Deliberative Agents
Knowledge Representation: Probabilistic

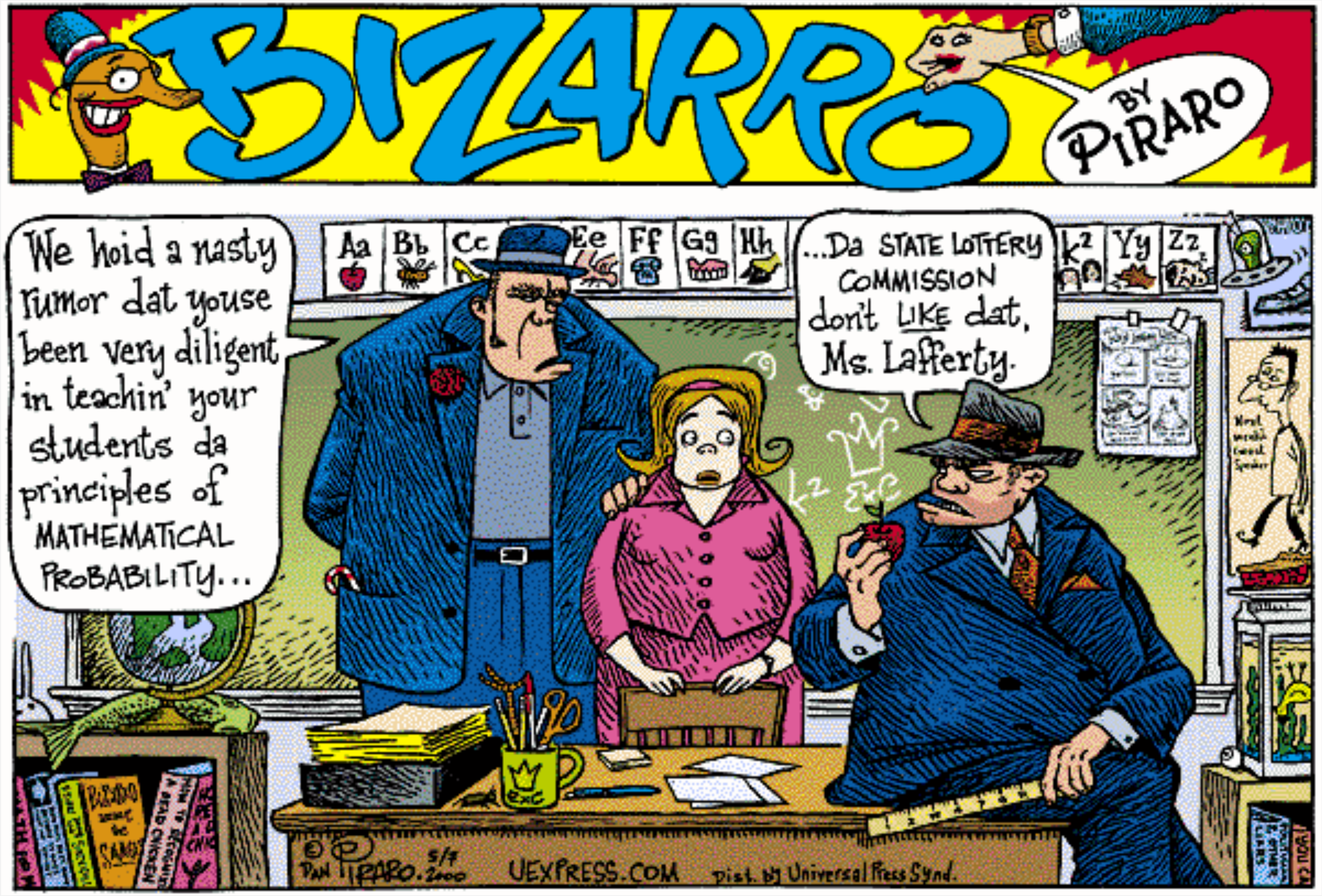
Vasant Honavar
Artificial Intelligence Research Laboratory
Informatics Graduate Program
Computer Science and Engineering Graduate Program
Bioinformatics and Genomics Graduate Program
Neuroscience Graduate Program

Center for Big Data Analytics and Discovery Informatics
Huck Institutes of the Life Sciences
Institute for Cyberscience
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University

vhonavar@ist.psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

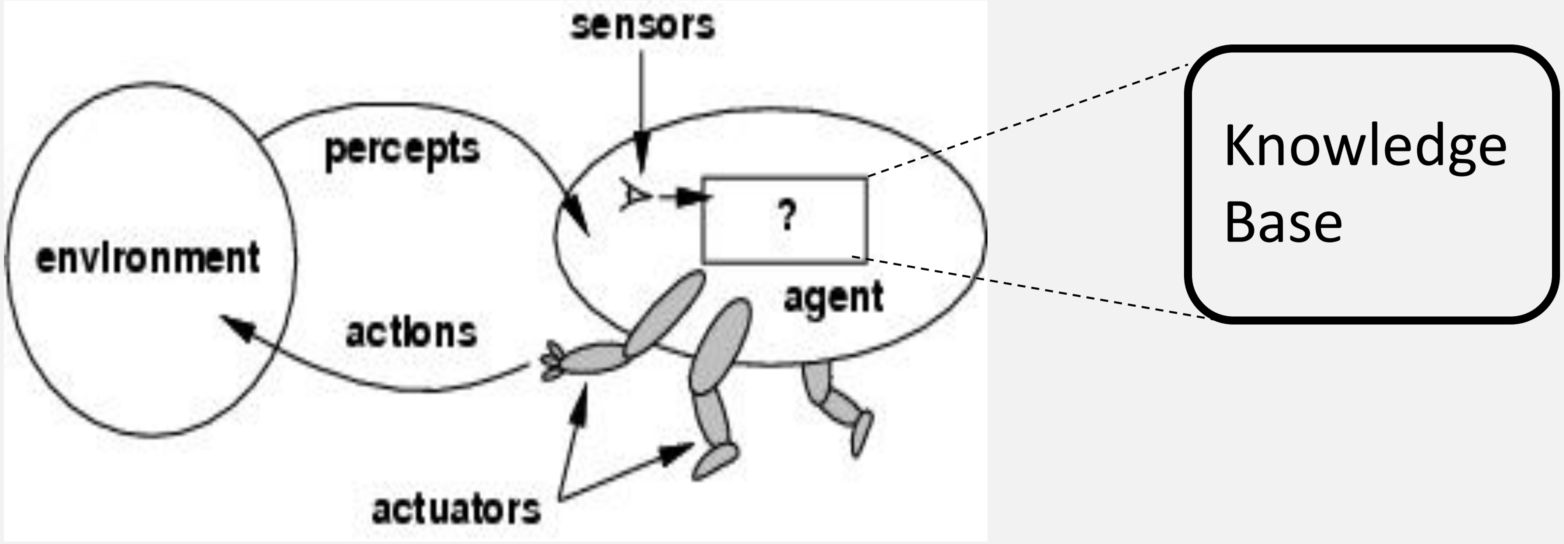
Probabilistic Knowledge Representation

- Basic probability theory
- Syntax and Semantics
- Random variables
- Distributions over random variables
- Independence and conditional independence
- Bayesian Network Representation
- Inference Using Bayesian Networks


BIZARRO
BY PIRARO
We hoid a nasty rumor dat youse been very diligent in teachin' your students da principles of MATHEMATICAL PROBABILITY...
...Da STATE LOTTERY COMMISSION don't LIKE dat, Ms. Lafferty.
© DAN PIRARO 5/7 2000
UEXPRESS.COM
Dist. by Universal Press Synd.

Agents That Represent and Reason Under Uncertainty

- Intelligent behavior requires knowledge about the world
- Often, we are uncertain about the state of the world



Representing and Reasoning under Uncertainty

- Probability Theory provides a framework for representing and reasoning under uncertainty
  - Represent beliefs about the world as sentences (much like in propositional logic)
  - Associate probabilities with sentences
  - Reason by manipulating sentences according to sound rules of probabilistic inference
  - Results of inference are probabilities associated with conclusions that are justified by beliefs and data (observations)
- Allows agents to substitute thinking for acting in the world

Representing and Reasoning under Uncertainty

- Beliefs:
  - If Oksana studies, there is an 60% chance that she will pass the test; and a 40 percent chance that she will not.
  - If she does not study, there is 20% percent chance that she will pass the test and 80% chance that she will not.
- Observation: Oksana did not study.
- Example Inference task:
  - What is the chance that Oksana will pass the test?
  - What is the chance that she will fail?
- Probability theory generalizes propositional logic
  - Probability theory associates probabilities that lie in the interval [0,1] as opposed to 0 or 1 (exclusively)

Probability Theory as a Knowledge Representation

- Ontological commitments (what do we want to talk about?)
  - Propositions that represent the agent's beliefs about the world
- Epistemological Commitments (what can we believe?)
  - What is the *probability* that a given proposition true (given the beliefs and observations)?
- Syntax
  - Much like propositional logic
- Semantics
  - Relative frequency interpretation
  - Bayesian interpretation
- Proof Theory
  - Based on laws of probability

Sources of uncertainty

Uncertainty modeled by Probabilistic assertions may

- In a deterministic world be due to
  - Laziness: failure to enumerate exceptions, qualifications, etc. that may be too numerous to state explicitly
  - Sensory limitations
  - Ignorance: lack of relevant facts etc.
- In a stochastic world be due to
  - Inherent uncertainty (as in quantum physics)

The framework is agnostic about the source of uncertainty

The world according to Agent Bob

- An atomic event or world state is a complete specification of the state of the agent's world.
- Event set is a set of mutually exclusive and exhaustive possible world states (relative to an agent's representational commitments and sensing abilities)
- From the point of view of an agent Bob who can sense only 3 colors and 2 shapes, the world can be in only one of 6 states
- Atomic events (world states) are
  - mutually exclusive
  - exhaustive

Semantics: Probability as a subjective measure of belief

- Suppose there are 3 agents – Oksana, Cornelia, Jun, in a world where a fair dice has been tossed.
- Oksana observes that the outcome is a "6" and whispers to Cornelia that the outcome is "even" but
- Jun knows nothing about the outcome.

Set of possible mutually exclusive and exhaustive world states = {1, 2, 3, 4, 5, 6}

Set of possible states of the world based on what Cornelia knows = {2, 4, 6}

Probability as a subjective measure of belief

Probability is a measure over all of the world states that are possible, or simply, possible worlds, given what an agent knows

$$Possibleworlds_{Oksana} = \{6\}, Possibleworlds_{Cornelia} = \{2,4,6\}$$
$$Possibleworlds_{Jun} = \{1,2,3,4,5,6\}$$

$$\Pr_{Oksana}(worldstate = 6) = 1$$

$$\Pr_{Cornelia}(worldstate = 6) = \frac{1}{3}$$

$$\Pr_{Jun}\left(worldstate = 6\right) = \frac{1}{6}$$

Oksana, Cornelia, and Jun assign different beliefs to the same world state because of differences in their knowledge!

Random variables

- The “domain” of a random variable is the set of values it can take. The values are mutually exclusive and exhaustive.
- The domain of a Boolean random variable X is {true, false} or {1, 0}
- Discrete random variables take values from a countable domain.
  - The domain of the random variable Color may be {Red, Green}.
  - If E = {(Red, Square), (Green, Circle), (Red, Circle), (Green, Square)}, the proposition (Color = Red) is True in the world states {(Red, Square), (Red, Circle)}.
  - Each state of a discrete random variable corresponds to a proposition e.g., (Color = Red)

Syntax

- Basic element: random variable
    - Similar to propositional logic: possible worlds defined by assignment of values to random variables.
    - *Cavity* (do I have a cavity?)
    - *Weather* is one of <*sunny, rainy, cloudy, snow*>
    - Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a random variable
    - *Weather = sunny=true* (abbreviated as *sunny), Cavity = false* (abbreviated as $\neg$*cavity*)
- Complex propositions formed from elementary propositions and standard logical connectives
    - *Weather = sunny* $\vee$ $\neg$*cavity*

Syntax and Semantics

- Atomic event: A complete specification of the state of the world about which the agent is uncertain
- Atomic events correspond to a possible worlds (much like in the case of propositional logic)

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events or 4 possible worlds:

  $Cavity = false \wedge Toothache = false$

  $Cavity = false \wedge Toothache = true$

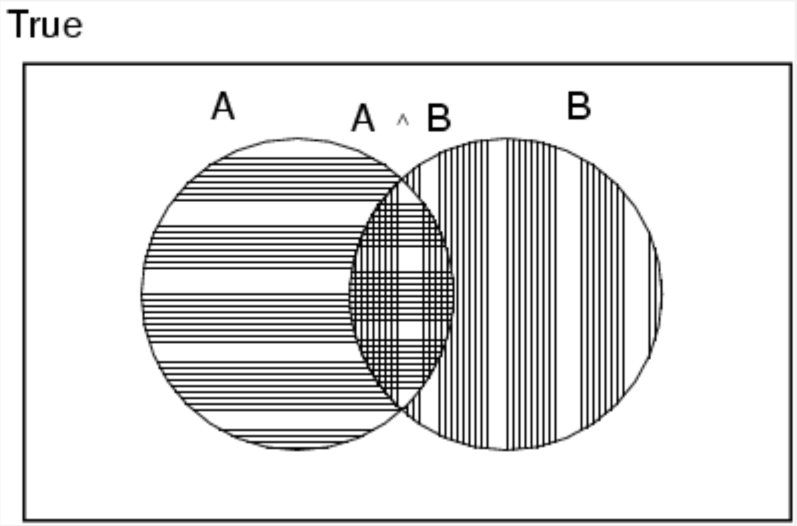
  $Cavity = true \wedge Toothache = false$

  $Cavity = true \wedge Toothache = true$

- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A, B*
  - $0 \leq P(A) \leq 1$
  - $P(\mathit{true}) = 1$ and $P(\mathit{false}) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Prior probability

- Prior or unconditional probabilities of propositions
  - P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence
- Probability distribution gives values for all possible assignments:
  - **P**(*Weather*) = <0.72, 0.1, 0.08, 0.1>
  - Note that the probabilities sum to 1
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
  - **P**(*Cavity,Play*) = a 4 × 2 matrix of values

Joint probability distribution

- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
  - **P**(*Weather, Cavity*) = a 4 × 2 matrix of values:
  -

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

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

Inference using the joint distribution

	Toothache	¬Toothache
Cavity	0.4	0.1
¬Cavity	0.1	0.4

$$P(cavity) = P(cavity, ache) + P(cavity, \neg ache)$$

Conditional probability

- Conditional or posterior probabilities
  - P(*Cavity* | *Toothache*) = 0.8
    (note *Cavity* is shorthand for *Cavity = True*)

  Probability of *Cavity* given *Toothache*

- Notation for conditional distributions:

  P(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors)

  P(*Cavity* | *Toothache, Cavity*) = 1

- New evidence may be irrelevant (Probability of Cavity given Toothache is independent of Weather)

  P(*Cavity* | *Toothache, Sunny*) = P(*Cavity* | *Toothache*) = 0.8

Conditional probability

- Definition of conditional probability:
  $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$

- Product rule gives an alternative formulation:
  - $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

Example:

- Suppose I have two coins – one a normal fair coin, and the other a rigged coin (with heads on both sides). I pick a coin at *random, toss it,* and tell you that the outcome of the toss is a Head.
- What is the probability that I am looking at a fair coin?

