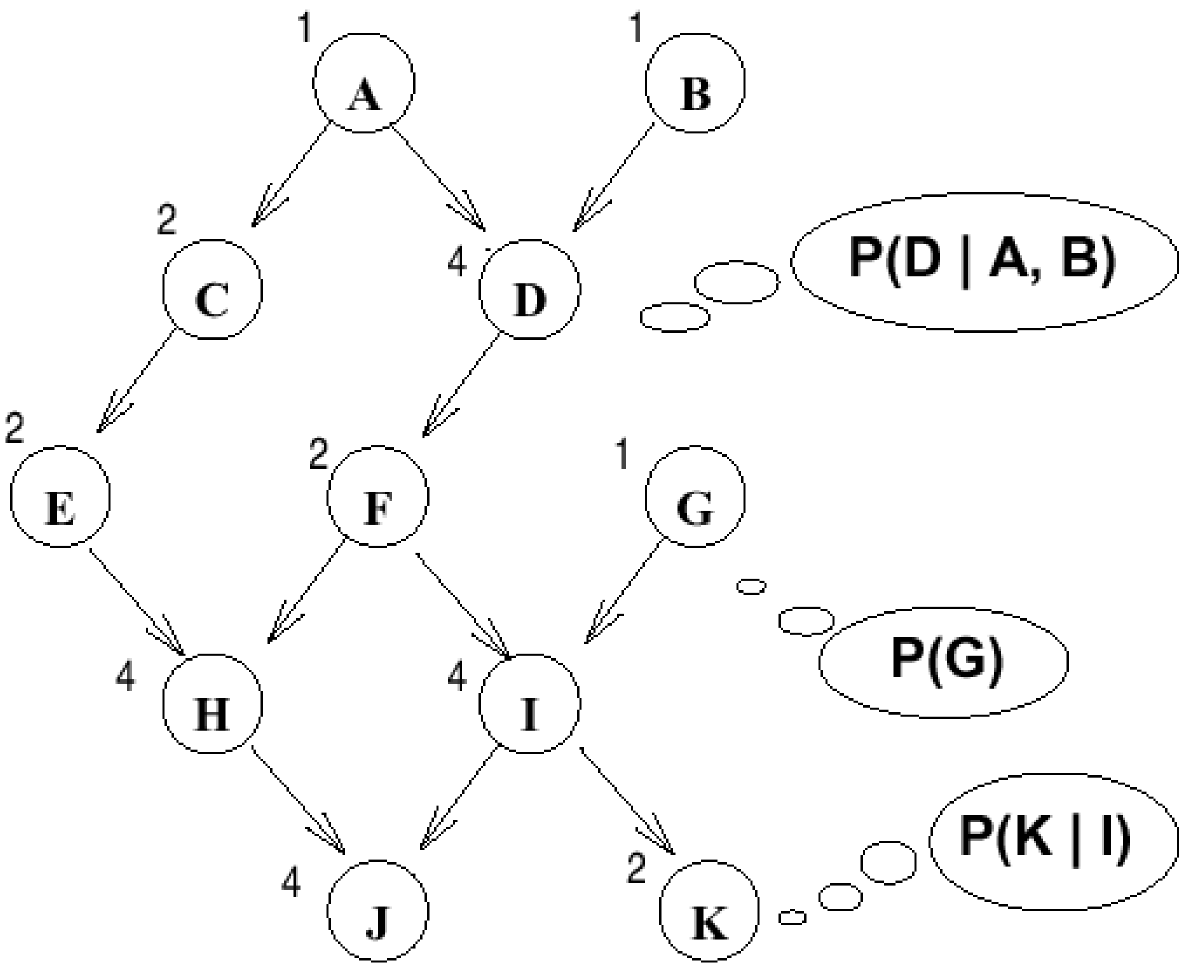
Bayesian Networks

- The structure above is a *Bayesian network*. A BN is a *graphical representation* of the direct dependencies over a set of variables, together with a set of *conditional probability tables* quantifying the strength of those influences.

- A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:
  - a DAG whose nodes are the variables
  - a set of CPTs $Pr(X_i \mid Par(X_i))$ for each X_i

- Key notions: parent, child, descendent, ancestor (all very intuitive)

An Example Bayes Net



- A couple CPTS are "shown"
- Explict joint requires 2^{11} -1 =2047 parameters *(assuming binary vars)*
- BN requires only 27 parameters (the number of entries for each CPT is listed)

Semantics of a Bayes Net

- The structure of the BN means: every X_i is *conditionally independent of all of its non-descendants given it parents*:

$$Pr(X_i \mid S \cup Par(X_i)) = Pr(X_i \mid Par(X_i))$$

for any subset $S \subseteq NonDescendents(X_i)$

Semantics of Bayes Nets (2)

- If we ask for $Pr(x_1, x_2, \ldots, x_n)$ we obtain
  - assuming an ordering consistent with network
- $Pr(x_1, x_2, \ldots, x_n)$
  $= Pr(x_n \mid x_{n-1}, \ldots, x_1)\, Pr(x_{n-1} \mid x_{n-2}, \ldots, x_1) \ldots Pr(x_1)$
  $= Pr(x_n \mid Par(X_n))\, Pr(x_{n-1} \mid Par(X_{n-1})) \ldots Pr(x_1)$
- Thus, any element of the joint is easily computable using the parameters specified in an arbitrary BN

Constructing a Bayes Net

- Given any distribution over variables $X_1, X_2, \ldots, X_n$, we can construct a Bayes net that faithfully represents that distribution.