Conditional probability

Example:

- Suppose I have two coins – one a normal fair coin, and the other a rigged coin (with heads on both sides). I pick a coin at *random, toss it,* and tell you that the outcome of the toss is a Head.
- What is the probability that I am looking at a fair coin?
- (F, H), (F,T),(R,H), (R,T)

  ¼ , ¼, ½, 0

  $P(F|H) = P(F,H)/P(H)=(1/4)/(3/4) = 1/3$

Conditional probability

- A general version holds for whole distributions, e.g.,
  P(*Weather,Cavity*) = **P**(*Weather* | *Cavity*) **P**(*Cavity*)

- View as a compact notation for a set of 4 × 2 equations, not matrix multiplication

- Chain rule is derived by successive application of product rule:

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \ldots$$
$$= \pi_i\, \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \text{ (i ranges from 1 to n)}$$

Possible worlds semantics

- A possible world is an assignment of Truth values to every simple proposition about the world. Let Ω be a set of possible worlds. Let $\omega \in \Omega$ and let *p, q* be propositions (atomic sentences or syntactically well formed logical formulae). Then *p* is True in ω (written $\omega \models p$) where

$$\omega \models p \text{ if } \omega \text{ assigns value } True \text{ to } p$$

$$\omega \models p \wedge q \text{ if } \omega \models p \text{ and } \omega \models q$$

$$\omega \models p \vee q \text{ if } \omega \models p \text{ or } \omega \models q \text{ (or both)}$$

$$\omega \models \neg p \text{ if } \omega \not\models p$$

Possible Worlds and Random Variables

- A possible world is an assignment of exactly one value to every random variable. Let Ω be a set of possible worlds. Let $\omega \in \Omega$ and let f be a (logical) formula. Then f is True in ω (written $\omega \models f$) where

$$\omega \models X = v \text{ if } \omega \text{ assigns value } v \text{ to } X$$

$$\omega \models f \wedge g \text{ if } \omega \models f \text{ and } \omega \models g$$

$$\omega \models f \vee g \text{ if } \omega \models f \text{ or } \omega \models g \text{ (or both)}$$

$$\omega \models \neg f \text{ if } \omega \not\models f$$

Probability as a Measure over Possible worlds

- Associated with each possible world is a measure. When there are only a finite number of possible worlds, the measure of the world ω, denoted by μ(ω) has the following properties:

$$\forall \omega \in \Omega,\ 0 \leq \mu(\omega)$$

$$\sum_{\omega \in \Omega} \mu(\omega) = 1$$

The probability of a formula or state of affairs described by a sentence f , written as P (f), is the sum of the measures of the possible words in which f is True. That is,

$$P(f) = \sum_{\omega \models f} \mu(\omega)$$

Probability as a measure over possible worlds

- Suppose I have two coins – one a normal fair coin, and the other with 2 heads. I pick a coin at *random* and toss it. What is the probability that the outcome is a head?

$$\Omega = \{(Fair, H), (Fair, T), (Rigged, H), (Rigged, T)\}$$

$$\mu = \left\{ \frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0 \right\}$$

$$\Pr(H) = \sum_{\omega \models H} \mu(\omega) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

Conditional probability as a Measure over Possible worlds not ruled out by evidence

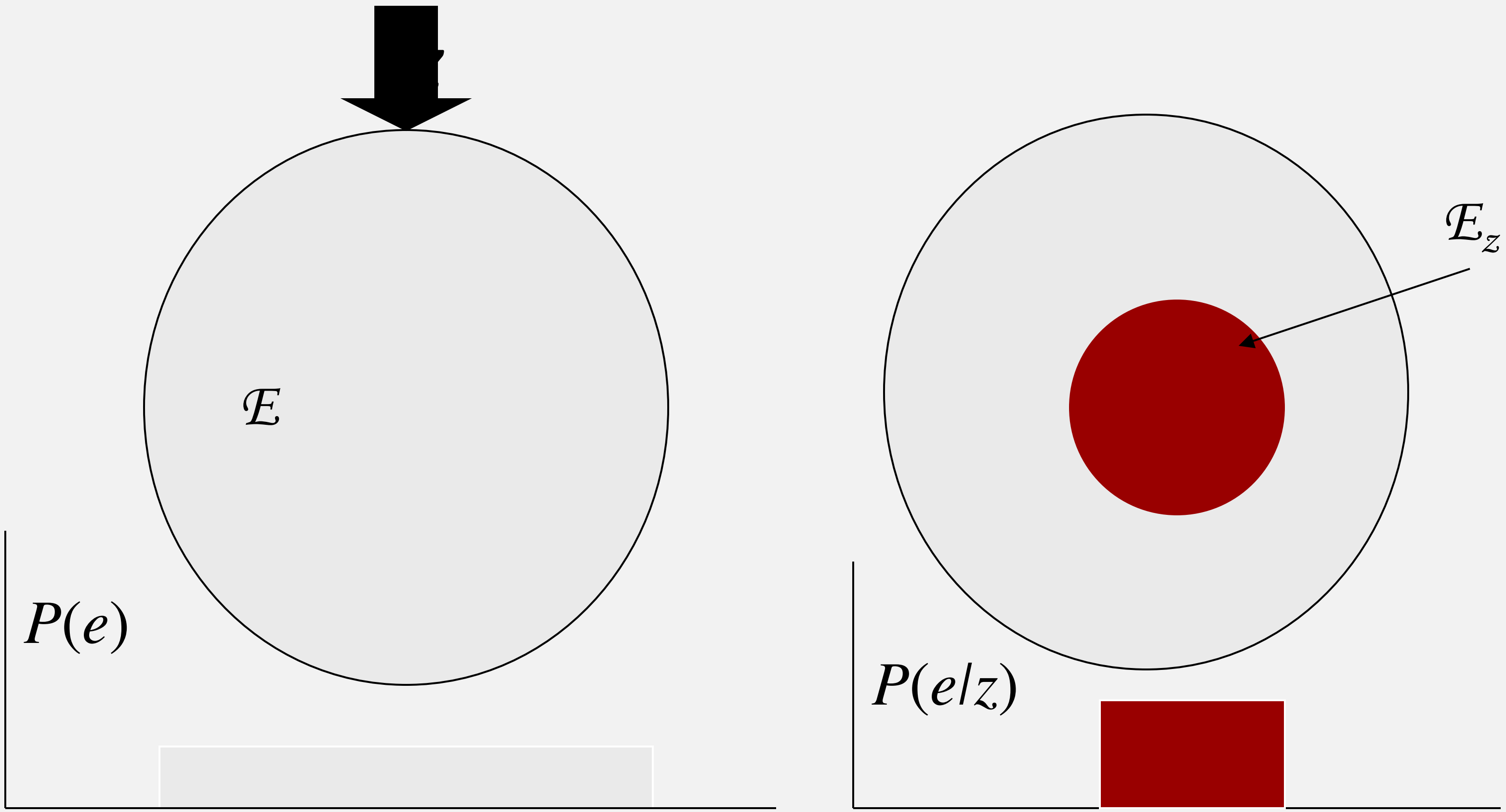
- A given piece of evidence e rules out all possible worlds that are incompatible with e or selects the possible worlds in which e is $True$. Evidence e *induces* a new measure μ_e.

$$\mu_e(\omega) = \begin{cases} \frac{1}{P(e)}\mu(\omega) \text{ if } \omega \models e \\ 0 \text{ if } \omega \not\models e \end{cases}$$

$$P(h|e) = \sum_{\omega \models h} \mu_e(\omega) = \frac{1}{P(e)} \sum_{\omega \models h \wedge e} \mu(\omega) = \frac{P(h \wedge e)}{P(e)}$$

Effect of Evidence on Possible worlds

Evidence z e.g., (color = red) rules out some assignments of values to some of the random variables



Evidence redistributes probability mass over possible worlds

- A given piece of evidence *z* rules out all possible worlds that are incompatible with *z* or selects the possible worlds in which *z* is *True*. Evidence *z induces* a distribution P_z

$$P_z(e) = \begin{cases} \frac{1}{P(z)} P(e) \text{ if } e \models z \\ 0 \text{ if } e \not\models z \end{cases}$$

$$P(h|z) = \sum_{e \models h} P_z(e) = \frac{1}{P(z)} \sum_{e \models h \wedge z} P(e) = \frac{P(h \wedge z)}{P(z)}$$

Defining probability as a Measure over Possible worlds – infinite sets of variables, continuous random variables

$$\forall \omega \in \Omega,\ 0 \leq \mu(\omega),\ \int_{\omega} \mu(\omega)\,\mathrm{d}\,\omega = 1, \quad P(f) = \int_{\omega \models f} \mu(\omega)\,\mathrm{d}\omega$$

When a random variable takes on real values the measure corresponds to a probability density function p. The probability that a random variable X takes values between *a* and *b* is given by

$$P(a \leq x \leq b) = \int_a^b p(x)\,\mathrm{d}\,x$$

This definition can be generalized to handle vector valued random variables

Example:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Note: we now have an infinite set of models

Inference by enumeration

- Start with the joint probability distribution:

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

- For any proposition ϕ, sum the measures of atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

Inference by enumeration

- Start with the joint probability distribution:

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

- For any proposition φ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability distribution:

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

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

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

$$= 0.4$$

Normalization

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

- Denominator can be viewed as a normalization constant α
- **P**(*Cavity* | *toothache*) = α **P**(*Cavity,toothache*)

  = α[**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache,¬ catch*)]

  = α[<0.108,0.016> + <0.012,0.064>]

  = α <0.12,0.08> = <0.6,0.4>
- General idea: compute distribution on query variable by fixing evidence variables and summing over unobserved variables

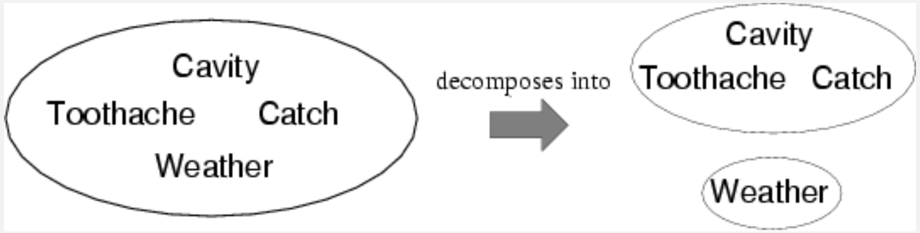
Inference by enumeration, continued

- Obvious problems:
  - Worst-case time complexity $O(d^n)$ where d is the largest arity
  - Space complexity $O(d^n)$ to store the joint distribution
  - How to find the numbers for $O(d^n)$ entries?

Independence

- A and B are independent iff

  $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



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

- 32 entries reduced to 12;
- n independent variables, $O(2^n)$ reduced to $O(n)$
- Absolute independence powerful but rare
- How can we manage a large numbers of variables?

Conditional independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)
- The same independence holds if I haven't got a cavity:
  - **P**(*catch* | *toothache,¬cavity*) = **P**(*catch* | ¬*cavity*)
- *Catch* is conditionally independent of *Toothache* given *Cavity*:
  - **P**(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)

Conditional independence

- *Catch* is conditionally independent of *Toothache* given *Cavity*:
  - **P**(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)
- Equivalent statements:
  - **P**(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)
  - **P**(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Conditional independence

- Write out full joint distribution using chain rule:

  P(*Toothache, Catch, Cavity*)

  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch, Cavity*)

  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch* | *Cavity*) **P**(*Cavity*)

  = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity)

  i.e., 2 + 2 + 1 = 5 independent numbers

- Conditional independence
  - often reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*
  - Is one of the most basic and robust form of knowledge about uncertain environments

Conditional Independence

- *X* is conditionally independent of Y given *Z* (written I(X,Z,Y)) if the probability distribution governing *X* is independent of the value *of Y* given the value of *Z*:

- $P(X \mid Y, Z) = P(X \mid Z)$ that is,

$$(\forall x_i, y_j, z_k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Independence is symmetric: I(X Y Z)=I(Z,Y,X)

- Assume: $P(X|Y, Z) = P(X|Y)$
- X and Z are independent given Y

$$P(Z \mid X,Y) = \frac{P(X,Y \mid Z)P(Z)}{P(X,Y)} \quad \text{(Bayes' s Rule)}$$

$$= \frac{P(Y \mid Z)P(X \mid Y,Z)P(Z)}{P(X \mid Y)P(Y)} \quad \text{(Chain Rule)}$$

$$= \frac{P(Y \mid Z)P(X \mid Y)P(Z)}{P(X \mid Y)P(Y)} \quad \text{(By Assumption)}$$

$$= \frac{P(Y \mid Z)P(Z)}{P(Y)} = P(Z \mid Y) \quad \text{(Bayes' s Rule)}$$

Bayes Rule

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$P(cancer) = \qquad P(\neg cancer) =$$

$$P(+ \mid cancer) = \qquad P(- \mid cancer) =$$

$$P(+ \mid \neg cancer) = \qquad P(- \mid \neg cancer) =$$

Bayes Rule

Does patient have cancer or not?

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$

$$P(+ \mid cancer) = 0.98 \qquad P(- \mid cancer) = 0.02$$

$$P(+ \mid \neg cancer) = 0.03 \qquad P(- \mid \neg cancer) = 0.97$$

$$P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)};$$

$$P(\neg cancer|+) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)}$$

$$P(cancer|+)P(+) = 0.98 \times 0.008 = 0.0078;$$

$$P(\neg cancer|+)P(+) = 0.03 \times 0.992 = 0.0298$$

$$P(+) = 0.0078 + 0.0298$$

$$P(cancer \mid +) = 0.21; \qquad P(\neg cancer \mid +) = 0.79$$

The patient, more likely than not, does not have cancer

Bayes Rule

- Product rule
  - $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
  - Bayes' rule: $P(a \mid b) = P(b \mid a) P(a) / P(b)$
- In distribution form

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\, \mathbf{P}(Y) / \mathbf{P}(X) = \alpha \mathbf{P}(X|Y)\, \mathbf{P}(Y)$$

Probabilistic KR: The story so far

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- Independence and conditional independence provide the basis for compact representation of joint probability distributions
- Graph theory provides a basis for efficient computation
-

Building Probabilistic Models – Conditional Independence

- Random variable *X* is conditionally independent of *Y* given *Z* if the probability distribution governing X is independent of the value of Y given the value of *Z*:
- $P(X \mid Y, Z) = P(X \mid Z)$ that is, if

$$(\forall x_i, y_i, z_k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Conditional Independence

$$P(Thunder=1 \mid Rain=1, Lightning=1) = P(Thunder=1 \mid Lightening=1)$$
$$= P(Thunder=1 \mid Rain=0, Lightening=1)$$

$$P(Thunder=1 \mid Rain=1, Lightning=0) = P(Thunder=1 \mid Lightening=0)$$
$$= P(Thunder=1 \mid Rain=0, Lightening=0)$$

$$P(Thunder=0 \mid Rain=1, Lightning=1) = P(Thunder=0 \mid Lightening=1)$$
$$= P(Thunder=0 \mid Rain=0, Lightening=1)$$

$$P(Thunder=0 \mid Rain=1, Lightning=0) = P(Thunder=0 \mid Lightening=0)$$
$$= P(Thunder=0 \mid Rain=0, Lightening=0)$$

Bayesian Networks

$S \in \{no, light, heavy\}$

P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

$C \in \{none, benign, malignant\}$

Smoking=	no	light	heavy
P(C=none)	0.96	0.88	0.60
P(C=benign)	0.03	0.08	0.25
P(C=malig)	0.01	0.04	0.15

Product Rule

- $P(C,S) = P(C|S)\,P(S)$

$S\Downarrow$ $C\Rightarrow$	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

Marginalization

$S \Downarrow \quad C \Rightarrow$	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

Row totals: *P(Smoke)*

Column totals: *P(Cancer)*

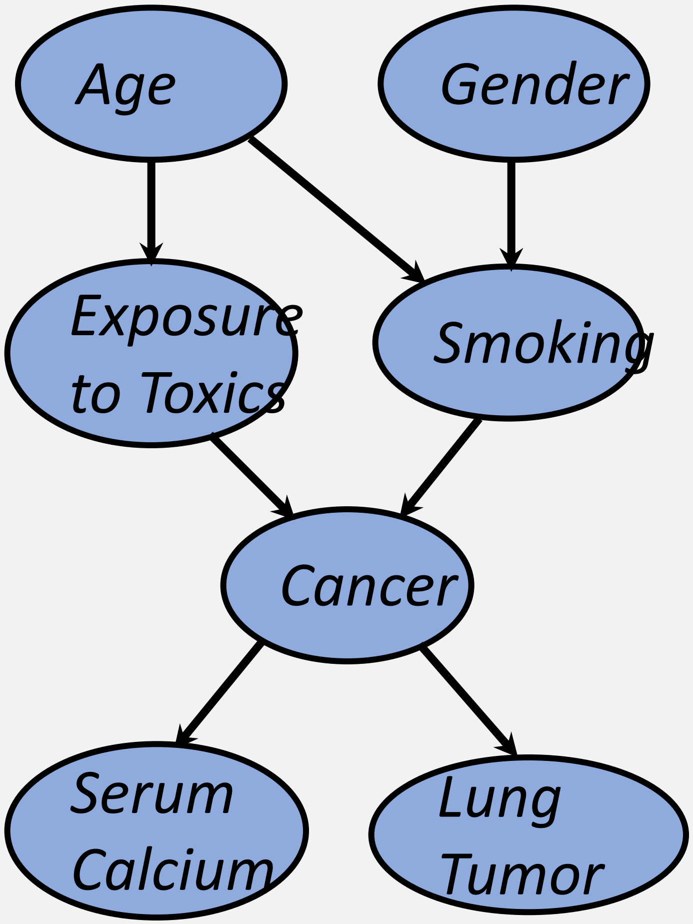
Bayes Rule Revisited

$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C)} = \frac{P(C,S)}{P(C)}$$

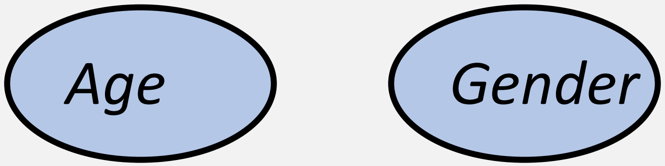
$S \Downarrow \quad C \Rightarrow$	$none$	$benign$	$malig$
no	0.768/.935	0.024/.046	0.008/.019
$light$	0.132/.935	0.012/.046	0.006/.019
$heavy$	0.030/.935	0.015/.046	0.005/.019