Take *any ordering* of the variables (say, the order given), and go through the following procedure for X_n down to X_1. Let $Par(X_n)$ be any subset $S \subseteq \{X_1, \ldots, X_{n-1}\}$ such that X_n is independent of $\{X_1, \ldots, X_{n-1}\}$ - S given S. Such a subset must exist (convince yourself).

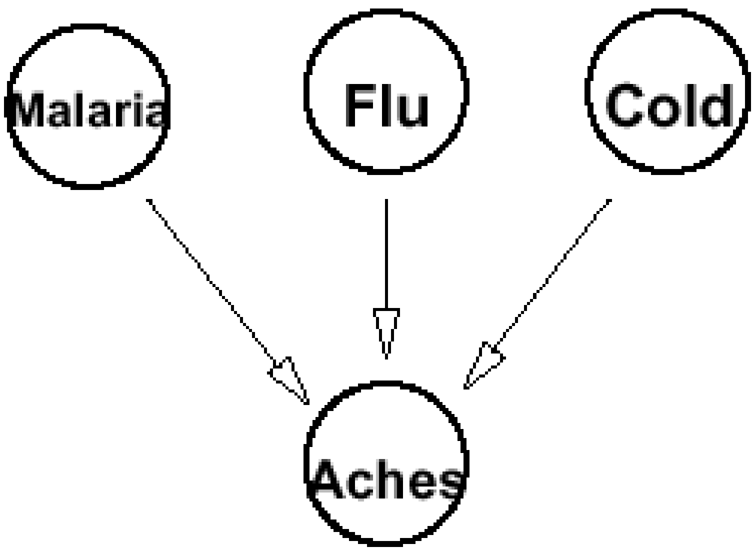
Then determine the parents of X_{n-1} the same way, finding a similar $S \subseteq \{X_1, \ldots, X_{n-2}\}$, and so on.

In the end, a DAG is produced and the BN semantics must hold by construction.

- *(Some other formal requirements must hold.)*

Causal Intuitions

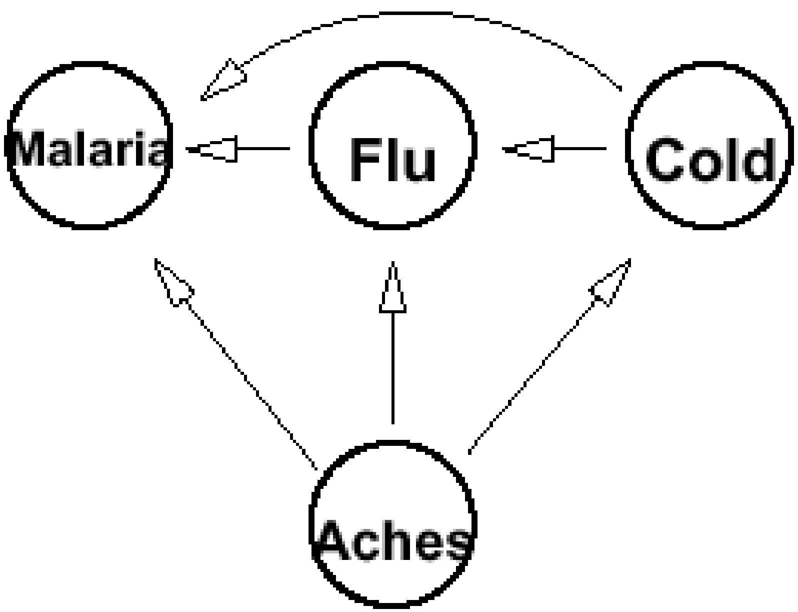
- The construction of a BN is simple
  - works with *arbitrary* orderings of variable set
  - but some orderings much better than others!
  - generally, if ordering/dependence structure reflects causal intuitions, a more natural, compact BN results



- In this BN, we've used the ordering *Mal, Cold, Flu, Aches* to build BN for distribution *P*
  - Variable can only have parents that come earlier in the ordering

Causal Intuitions

- Suppose we build the BN for distribution P using the opposite ordering
  - i.e., we use ordering *Aches, Cold, Flu, Malaria*
  - resulting network is more complicated!



- *Mal* depends on *Aches*; but it also depends on *Cold, Flu given Aches*
  - *Cold*, *Flu* explain away *Mal* given *Aches*
- *Flu* depends on *Aches*; but also on *Cold given Aches*
- *Cold* depends on *Aches*

Testing Independence

- Given BN, how do we determine if two variables *X*, *Y* are independent (given evidence *E*)?
  - we use a (simple) graphical property

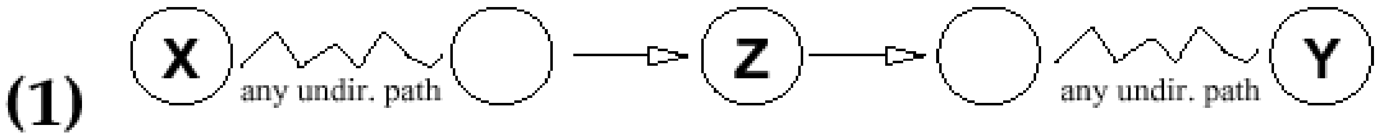
- **D-separation**: A set of variables ***E*** *d-separates* *X* and *Y* if it *blocks every undirected path* in the BN between *X* and *Y*. (We'll define *blocks* next.)

- *X* and *Y* are conditionally independent given evidence if ***E*** d-separates *X* and *Y*
  - thus BN gives us an easy way to tell if two variables are independent (set $E = \emptyset$) or cond. independent given *E*

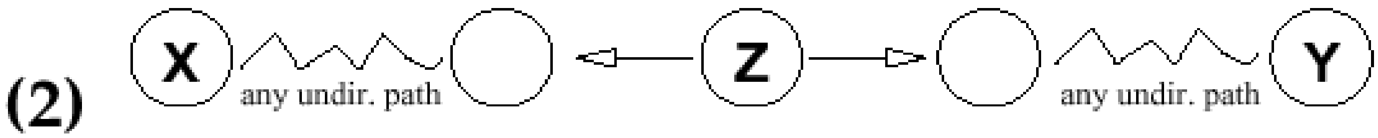
Blocking in D-Separation

- Let *P* be an undirected path from *X* to *Y* in a BN. Let ***E*** be an evidence set. We say ***E*** *blocks path P* iff there is some node *Z* on the path such that:
  - **Case 1:** one arc on *P* goes into *Z* and one goes out, and $Z \in$ ***E***; or
  - **Case 2:** both arcs on *P* leave *Z*, and $Z \in$ ***E***; or
  - **Case 3:** both arcs on *P* enter *Z* and *neither Z, nor any of its descendents*, are in ***E***.