Cancer=	*none*	*benign*	*malignant*
P(S=no)	0.821	0.522	0.421
P(S=light)	0.141	0.261	0.316
P(S=heavy)	0.037	0.217	0.263

A Bayesian Network



Independence



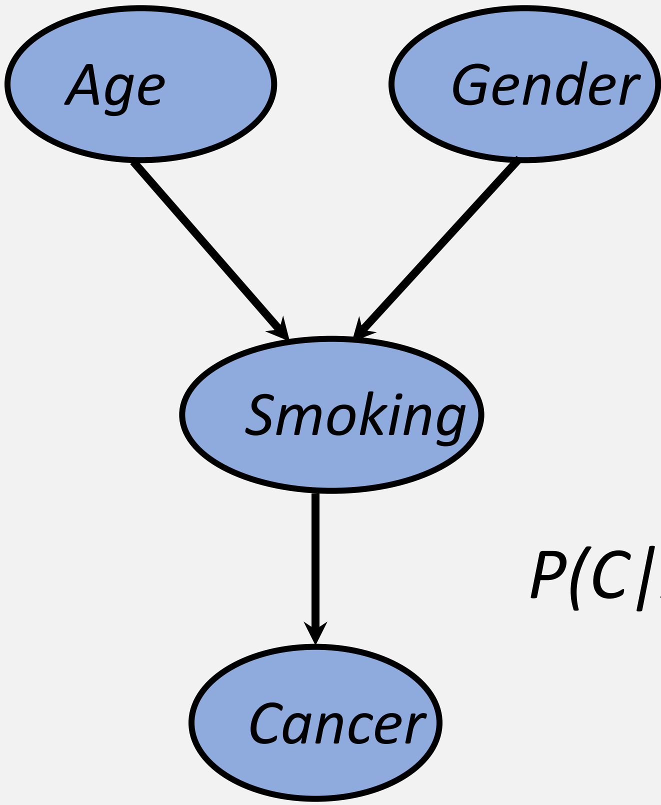
Age and *Gender* are independent.

$$P(A,G) = P(G)P(A)$$

$$P(A|G) = P(A) \quad A \perp G$$
$$P(G|A) = P(G) \quad G \perp A$$

$$P(A,G) = P(G|A)\,P(A) = P(G)P(A)$$
$$P(A,G) = P(A|G)\,P(G) = P(A)P(G)$$

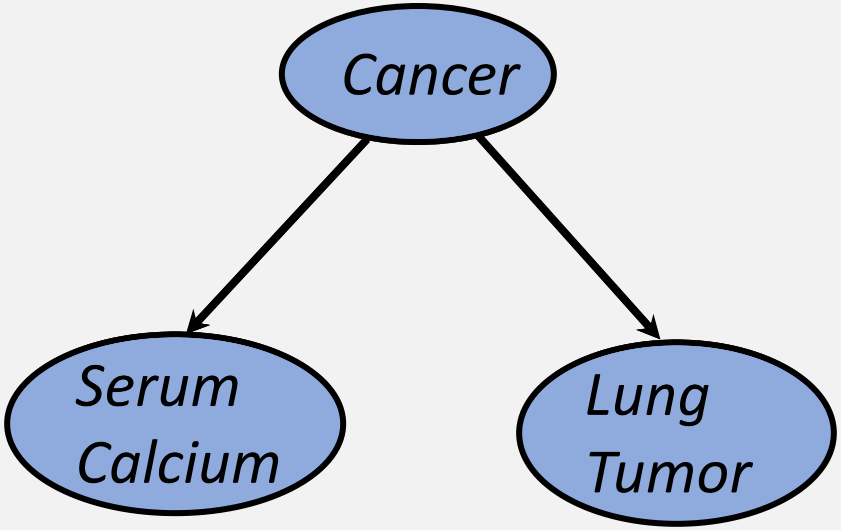
Conditional Independence



Cancer is independent of *Age* and *Gender* given *Smoking*.

$$P(C|A,G,S) = P(C|S) \quad C \perp A,G \mid S$$

More Conditional Independence: Naïve Bayes



Serum Calcium and *Lung Tumor* are dependent

Serum Calcium is independent of *Lung Tumor*, given *Cancer*

$$P(L|SC,C) = P(L|C)$$

Probabilistic Graphical Models

- The Probabilistic graphical models e.g., Bayes networks, explicitly model conditional independence among subsets of variables to yield a graphical representation of probability distributions that admit such independence

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{Pa_i})$$

$\boldsymbol{Pa_i} = parents(X_i)$

Bayesian network

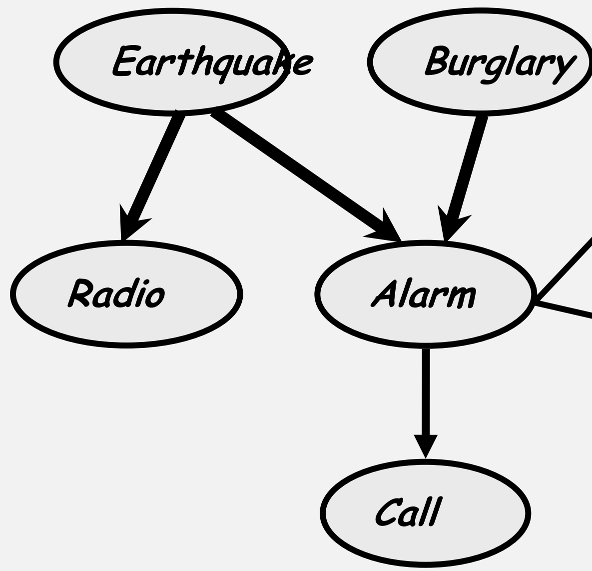
- Bayesian network is a directed acyclic graph (DAG) in which the nodes represent random variables
- Each node is annotated with a probability distribution $P\,(X_i \mid Parents(X_i)\,)$ representing the dependency of that node on its parents in the DAG
- Each node is asserted to be conditionally independent of its non-descendants, given its immediate predecessors
- Arcs represent direct dependencies

Conditional Independence

- *X* is conditionally independent of Y given *Z* if the probability distribution governing *X* is independent of the value *of Y* given the value of *Z*:
- $P(X \mid Y, Z) = P(X \mid Z)$ that is,

$$(\forall x_i, y_j, z_k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Bayesian Networks



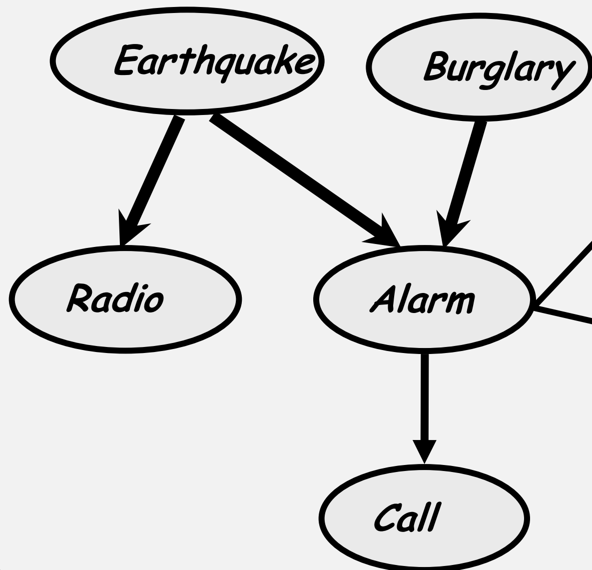
E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

Bayesian Networks

- Qualitative part
  statistical independence statements represented in the form of a directed acyclic graph (DAG)
  - Nodes - random variables
  - Edges – direct influence

Quantitative part

Conditional probability distributions – one for each random variable conditioned on its parents



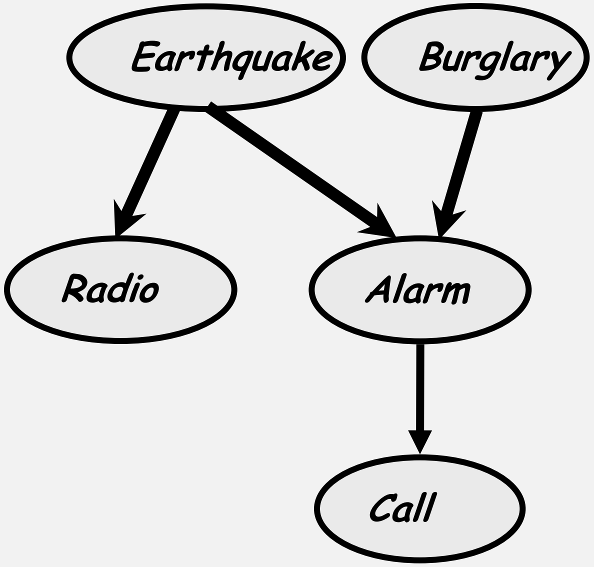
E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

Efficient factorized representation of probability distributions via conditional independence

- Nodes are independent of non-descendants given their parents

d-separation:

- a graph theoretic criterion for checking implicit independence assertions
- can be computed in linear time (in the number of edges)



What independences does a Bayes Net model?

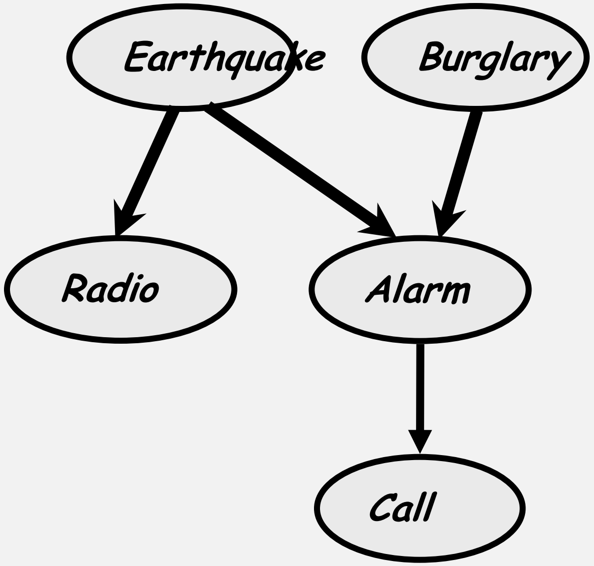
- In order for a Bayesian network to model a probability distribution, the following must be true by definition:
- Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.

This implies

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

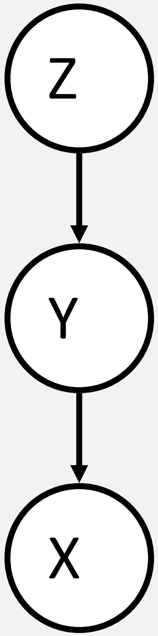
$$P(E,B,R,A,C) =$$
$$P(E)P(B)P(R \mid E)P(A \mid E,B)P(C \mid A)$$

But what else does it imply?



What Independences does a Bayes Network model?

Example:



Given Y, does learning the value of Z tell us nothing new about X?

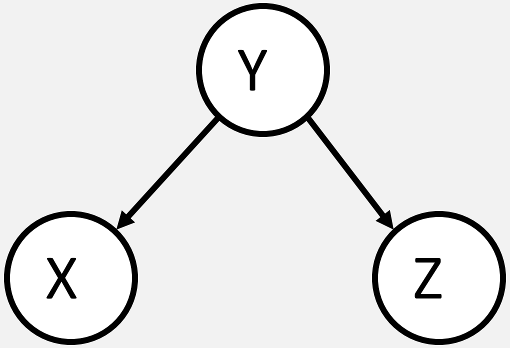
i.e., is $P(X|Y, Z)$ equal to $P(X \mid Y)$?

Yes. Since we know the value of all of X's parents (namely, Y), and Z is not a descendant of X, X is conditionally independent of Z.

Also, since independence is symmetric, $P(Z|Y, X) = P(Z|Y)$.

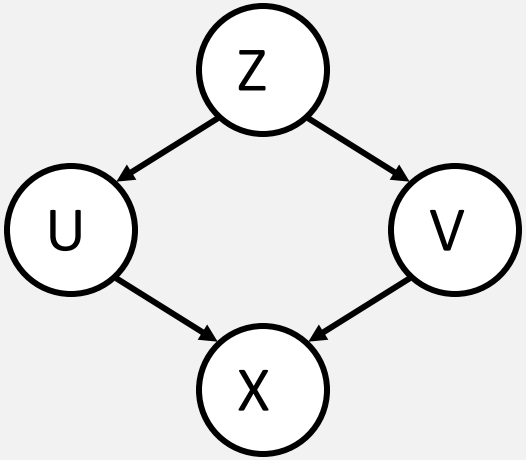
What Independences does a Bayes Network model?

- Let $I(X,Y,Z)$ represent X and Z being conditionally independent given Y.



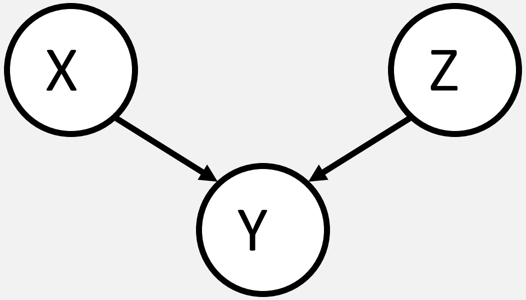
- $I(X,Y,Z)$? Yes, just as in previous example: All X's parents given, and Z is not a descendant.

What Independences does a Bayes Network model?



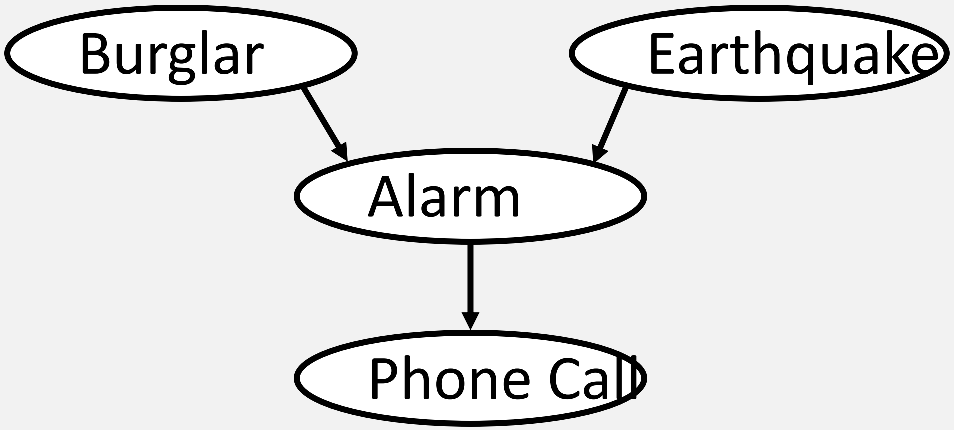
- $I(X, \{U\}, Z)$? No.
- $I(X, \{U, V\}, Z)$? Yes.

Dependency induced by V-structures

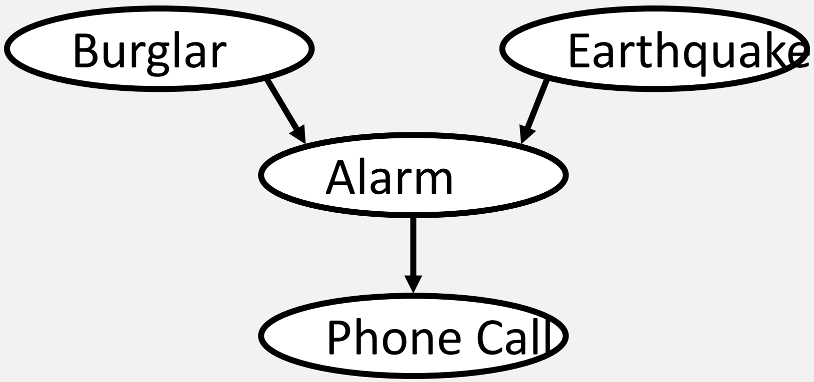


- X has no parents, so we know all its parents' values trivially
- Z is not a descendant of X
- So, $I(X,\{\},Z)$, even though there is a undirected path from X to Z through an unknown variable Y.
- What if we do know the value of Y ? Or one of its descendants?

The Burglar Alarm example



- Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.
- Earth arguably doesn't care whether your house is currently being burgled
- While you are on vacation, one of your neighbors calls and tells you your home's burglar alarm is ringing.



- But now suppose you learn that there was a medium-sized earthquake in your neighborhood. …Probably not a burglar after all.
- Earthquake "explains away" the hypothetical burglar.
- But then it must NOT be the case that

$I(\text{Burglar}, \{\text{Phone Call}\}, \text{Earthquake})$,

even though $I(\text{Burglar}, \{\}, \text{Earthquake})$!

d-separation

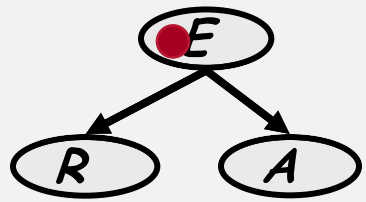
- Fortunately, there is a relatively simple algorithm for determining whether two variables in a Bayesian network are conditionally independent given some other variables:
  - *d-separation*.
- Two variables are independent if all paths between them are blocked by evidence
- Three cases:
  - Common cause
  - Intermediate cause
  - Common Effect

d-separation

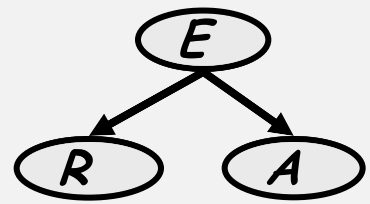
- Two variables are independent if all paths between them are blocked by evidence
- Three cases:
  - Common cause
  - Intermediate cause
  - Common Effect

Evidence may be transmitted through a diverging connection unless it is instantiated.

Blocked



Unblocked



- If we do not know whether an earthquake occurred, then radio announcement can influence our belief about the alarm having gone off.
- If we know that earthquake occurred, then radio announcement gives no information about the alarm

d-separation

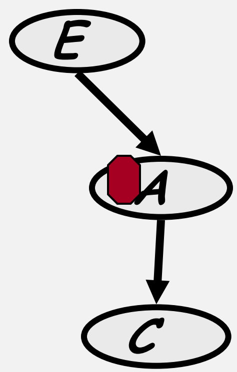
Blocked Unblocked

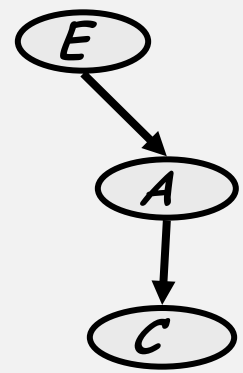
Blocked Unblocked

Common cause

Intermediate cause

Common Effect





Information may be transmitted through a serial connection unless it is blocked (value set)

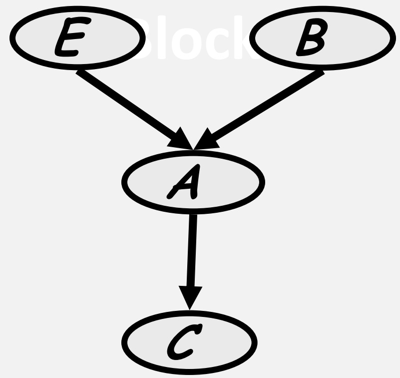
d-separation

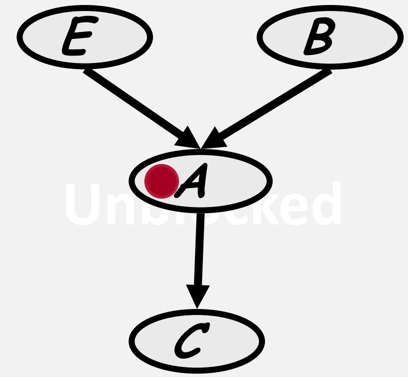
Blocked | Unblocked

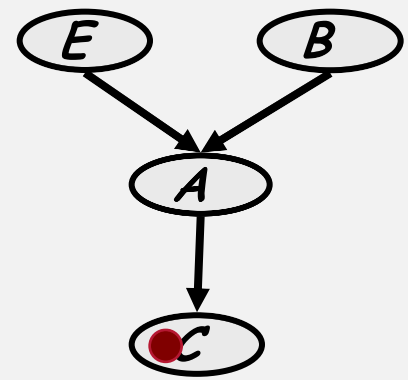
Common cause

Intermediate cause

Common Effect







Information may be transmitted through a converging connection only if either the variable or one of its descendants has been set

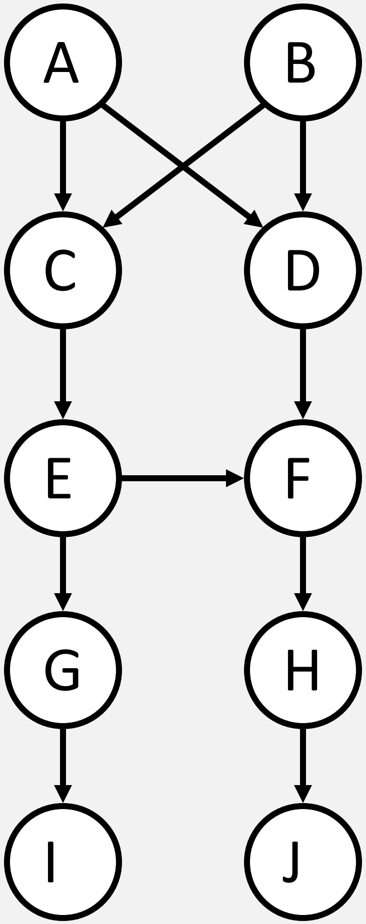
d-separation

- Definition: *X* and *Z* are *d-separated* by a set of evidence variables *E* iff every undirected path from *X* to *Z* is "blocked" by evidence *E*

d-separation

- Theorem [Verma & Pearl, 1998]: If a set of evidence variables *E d*-separates *X* and *Z* in a Bayesian network's graph, then $I(X, E, Z)$.
- *d*-separation can be computed in linear time using a depth-first search like algorithm.
- We now have a fast algorithm for automatically inferring whether finding out about the value of one variable might give us any additional hints about some other variable, given what we already know.
- *d*-separation of *X* and *Z* by *E* is sufficient for asserting $I(X, E, Z)$, but not necessary.
  - Variables may actually be independent when they are not *d*-separated, depending on the actual probabilities involved

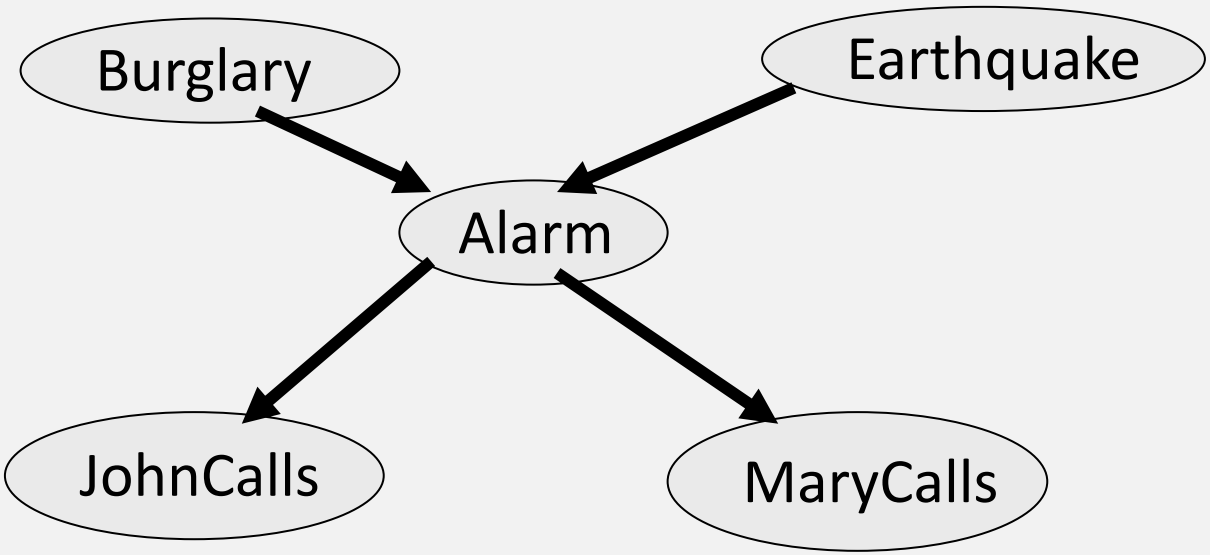
d-separation



I(C, {}, D)?
I(C, {A}, D)?
I(C, {A, B}, D)?
I(C, {A, B, J}, D)?

Markov Blanket

- A node is conditionally independent of all other nodes in the network given its parents, children, and children’s parents -



Burglary is independent of John Calls and Mary Calls given Alarm and Earth Quake

Bayesian Networks: Summary

- Bayesian networks offer an efficient representation of probability distributions
- Efficient:
  - Local models
  - Independence (*d*-separation)
- Effective: Algorithms take advantage of structure to
  - Compute posterior probabilities
  - Compute most probable instantiation
  - Decision making

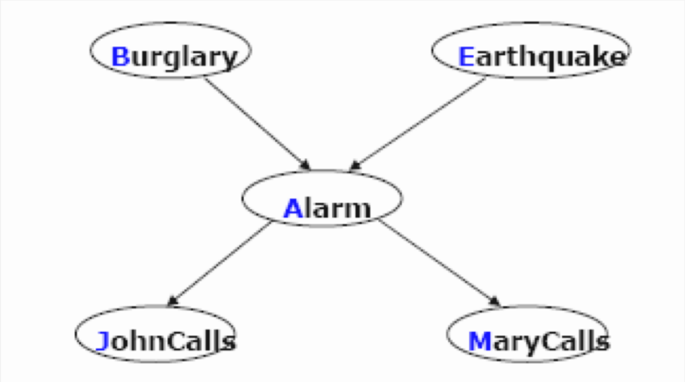
Inference in Bayesian network

Bad news:

- – Exact inference problem in BNs is NP-hard (Cooper)
- – Approximate inference is NP-hard (Dagum, Luby)

In practice, things are not so bad

- Exact inference
  - Inference in Simple Chains
  - Variable elimination
  - Clustering / join tree algorithms
- Approximate inference
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods
  - Mean field theory



Computing joint probability distributions using a Bayesian network

- Any entry in the joint probability distribution can be calculated from the Bayesian network.
- We're just using the chain rule and conditional independence.

$$
\begin{aligned}
P(J, M, A, \neg B, \neg E) &= P(J \mid M, A, \neg B, \neg E) P(M, A, \neg B, \neg E) \\
&= P(J \mid A) P(M \mid A, \neg B, \neg E) P(A, \neg B, \neg E) \\
&= P(J \mid A) P(M \mid A) P(A \mid \neg B, \neg E) P(\neg B, \neg E) \\
&= P(J \mid A) P(M \mid A) P(A \mid \neg B, \neg E) P(\neg B) P(\neg E)
\end{aligned}
$$

Computing joint probabilities

General formula:

$$P(X_1,\ldots,X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid Parents(X_i))$$

- Joint distribution can be used to answer any query about the domain.
- Bayesian network represents the joint distribution
- Any query about the domain can be answered using a BN
- Tradeoff: A BN can be much more concise, but you need to calculate, rather than look up in a table, probabilities from the joint distribution

Inference in Bayesian Networks

- Bayesian networks are a compact encoding of the full joint probability distribution over *N* variables that makes conditional independence assumptions between these variables explicit.
- We can use Bayesian networks to compute any probability of interest over the given variables.
- Now we look at Inference in more detail

Inference in Bayesian Networks

Find $P(Q=q|\boldsymbol{E}=e)$

- Q the query variable(s)
- $\boldsymbol{E}$ set of evidence variables

$$P(q|e) = P(q,e) / P(e)$$

$X_1,.. X_n$ are network variables except Q,E

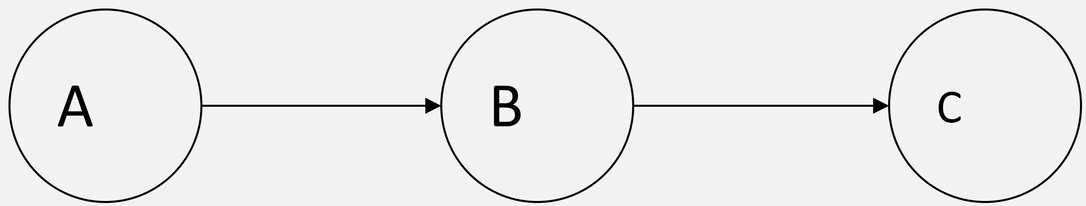
$$P(q,e) = \sum_{x_1,x_2 \ldots x_n} (q,e,X_1,X_2 \ldots X_n)$$

Basic Inference

$P(b) = ?$

$$P(b) = \sum_a P(a, b) = \sum_a P(b \mid a)\, P(a)$$

Basic Inference



$$P(b) = \sum_a \mathrm{P(a, b)} = \sum_a \mathrm{P(b \mid a)\, P(a)}$$

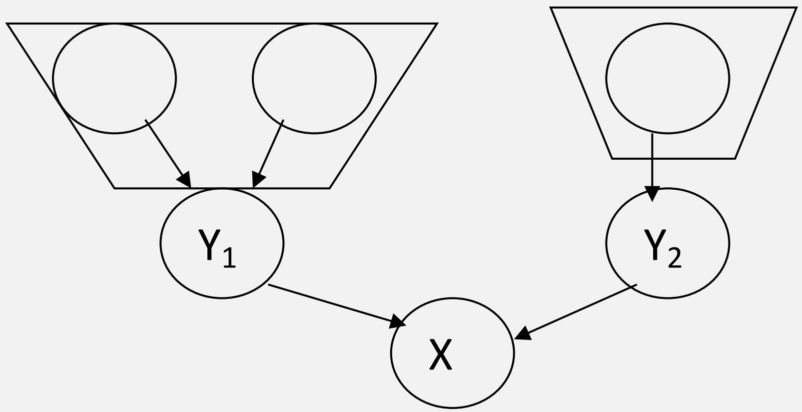
$$P(c) = \sum_b P(c \mid b) P(b)$$

$$P(c) = \sum_{a,b} P(a,b,c) = \sum_{a,b} P(c \mid b,a) P(b \mid a) P(a)$$

$$= \sum_{a,b} P(c \mid b) P(b \mid a) P(a)$$

$$= \sum_{a,b} P(c \mid b) P(b)$$

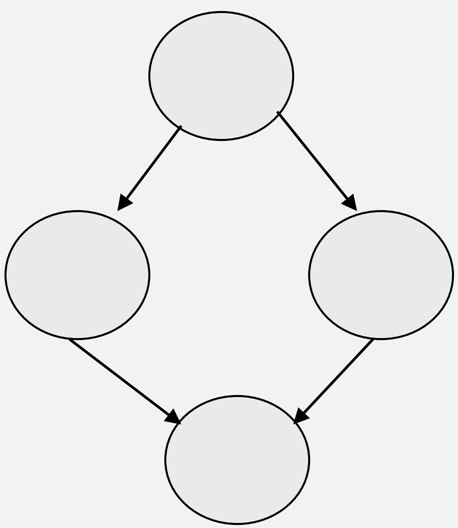
Inference in trees



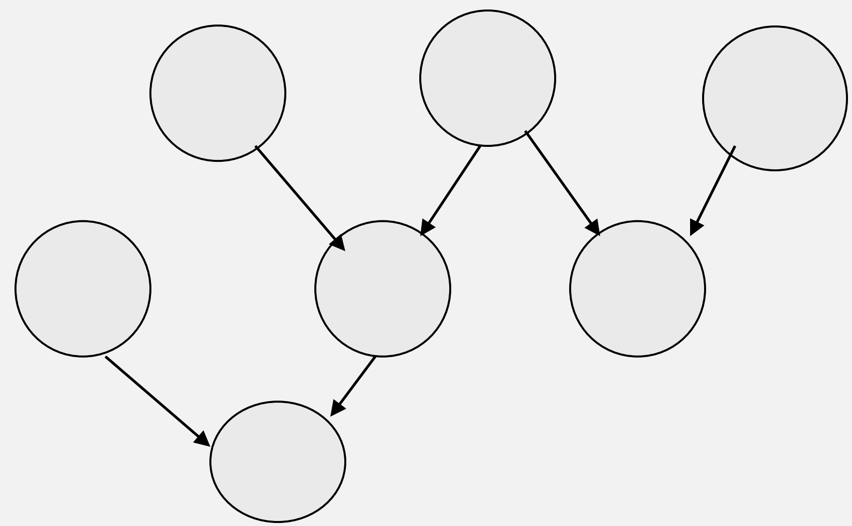
$$P(X)=\sum_{y_1,y_2}P(X,Y_1,Y_2)=\sum_{y_1,y_2}P(X\mid Y_1,Y_2)P(Y_1,Y_2)=\sum_{y_1,y_2}P(X\mid Y_1,Y_2)P(Y_1)P(Y_2)$$

Polytrees

- A network is *singly connected* (*a polytree*) if it contains no undirected loops.



Not a polytree



Polytree

Inference in polytrees

- Theorem: Inference in polytrees can be performed in time that is polynomial in the number of variables.
- Main idea: in variable elimination, need only maintain distributions over single nodes at any step.