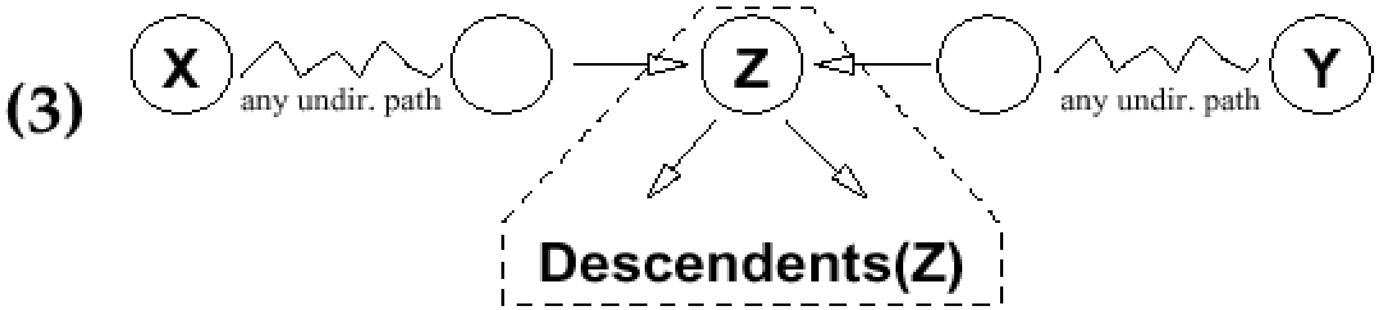
Blocking: Graphical View



If Z in evidence, the path between X and Y blocked

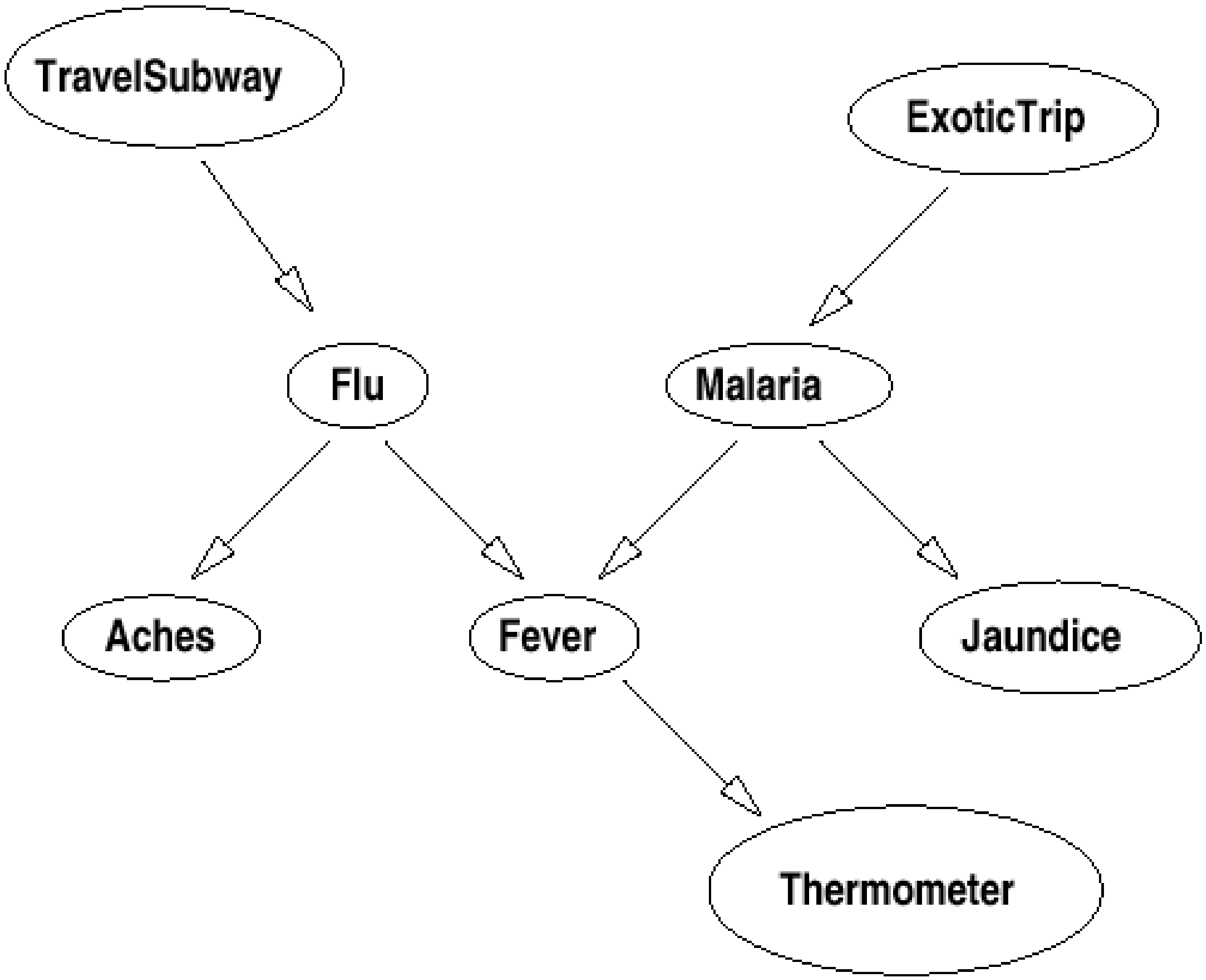


If Z in evidence, the path between X and Y blocked



If Z is ***not*** in evidence and ***no*** descendent of Z is in evidence, then the path between X and Y is blocked

D-Separation: Intuitions



D-Separation: Intuitions

- Subway and Therm are dependent; but are independent given Flu (since Flu blocks the only path)
- Aches and Fever are dependent; but are independent given Flu (since Flu blocks the only path). Similarly for Aches and Therm (dependent, but indep. given Flu).
- Flu and Mal are indep. (given no evidence): Fever blocks the path, since it is *not in evidence*, nor is its decsendant Therm. Flu,Mal are dependent given Fever (or given Therm): nothing blocks path now.
- Subway,ExoticTrip are indep.; they are dependent given Therm; they are indep. given Therm and Malaria. This for exactly the same reasons for Flu/Mal above.

Inference in Bayes Nets

- The independence sanctioned by d-separation allows us to compute prior and posterior probabilities quite effectively.

- We'll look at a couple simple examples to illustrate. We'll focus on networks without *loops*. (A loop is a cycle in the underlying *undirected* graph. Recall the directed graph has no cycles.)

Simple Forward Inference (Chain)

- Computing prior require simple forward "propagation" of probabilities (using Subway net)

$$P(J) = \Sigma_{M,ET}\, P(J|M,ET)\, P(M,ET)$$
$$= \Sigma_{M,ET}\, P(J|M)\, P(M|ET)\, P(ET)$$
$$= \Sigma_{M}\, P(J|M)\, \Sigma_{ET}\, P(M|ET)\, P(ET)$$

- (1) follows by summing out rule; (2) by chain rule and independence; (3) by distribution of sum
- Note: only ancestors of *J* considered

Simple Forward Inference (Chain)

- Same idea applies when we have "upstream" evidence

$$P(J \mid et) = \Sigma_M P(J \mid M,et)\, P(M \mid et)$$
$$= \Sigma_M P(J \mid M)\, P(M \mid et)$$

Simple Forward Inference (Pooling)

- Same idea applies with multiple parents

$$P(Fev) = \Sigma_{Flu,M}\, P(Fev|Flu,M)\, P(Flu,M)$$

$$= \Sigma_{Flu,M}\, P(Fev|Flu,M)\, P(Flu)\, P(M)$$

$$= \Sigma_{Flu,M}\, P(Fev|Flu,M)\, \Sigma_{TS}\, P(Flu|TS)\, P(TS)\, \Sigma_{ET}\, P(M|ET)\, P(ET)$$

- (1) follows by summing out rule; (2) by independence of *Flu, M*; (3) by summing out
  - note: all terms are CPTs in the Bayes net

Simple Forward Inference (Pooling)

- Same idea applies with evidence

$$P(Fev|ts,{\sim}m) = \Sigma_{Flu}\, P(Fev\,|Flu,ts,{\sim}m)\; P(Flu|\, ts,{\sim}m)$$

$$= \Sigma_{Flu}\, P(Fev|Flu,{\sim}m)\; P(Flu|ts,{\sim}m)$$

Simple Backward Inference

- When evidence is downstream of query variable, we must reason "backwards." This requires the use of Bayes rule:

$$P(ET \mid j) = \alpha P(j \mid ET)\, P(ET)$$

$$= \alpha \Sigma_M P(j \mid M,ET)\, P(M|ET)\, P(ET)$$

$$= \alpha \Sigma_M P(j \mid M)\, P(M|ET)\, P(ET)$$

- First step is just Bayes rule
  - normalizing constant α is $1/P(j)$; but we needn't compute it explicitly if we compute $P(ET \mid j)$ for each value of ET: we just add up terms $P(j \mid ET)\, P(ET)$ for all values of ET (they sum to $P(j)$)

Backward Inference (Pooling)

- Same ideas when several pieces of evidence lie "downstream"

$$P(ET \mid j,fev) = \alpha P(j,fev \mid ET)\, P(ET)$$

$$= \alpha \Sigma_M P(j,fev \mid M,ET)\, P(M|ET)\, P(ET)$$

$$= \alpha \Sigma_M P(j,fev \mid M)\, P(M|ET)\, P(ET)$$

$$= \alpha \Sigma_M P(j \mid M)\, P(fev \mid M)\, P(M|ET)\, P(ET)$$

- Same steps as before; but now we compute prob of both pieces of evidence given hypothesis *ET* and combine them. Note: they are independent given *M*; but not given *ET*.