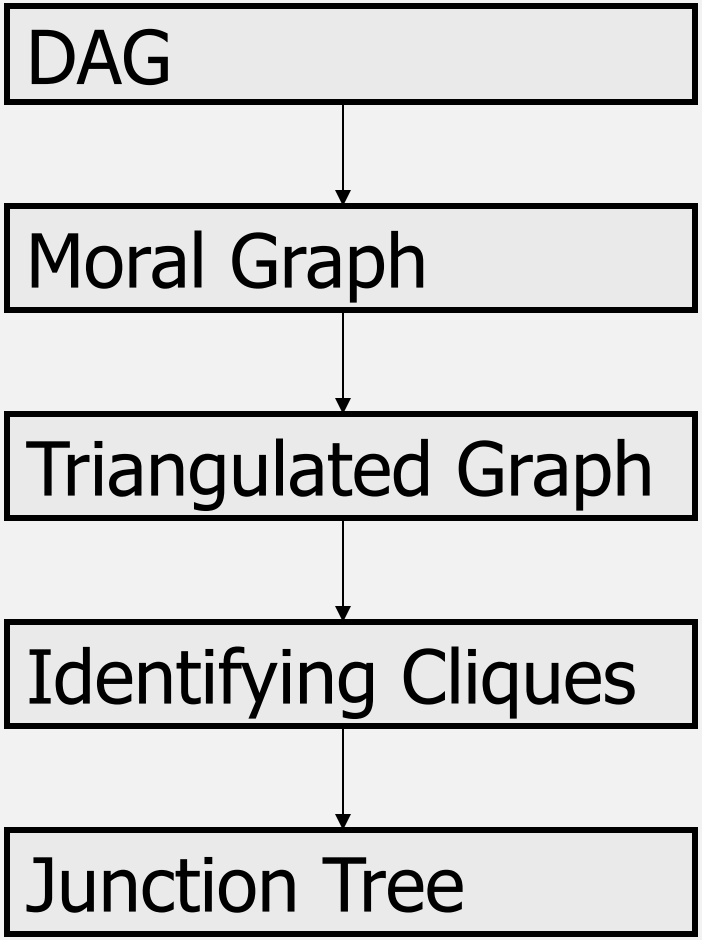
Inference with Bayesian Networks

- Inference in polytrees can be performed efficiently
- Inference with DAG is NP-Hard
  - Proof by reduction of SAT to Bayesian network inference

Approaches to inference

- Exact inference
  - Inference in Simple Chains
  - Variable elimination
  - Clustering / join tree algorithms
- Approximate inference
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods
  - Mean field theory

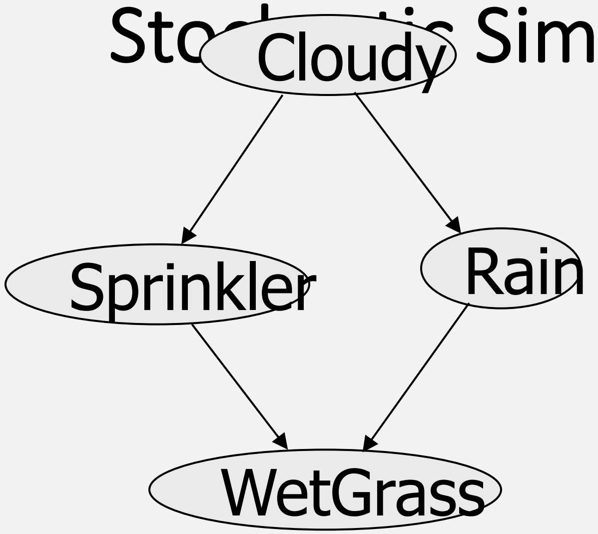
Building Junction Trees



Approximate Inference: Stochastic simulation

- Suppose you are given values for some subset of the variables, G, and want to infer values for unknown variables, U
- Randomly generate a very large number of instantiations from the BN
  - Generate instantiations for **all** variables – start at root variables and work your way "forward"
- Only keep those instantiations that are consistent with the values for G
- Use the frequency of values for U to get estimated probabilities
- Accuracy of the results depends on the size of the sample (asymptotically approaches exact results)

Stochastic Simulation



P(WetGrass|Cloudy)?

P(WetGrass|Cloudy)
= P(WetGrass, Cloudy) / P(Cloudy)

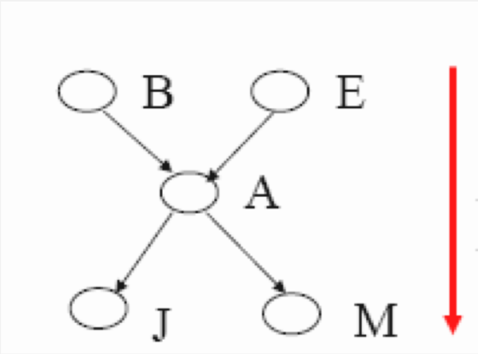
1. Draw N samples from the BN by repeating 1.1 and 1.2
   1.1. Guess Cloudy at random according to P(Cloudy)
   1.2. For each guess of Cloudy, guess Sprinkler and Rain, then WetGrass
2. Compute the ratio of the # runs where WetGrass and Cloudy are True over the # runs where Cloudy is True

Stochastic simulation

- The probability is approximated using sample frequencies

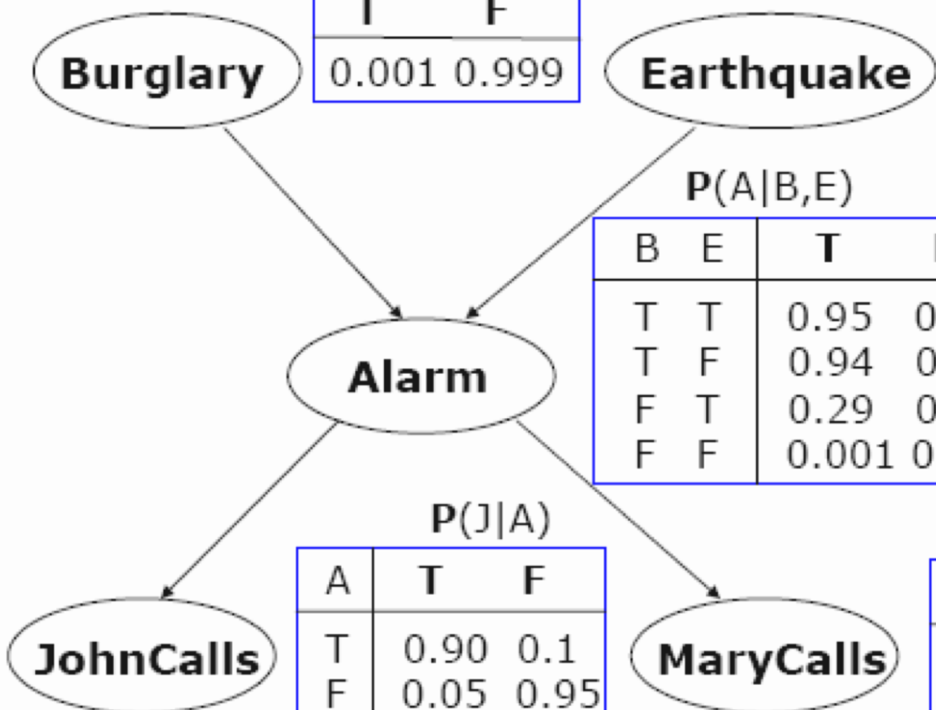
BN sampling:

- Generate sample in a top down manner, following the links in BN
- A sample is an assignment of values to all variables



BN Sampling Example

Goal: To infer $P(B \mid J = T, M = F)$



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

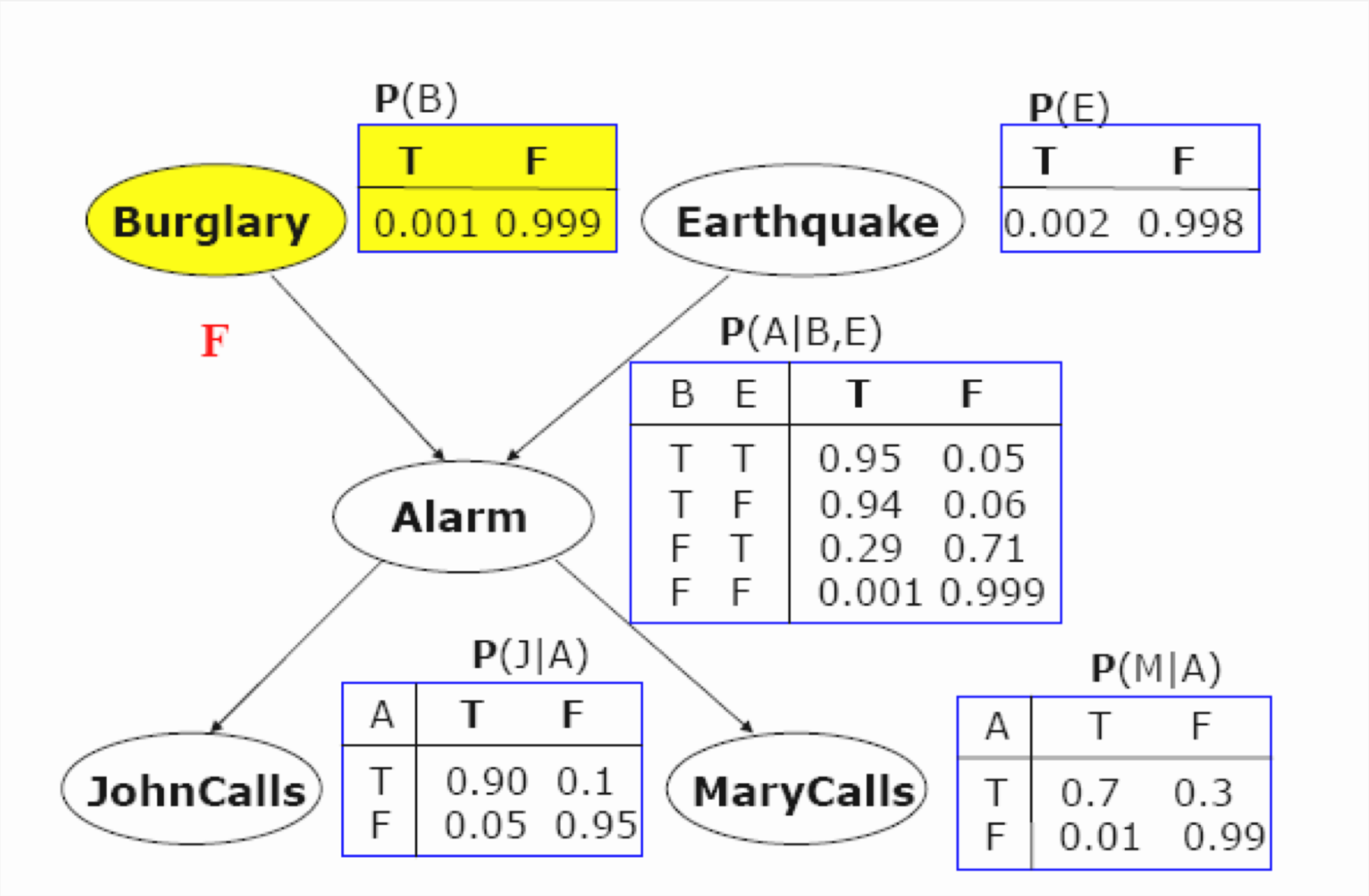
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

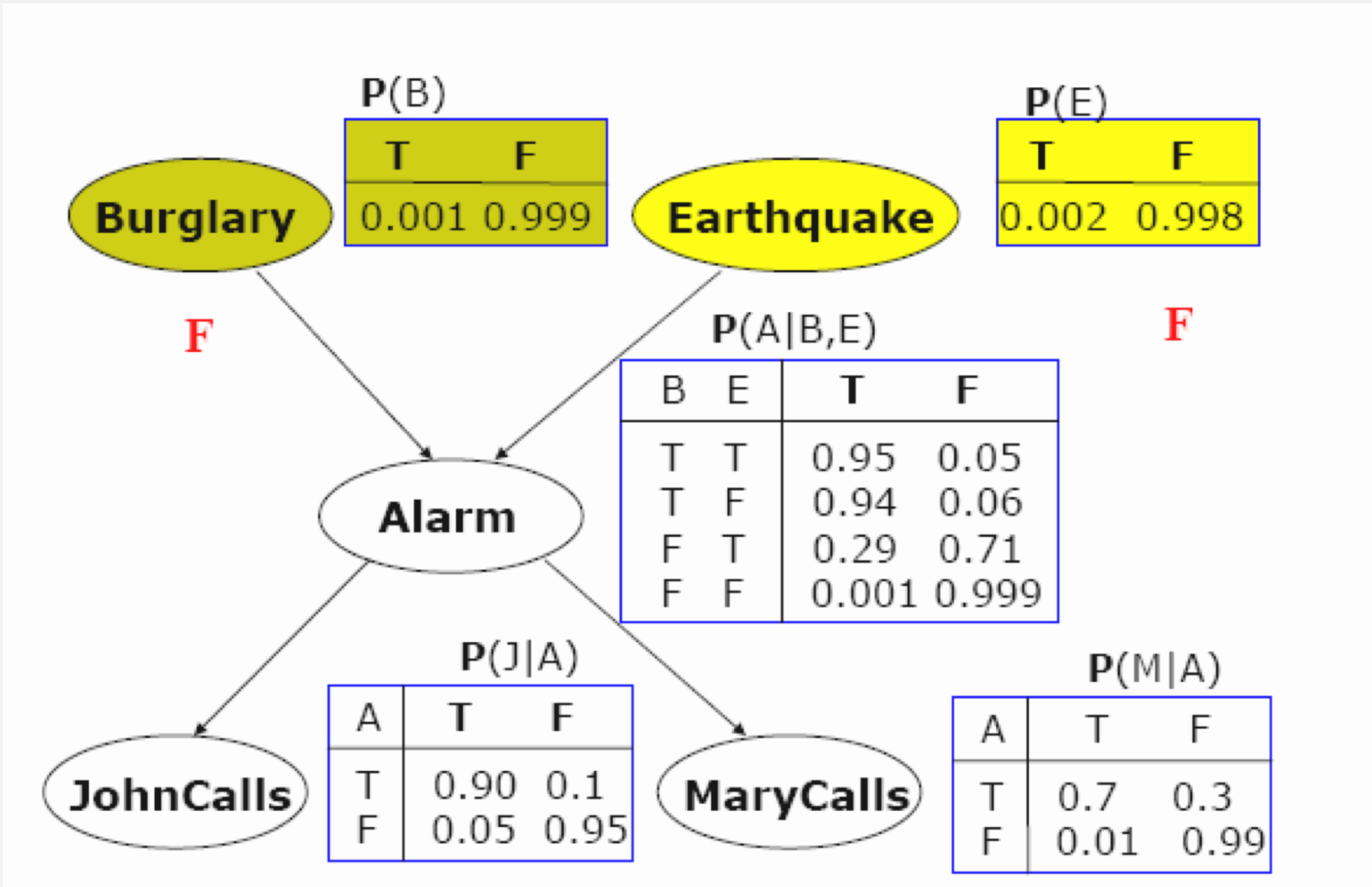
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