Variable Elimination

- The intuitions in the above examples give us a simple inference algorithm for networks without loops: the *polytree* algorithm. We won't discuss it further. But be comfortable with the intuitions.

- Instead we'll look at a more general algorithm that works for general BNs; but the propagation algorithm will more or less be a special case.

- The algorithm, ***variable elimination***, simply applies the summing out rule repeatedly. But to keep computation simple, it exploits the independence in the network and the ability to distribute sums inward.

Factors

- A function $f(X_1, X_2,\ldots, X_k)$ is also called a *factor*. We can view this as table of numbers, one for each instantiation of the variables $X_1, X_2,\ldots, X_k$.
- A tabular rep'n of a factor is exponential in k
- Each CPT in a Bayes net is a factor:
  - e.g., $Pr(C|A,B)$ is a function of three variables, A, B, C
- Notation: $f(\mathbf{X},\mathbf{Y})$ denotes a factor over the variables $\mathbf{X} \cup \mathbf{Y}$. (Here $\mathbf{X}$, $\mathbf{Y}$ are *sets* of variables.)

The Product of Two Factors

- Let $f(\boldsymbol{X},\boldsymbol{Y})$ and $g(\boldsymbol{Y},\boldsymbol{Z})$ be two factors with variables $\boldsymbol{Y}$ in common
- The *product* of f and g, denoted $h = f \times g$ (or sometimes just $h = fg$), is defined:

$$h(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = f(\boldsymbol{X},\boldsymbol{Y}) \times g(\boldsymbol{Y},\boldsymbol{Z})$$

f(A,B)		g(B,C)		h(A,B,C)			
ab	0.9	bc	0.7	abc	0.63	ab~c	0.27
a~b	0.1	b~c	0.3	a~bc	0.08	a~b~c	0.02
~ab	0.4	~bc	0.8	~abc	0.28	~ab~c	0.12
~a~b	0.6	~b~c	0.2	~a~bc	0.48	~a~b~c	0.12

Summing a Variable Out of a Factor

- Let $f(X,\mathbf{Y})$ be a factor with variable X ($\mathbf{Y}$ is a set)
- We *sum out* variable X from f to produce a new factor $h = \Sigma_X f$, which is defined:

$$h(\mathbf{Y}) = \Sigma_{x \in Dom(X)} f(x,\mathbf{Y})$$

f(A,B)		h(B)	
ab	0.9	b	1.3
a~b	0.1	~b	0.7
~ab	0.4		
~a~b	0.6		

Restricting a Factor

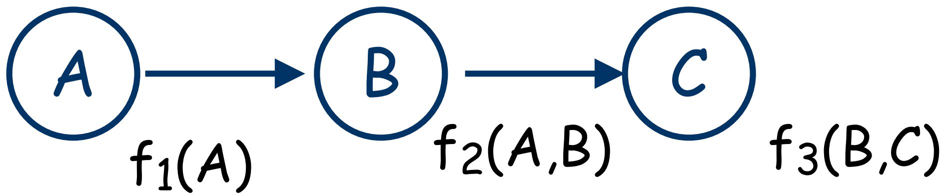
- Let $f(X,\mathbf{Y})$ be a factor with variable X ($\mathbf{Y}$ is a set)
- We *restrict* factor f *to* $X=x$ by setting X to the value x and "deleting". Define $h = f_{X=x}$ as:

$$h(\mathbf{Y}) = f(x,\mathbf{Y})$$

f(A,B)		$h(B) = f_{A=a}$	
ab	0.9	b	0.9
a~b	0.1	~b	0.1
~ab	0.4		
~a~b	0.6		

Variable Elimination: No Evidence

- Computing prior probability of query var X can be seen as applying these operations on factors

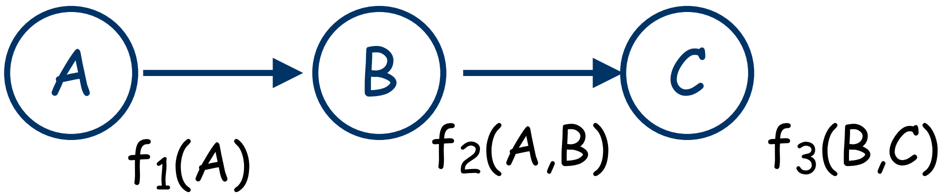


$$P(C) = \Sigma_{A,B}\, P(C|B)\, P(B|A)\, P(A)$$
$$= \Sigma_B\, P(C|B)\, \Sigma_A\, P(B|A)\, P(A)$$
$$= \Sigma_B\, f_3(B,C)\, \Sigma_A\, f_2(A,B)\, f_1(A)$$
$$= \Sigma_B\, f_3(B,C)\, f_4(B)$$
$$= f_5(C)$$