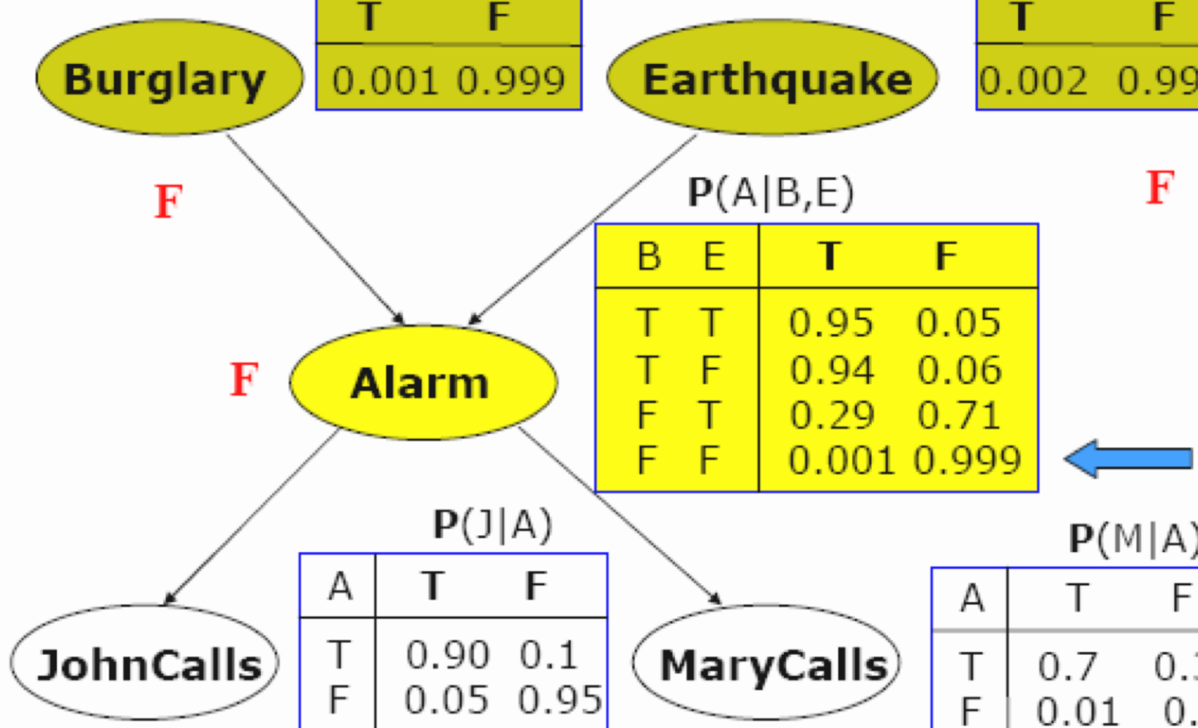
BN Sampling Example



BN Sampling Example



BN Sampling Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

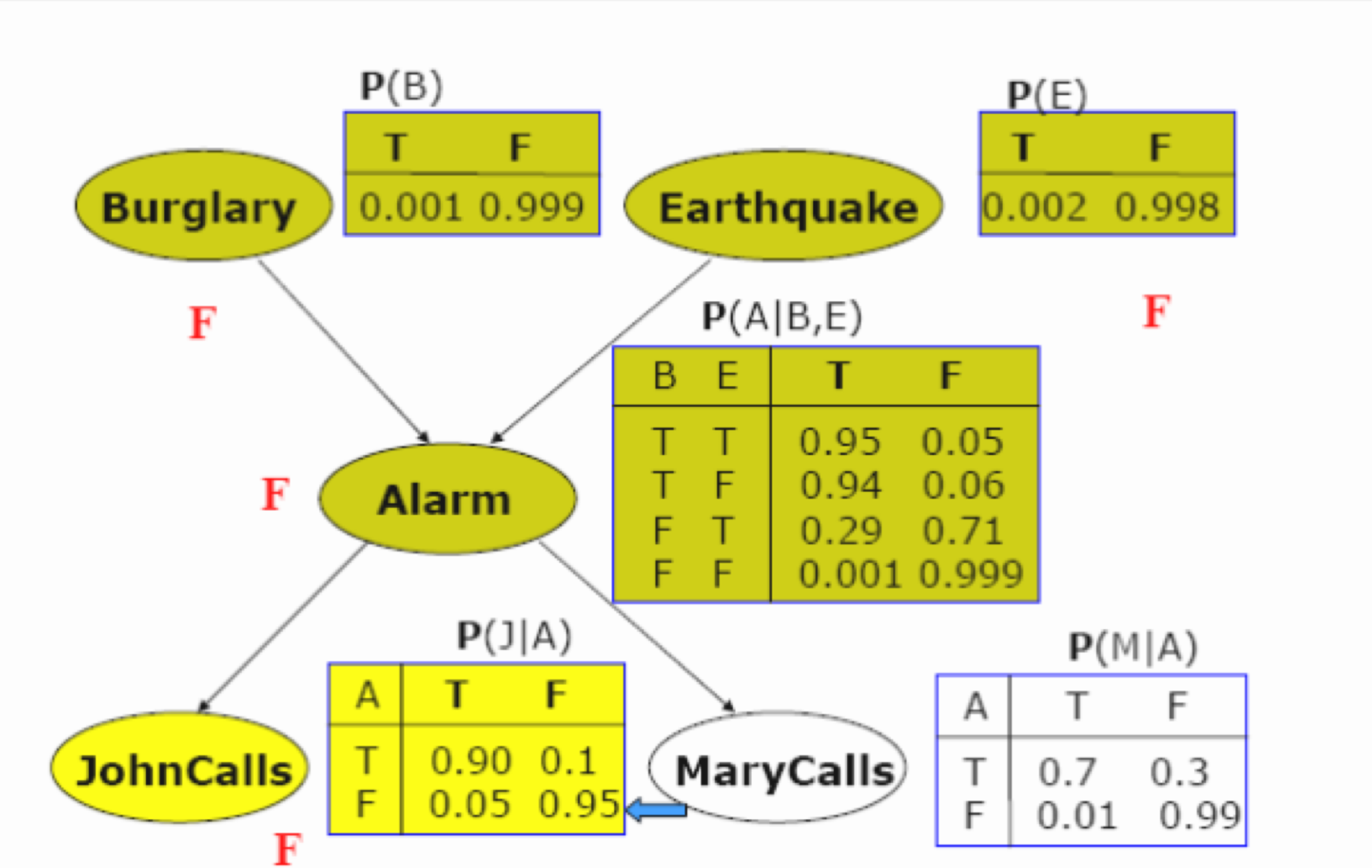
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

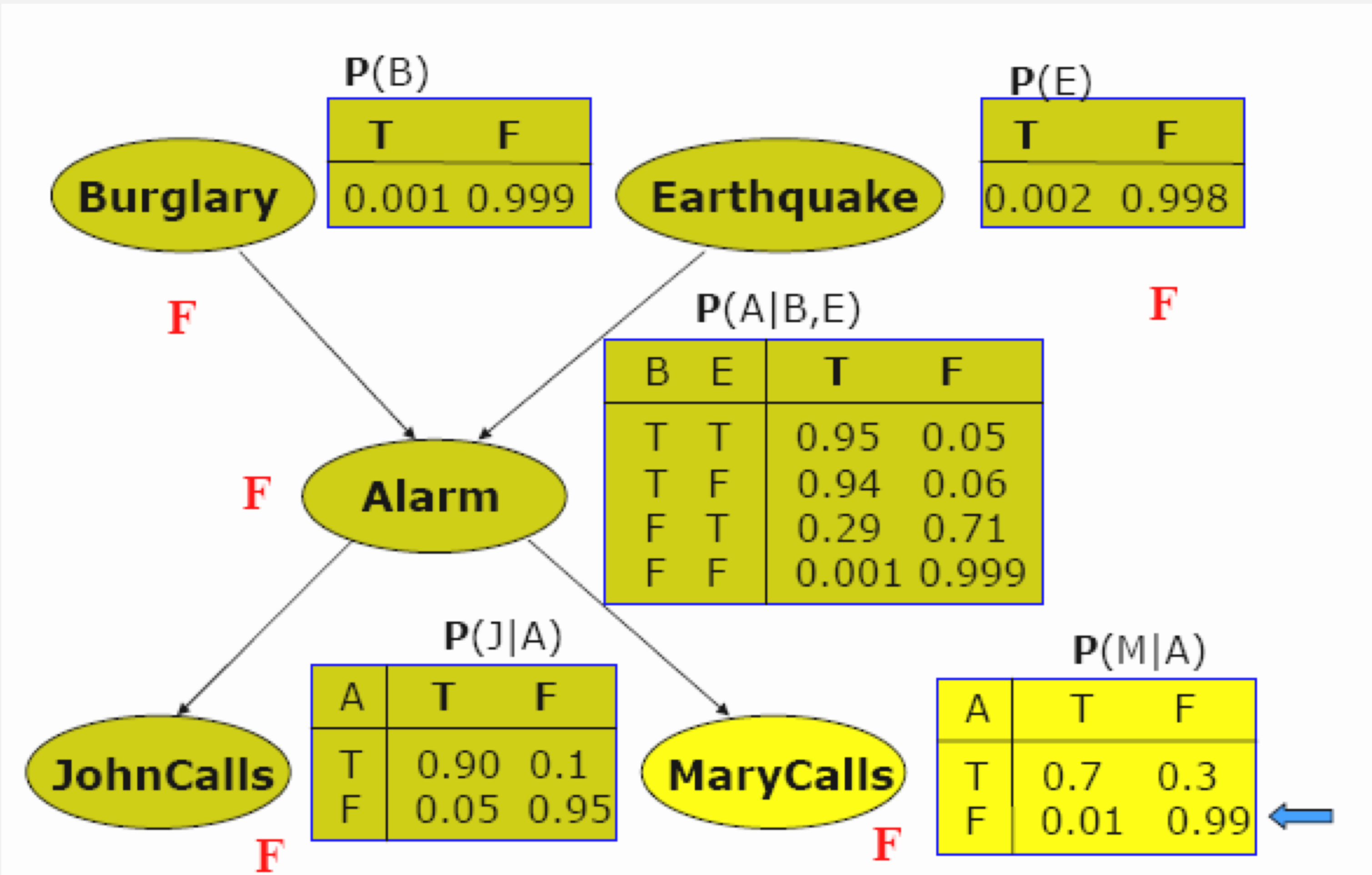
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BN Sampling Example



BN Sampling Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Rejection Sampling

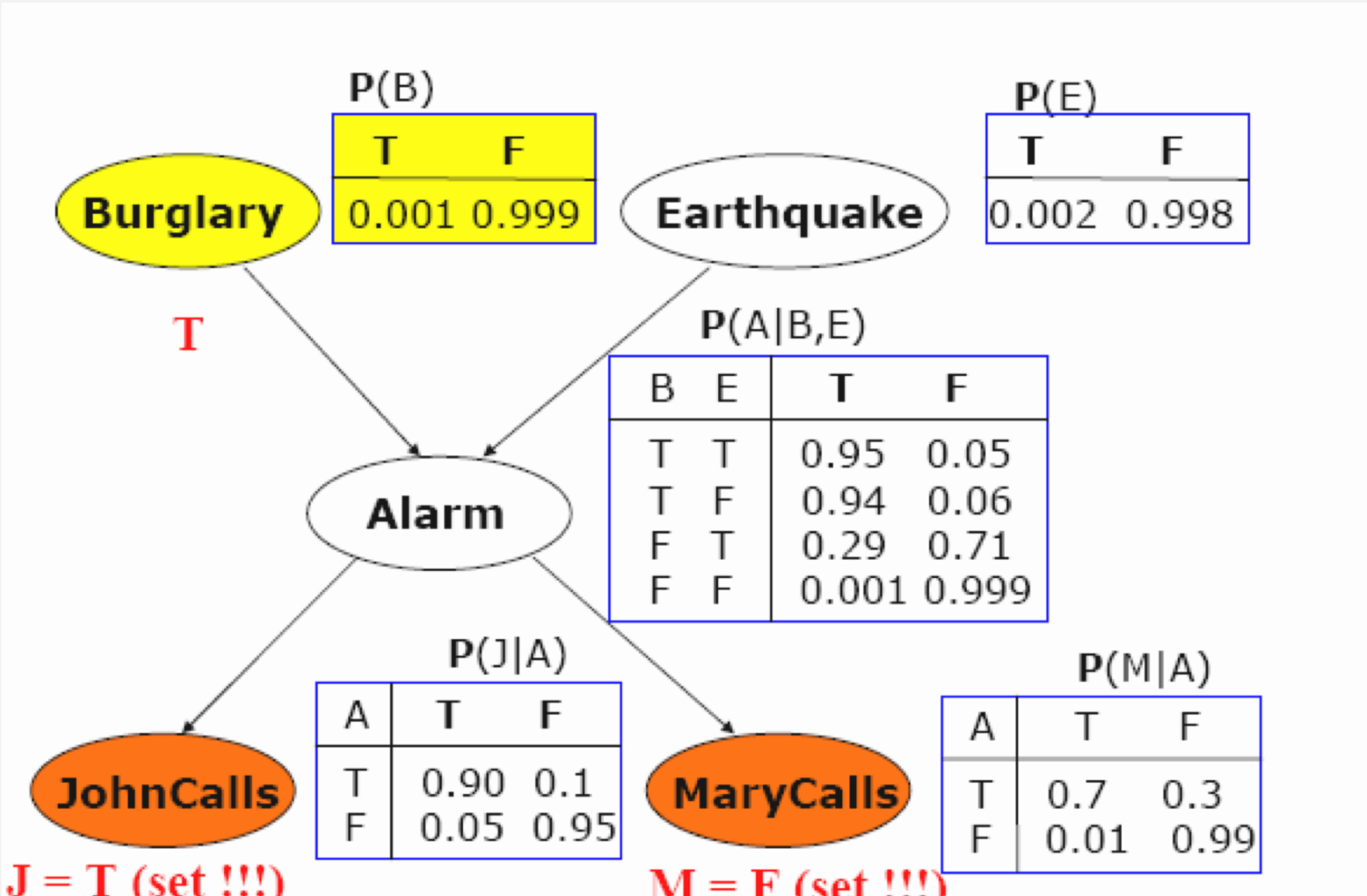
Rejection sampling:

- Generate sample for the full joint by sampling BN
- Use only samples that agree with the condition, the remaining samples are rejected
- Problem: many samples can be rejected

Likelihood weighting

- Avoids inefficiencies of rejection sampling
- Idea: generate only samples consistent with an evidence (or conditioning event)
- If the value is set by evidence, there is no sampling
- Problem: using simple counts is not enough since these may occur with different probabilities
- Likelihood weighting: with every sample keep a weight with which it should count towards the estimate

Likelihood weighting Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

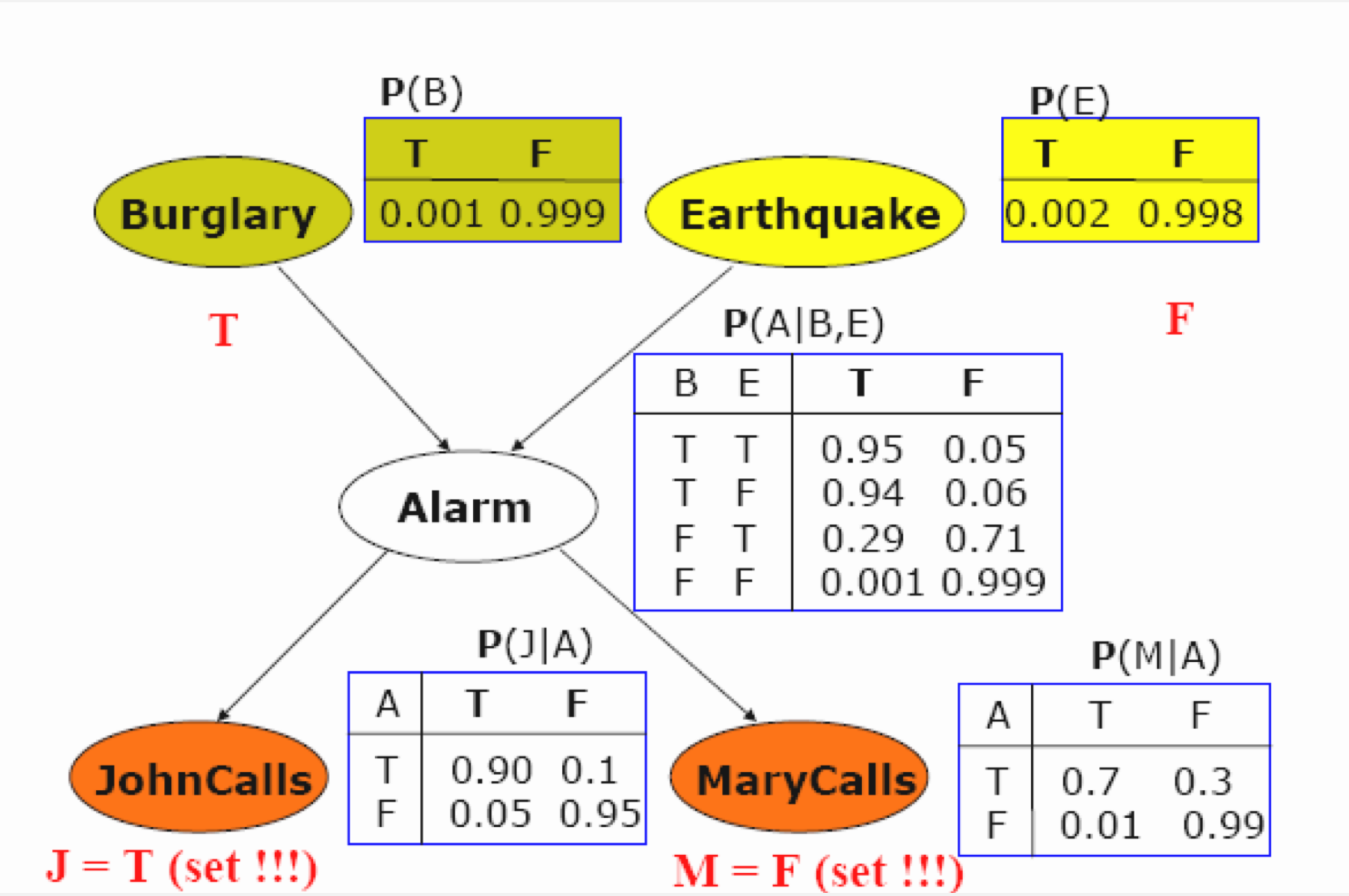
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Likelihood weighting Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