Define new factors: $f_4(B) = \Sigma_A\, f_2(A,B)\, f_1(A)$ and $f_5(C) = \Sigma_B\, f_3(B,C)\, f_4(B)$

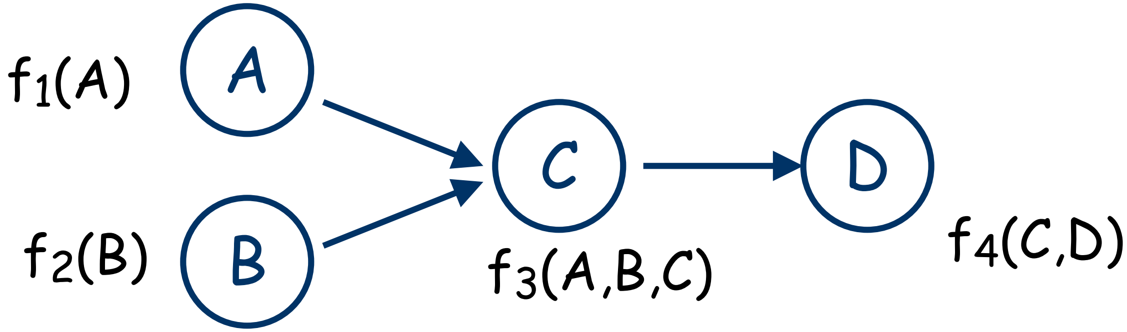
Variable Elimination: No Evidence

- Here's the example with some numbers



$f_1(A)$		$f_2(A,B)$		$f_3(B,C)$		$f_4(B)$		$f_5(C)$	
a	0.9	ab	0.9	bc	0.7	b	0.85	c	0.625
~a	0.1	a~b	0.1	b~c	0.3	~b	0.15	~c	0.375
		~ab	0.4	~bc	0.2				
		~a~b	0.6	~b~c	0.8				

VE: No Evidence (Example 2)



$$
\begin{aligned}
P(D) &= \Sigma_{A,B,C}\ P(D|C)\ P(C|B,A)\ P(B)\ P(A) \\
&= \Sigma_C\ P(D|C)\ \Sigma_B P(B)\ \Sigma_A P(C|B,A)\ P(A) \\
&= \Sigma_C f_4(C,D)\ \Sigma_B f_2(B)\ \Sigma_A f_3(A,B,C)\ f_1(A) \\
&= \Sigma_C f_4(C,D)\ \Sigma_B f_2(B)\ f_5(B,C) \\
&= \Sigma_C f_4(C,D)\ f_6(C) \\
&= f_7(D)
\end{aligned}
$$

Define new factors: $f_5(B,C)$, $f_6(C)$, $f_7(D)$, in the obvious way

Variable Elimination: One View

- One way to think of variable elimination:
  - write out desired computation using the chain rule, exploiting the independence relations in the network
  - arrange the terms in a convenient fashion
  - distribute each sum (over each variable) in as far as it will go
    - i.e., the sum over variable *X* can be "pushed in" as far as the "first" factor mentioning *X*
  - apply operations "inside out", repeatedly eliminating and creating new factors (note that each step/removal of a sum eliminates one variable)

Variable Elimination Algorithm

- Given query var Q, remaining vars $\boldsymbol{Z}$. Let F be set of factors corresponding to CPTs for $\{Q\} \cup \boldsymbol{Z}$.

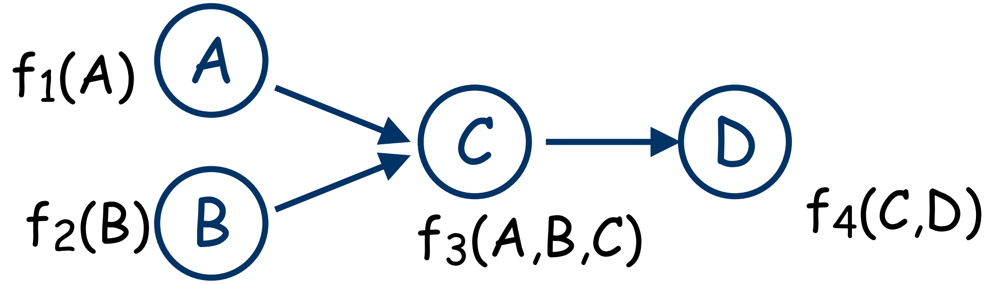
1. Choose an elimination ordering $Z_1, \ldots, Z_n$ of variables in $\boldsymbol{Z}$.
2. For each Z_j -- in the order given -- eliminate $Z_j \in \boldsymbol{Z}$ as follows:
   (a) Compute new factor $g_j = \Sigma_{Zj} f_1 \times f_2 \times \ldots \times f_k$,
   where the f_i are the factors in F that include Z_j
   (b) Remove the factors f_i (that mention Z_j) from F and add new factor g_j to F
3. The remaining factors refer only to the query variable Q. Take their product and normalize to produce $P(Q)$

VE: Example 2 again

Factors: $f_1(A)$ $f_2(B)$
$f_3(A,B,C)$ $f_4(C,D)$

Query: P(D)?