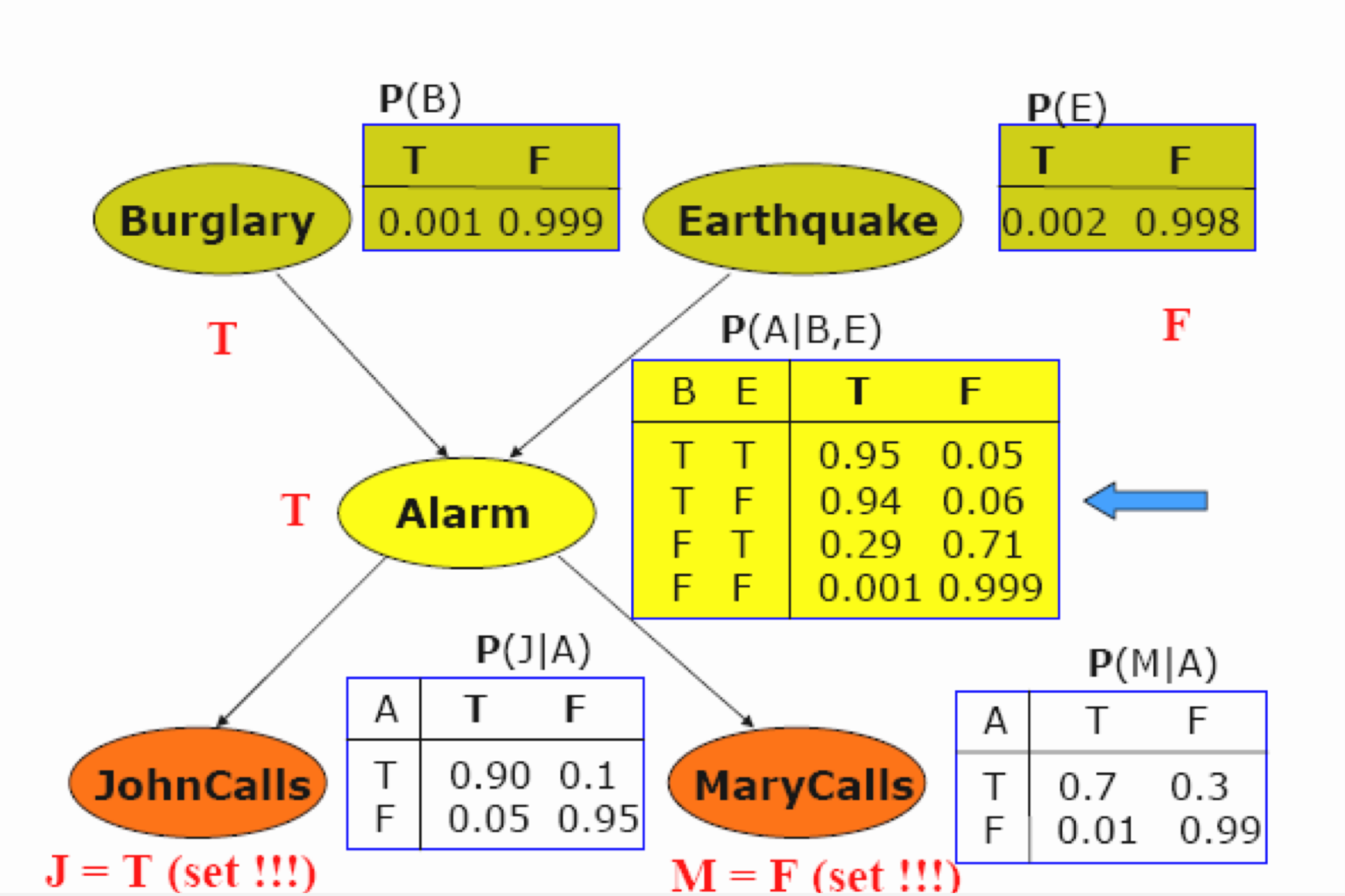
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

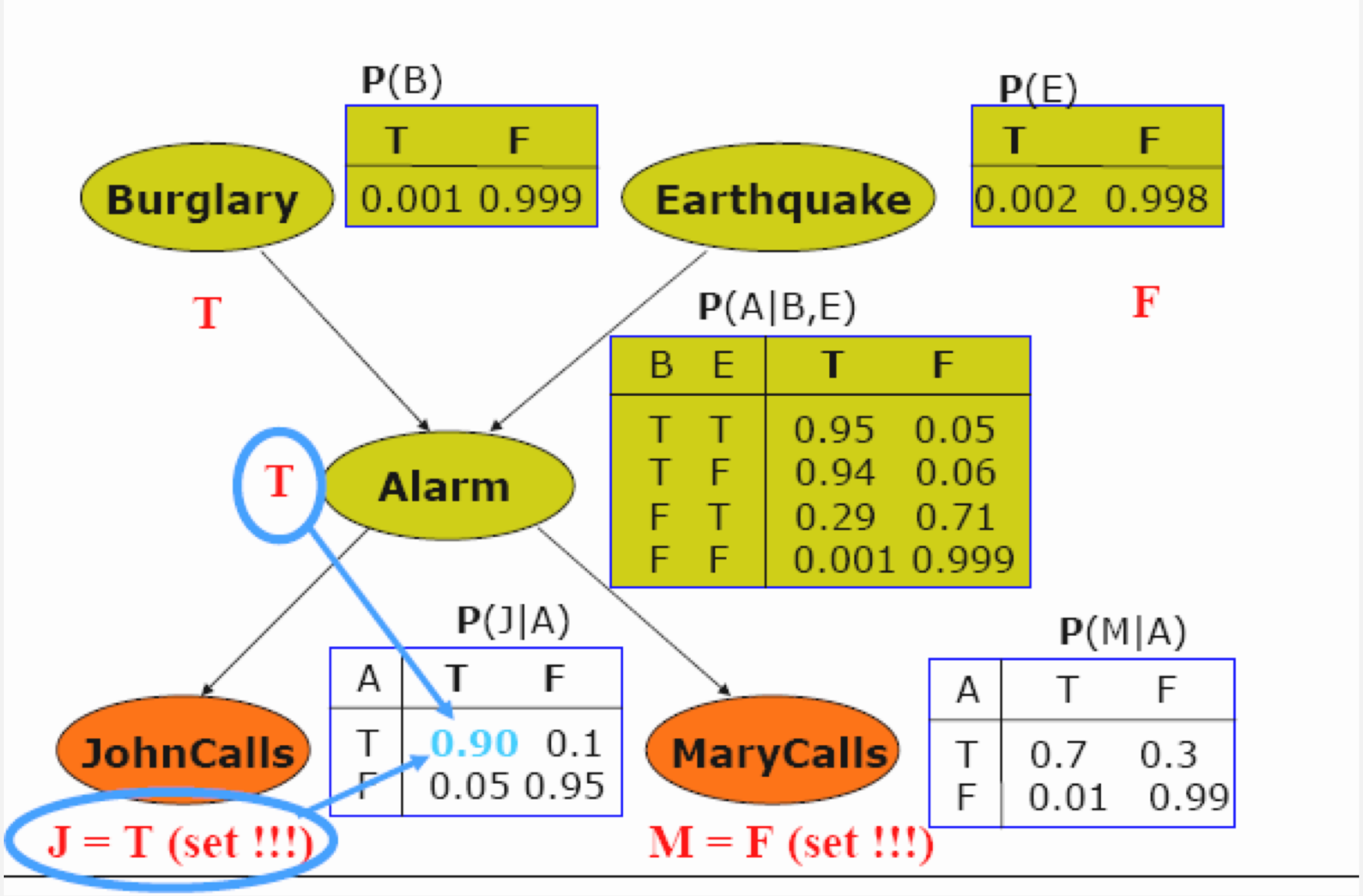
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