Elim. Order: A, B, C



Step 1: Add $f_5(B,C) = \Sigma_A f_3(A,B,C) f_1(A)$

Remove: $f_1(A)$, $f_3(A,B,C)$

Step 2: Add $f_6(C)= \Sigma_B f_2(B) f_5(B,C)$

Remove: $f_2(B)$, $f_5(B,C)$

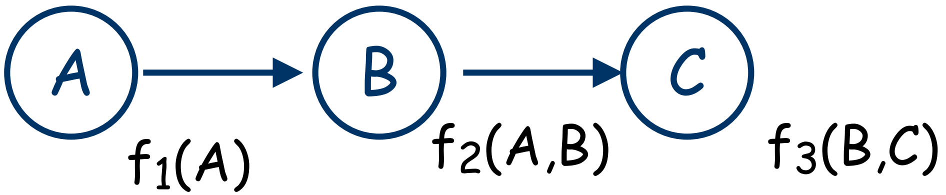
Step 3: Add $f_7(D) = \Sigma_C f_4(C,D) f_6(C)$

Remove: $f_4(C,D)$, $f_6(C)$

Last factor $f_7(D)$ is (possibly unnormalized) probability P(D)

Variable Elimination: Evidence

- Computing posterior of query variable given evidence is similar; suppose we observe *C*=*c*:



$$
\begin{aligned}
P(A|c) &= \alpha\, P(A)\, P(c|A) \\
&= \alpha\, P(A) \textstyle\sum_B P(c|B)\, P(B|A) \\
&= \alpha\, f_1(A) \textstyle\sum_B f_3(B,c)\, f_2(A,B) \\
&= \alpha\, f_1(A) \textstyle\sum_B f_4(B)\, f_2(A,B) \\
&= \alpha\, f_1(A)\, f_5(A) \\
&= \alpha\, f_6(A)
\end{aligned}
$$

New factors: $f_4(B)= f_3(B,c)$; $f_5(A)= \sum_B f_2(A,B)\, f_4(B)$; $f_6(A)= f_1(A)\, f_5(A)$

Variable Elimination with Evidence

Given query var Q, evidence vars $\mathbf{E}$ (observed to be $\mathbf{e}$), remaining vars $\mathbf{Z}$. Let F be set of factors involving CPTs for $\{Q\} \cup \mathbf{Z}$.

1. Replace each factor $f \in F$ that mentions a variable(s) in $\mathbf{E}$ with its restriction $f_{\mathbf{E}=\mathbf{e}}$ (somewhat abusing notation)
2. Choose an elimination ordering $Z_1, \ldots, Z_n$ of variables in $\mathbf{Z}$.
3. Run variable elimination as above.
4. The remaining factors refer only to the query variable Q. Take their product and normalize to produce $P(Q)$

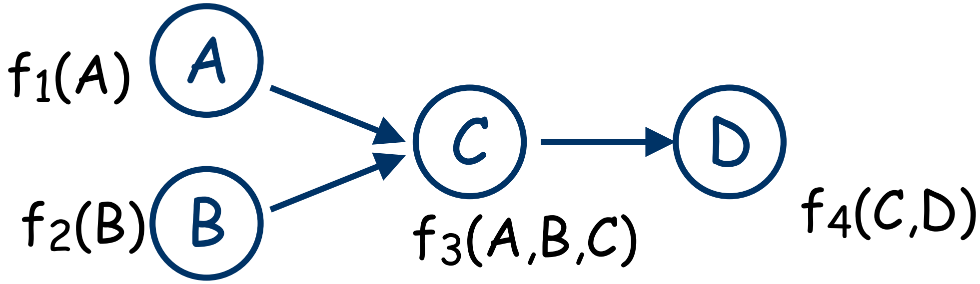
VE: Example 2 again with Evidence

Factors: $f_1(A)$ $f_2(B)$ $f_3(A,B,C)$ $f_4(C,D)$

Query: P(A)?

Evidence: D = d

Elim. Order: C, B



Restriction: replace $f_4(C,D)$ with $f_5(C) = f_4(C,d)$

Step 1: Add $f_6(A,B)= \Sigma_C f_5(C) f_3(A,B,C)$

Remove: $f_3(A,B,C)$, $f_5(C)$

Step 2: Add $f_7(A) = \Sigma_B f_6(A,B) f_2(B)$

Remove: $f_6(A,B)$, $f_2(B)$

Last factors: $f_7(A)$, $f_1(A)$. The product $f_1(A) \times f_7(A)$ is (possibly unnormalized) posterior. So… $P(A|d) = \alpha f_1(A) \times f_7(A)$.

Some Notes on the VE Algorithm

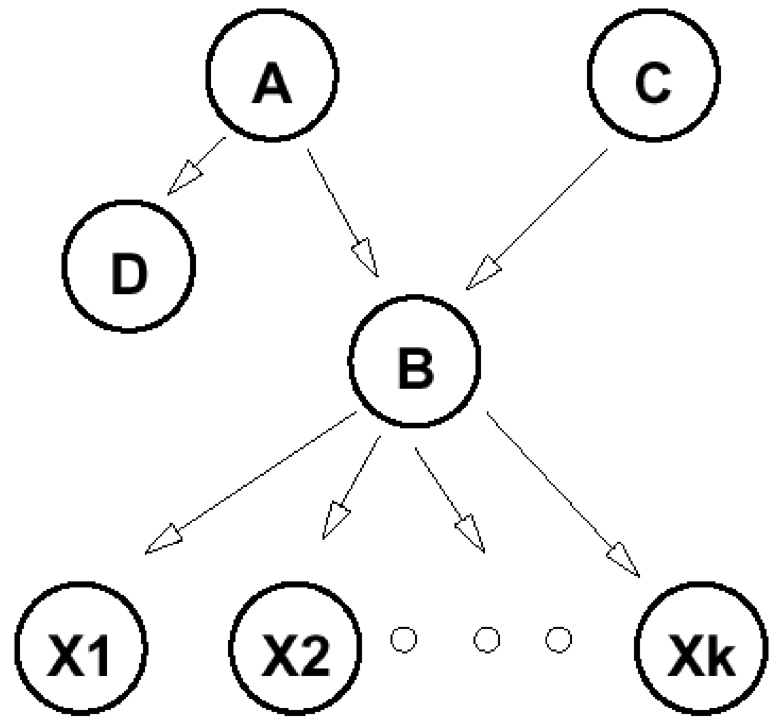
- After iteration *j* (elimination of Z_j), factors remaining in set *F* refer only to variables $X_{j+1}, \ldots Z_n$ and *Q*. No factor mentions an evidence variable *E* after the initial restriction.
- Number of iterations: linear in number of variables
- Complexity is linear in number of vars and exponential in size of the largest factor. (Recall each factor has exponential size in its number of variables.) Can't do any better than size of BN (since its original factors are part of the factor set). When we create new factors, we might make a set of variables larger.

Some Notes on the VE Algorithm

- The size of the resulting factors is determined by elimination ordering! (We'll see this in detail)
- For *polytrees*, easy to find good ordering (e.g., work outside in).
- For general BNs, sometimes good orderings exist, sometimes they don't (then inference is exponential in number of vars).
  - Simply *finding* the optimal elimination ordering for general BNs is NP-hard.
  - Inference in general is NP-hard in general BNs

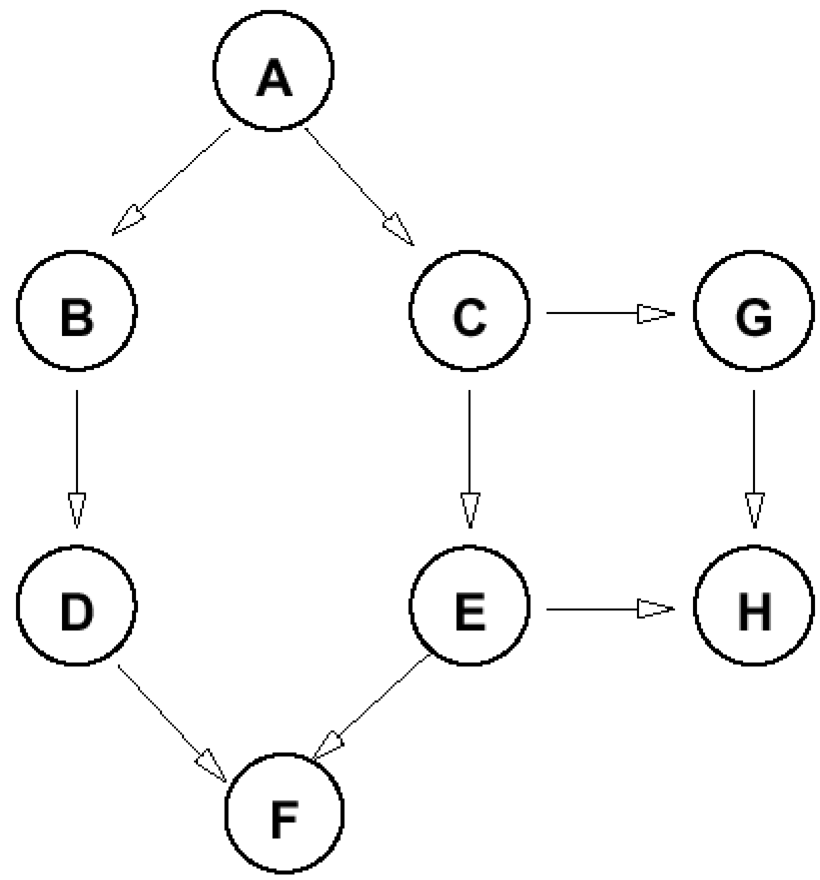
Elimination Ordering: Polytrees

- Inference is linear in size of network
  - ordering: eliminate only “singly-connected” nodes
  - e.g., in this network, eliminate *D, A, C, X1,…*; or eliminate *X1,… Xk, D, A, C*
  - result: no factor ever larger than original CPTs
  - eliminating *B* before these gives large factors!



Effect of Different Orderings

- Suppose query variable is *D*. Consider different orderings for this network
  - *A,F,H,G,B,C,E:*
    - good: why?
  - *E,C,A,B,G,H,F:*
    - bad: why?
- Which ordering creates smallest factors?
  - either max size or total
  - which creates largest?



Complexity of VE

- Given BN, elim. ordering. Let *induced graph* be the undirected graph obtained by joining any two variables that occur is some factor that occurs during VE.
- Each (maximal) clique in induced graph corresponds to a factor, and each factor is a subset of some clique.
- Hence: complexity is exponential in size of largest clique.
- Induced graph: moralized and triangulated

Relevance

- Certain variables have no impact on the query. In ABC network above, computing Pr(A given no evidence requires elimination of B and C. But when you sum out these vars, you compute a trivial factor (whose value are all ones).
- Can restrict attention to *relevant* variables. Given query Q, evidence **E**:
  - Q is relevant
  - if any node Z is relevant, its parents are relevant
  - if E∈**E** is a descendent of a relevant node, then E is relevant