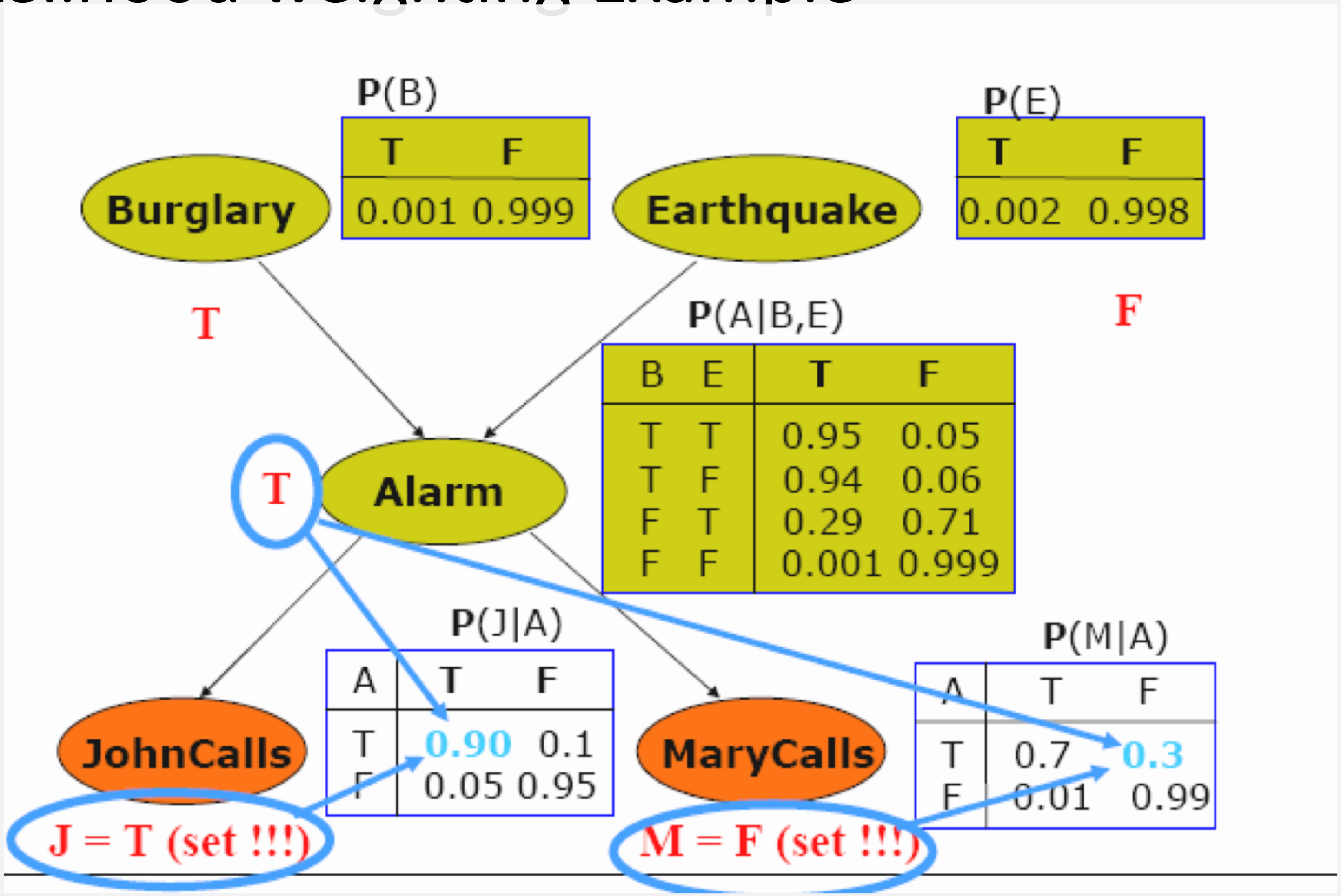
Likelihood weighting Example



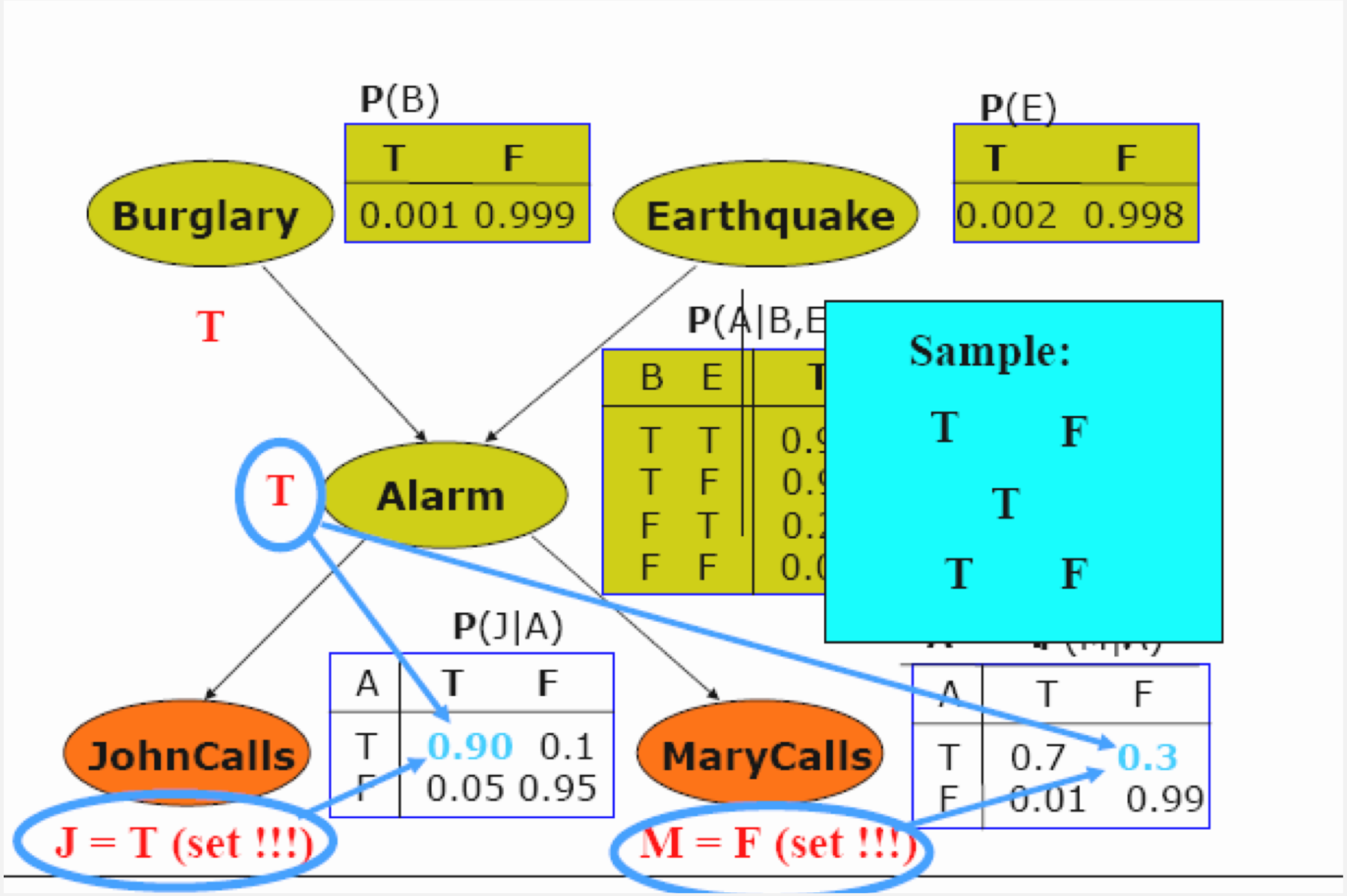
Likelihood weighting Example



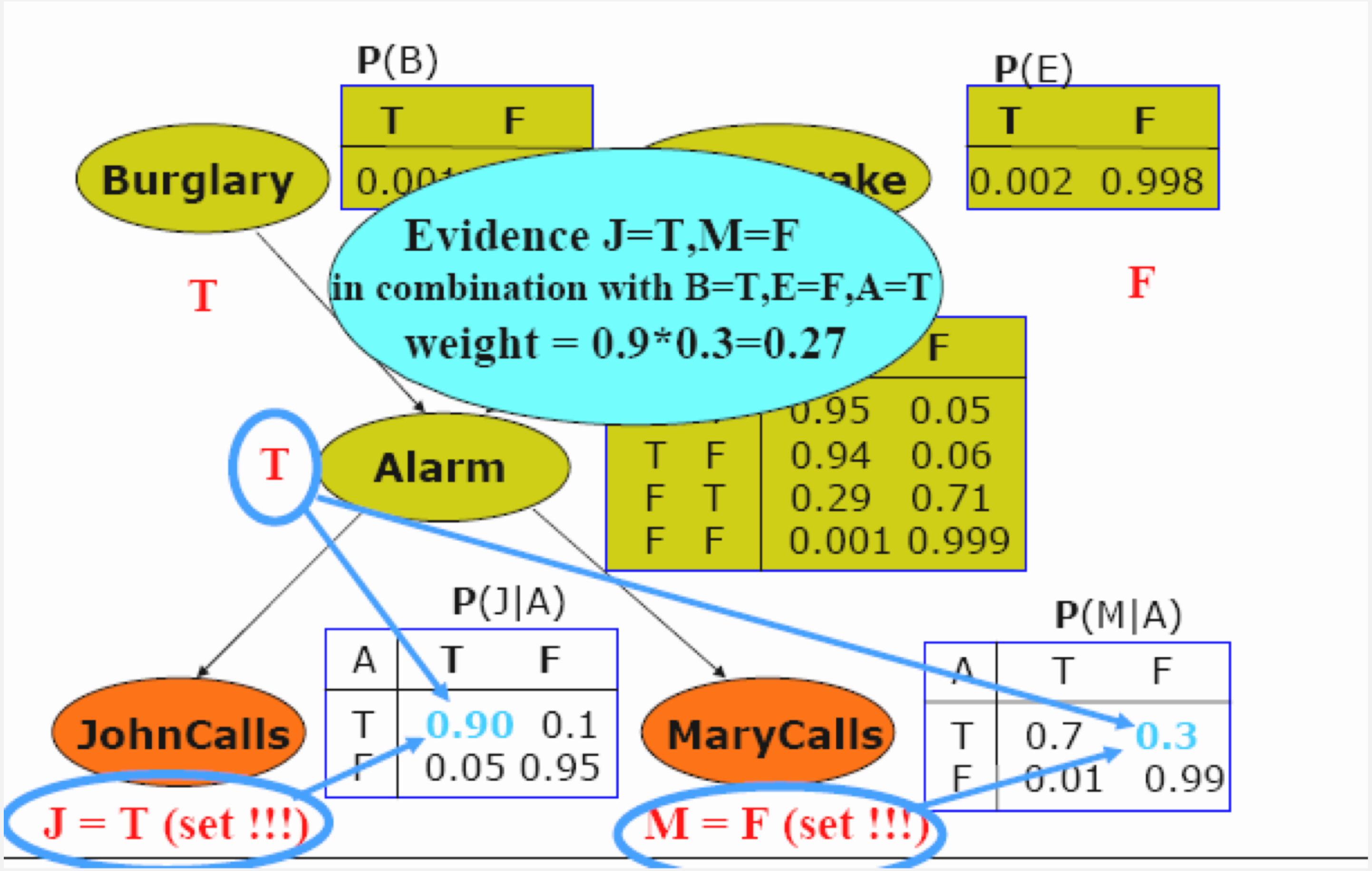
Likelihood weighting Example



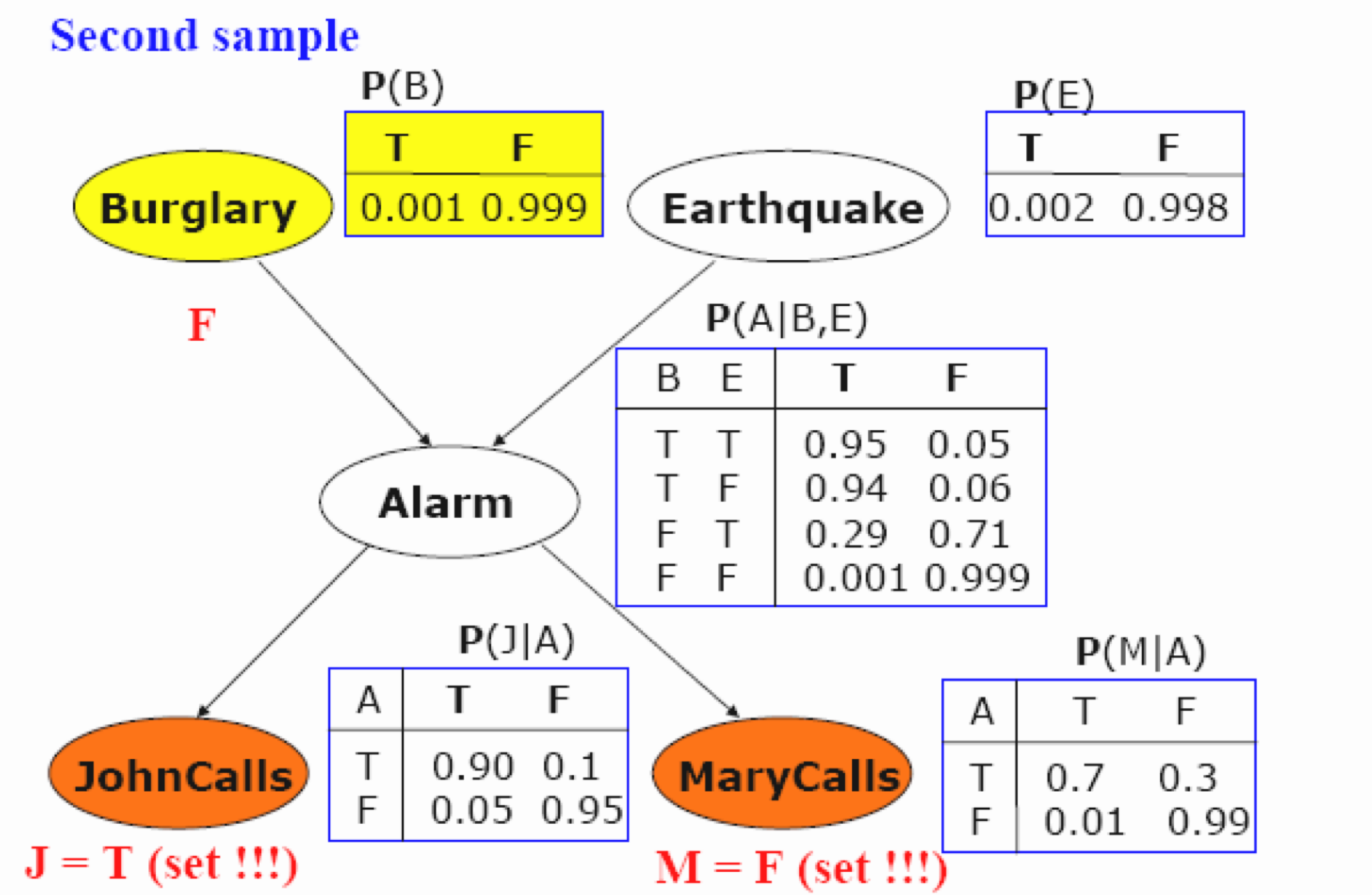
Likelihood weighting Example



Likelihood weighting Example



Likelihood weighting Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

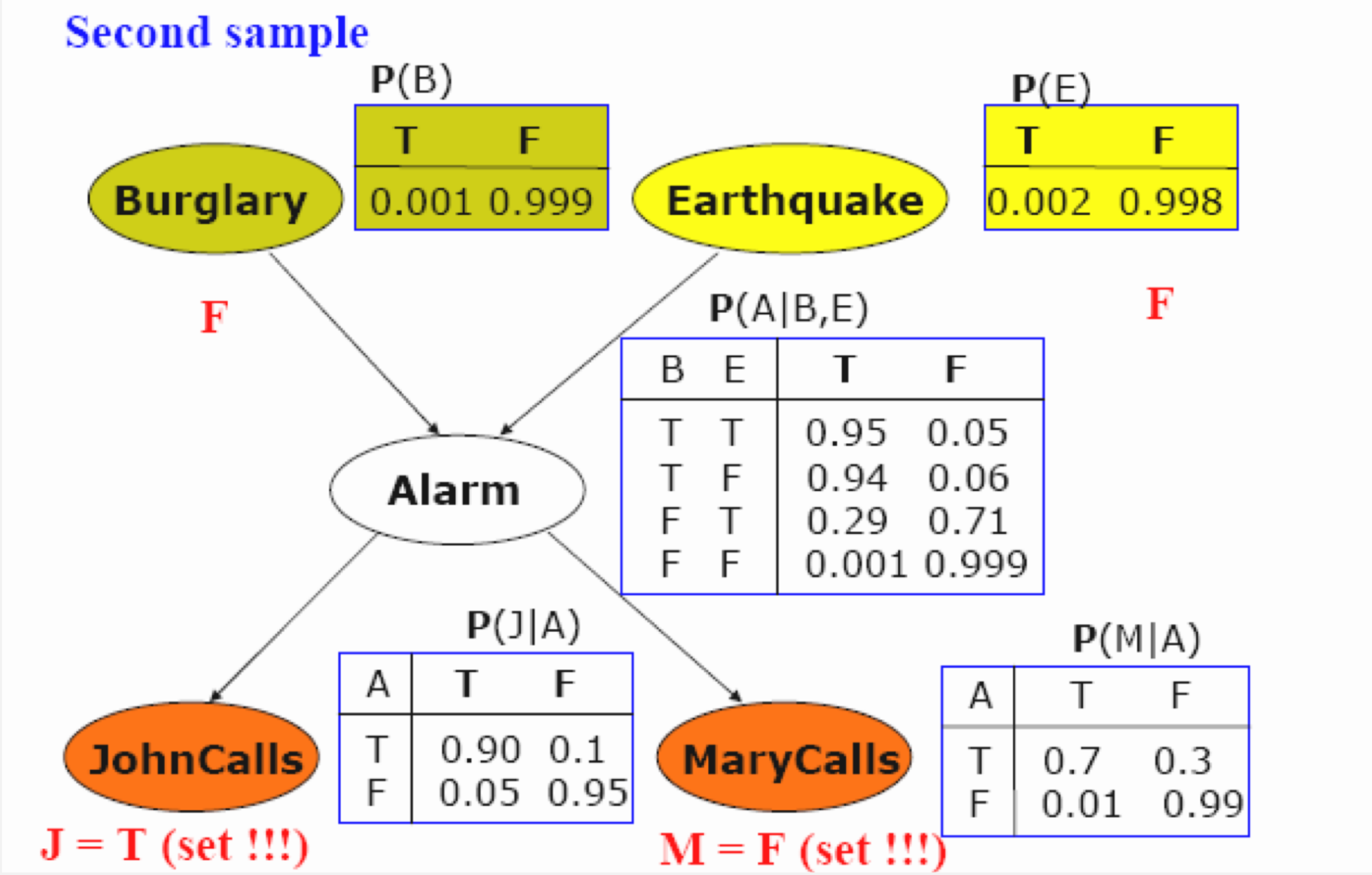
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Likelihood weighting Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

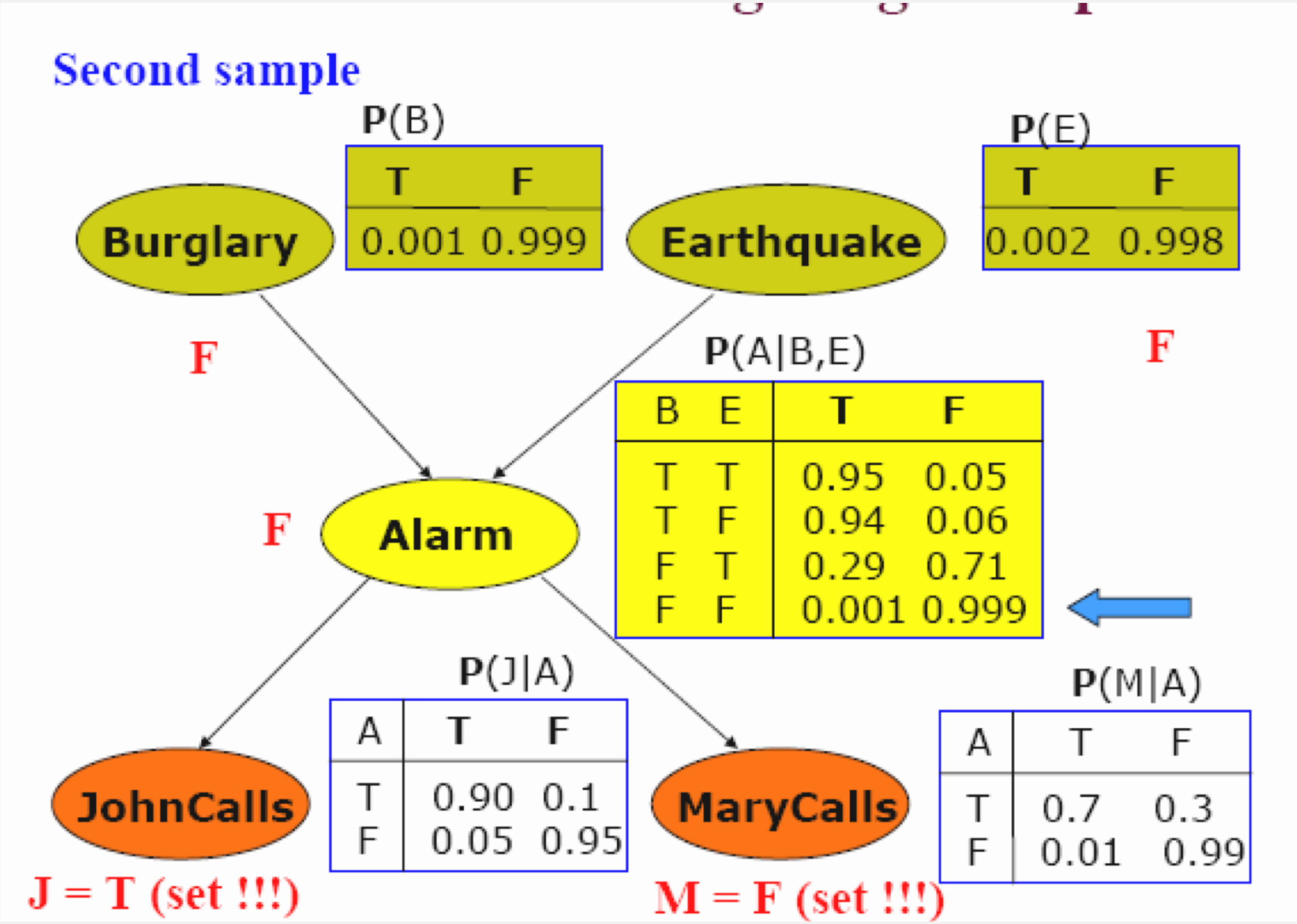
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

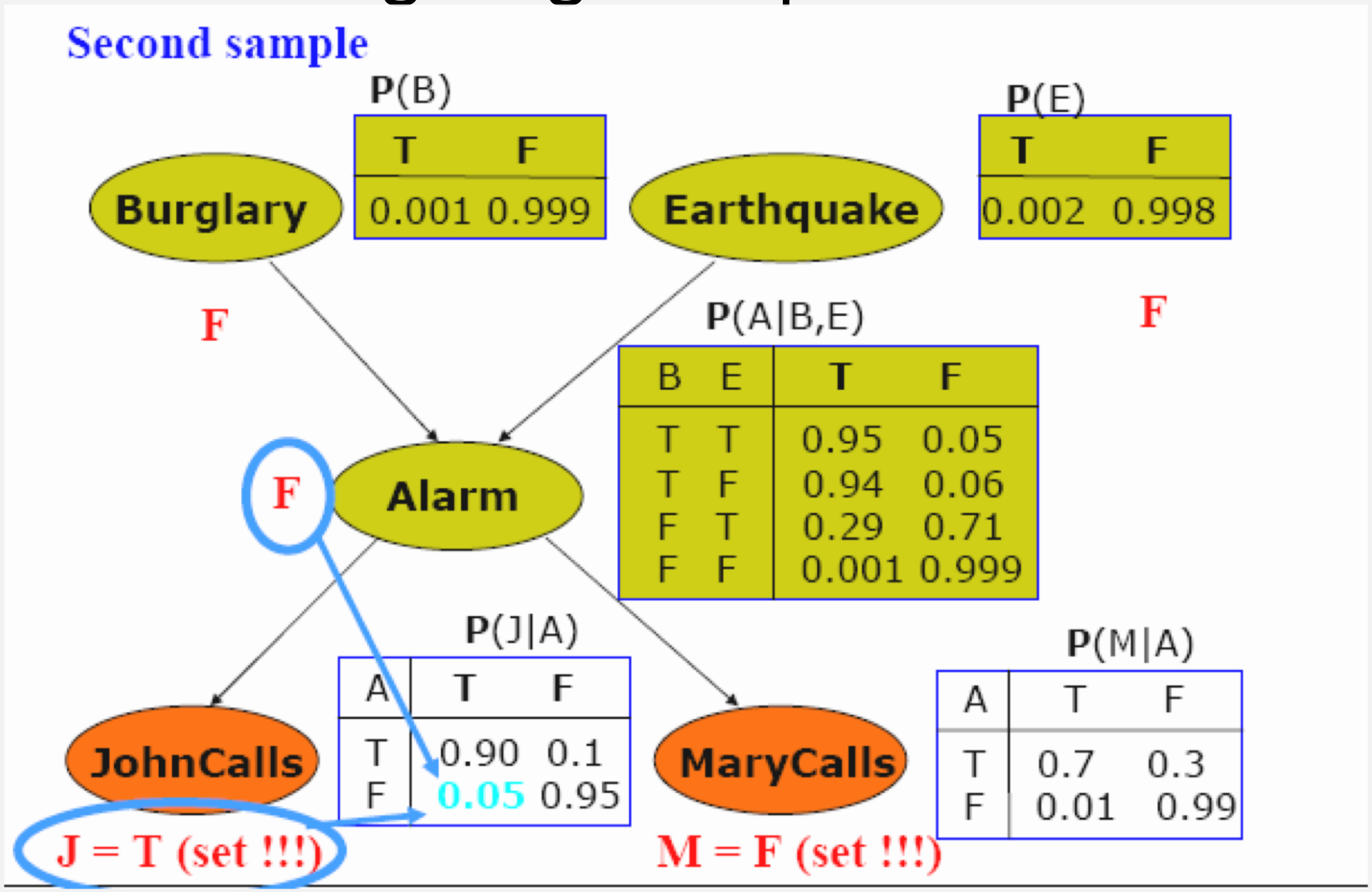
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Likelihood weighting Example



Likelihood weighting Example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

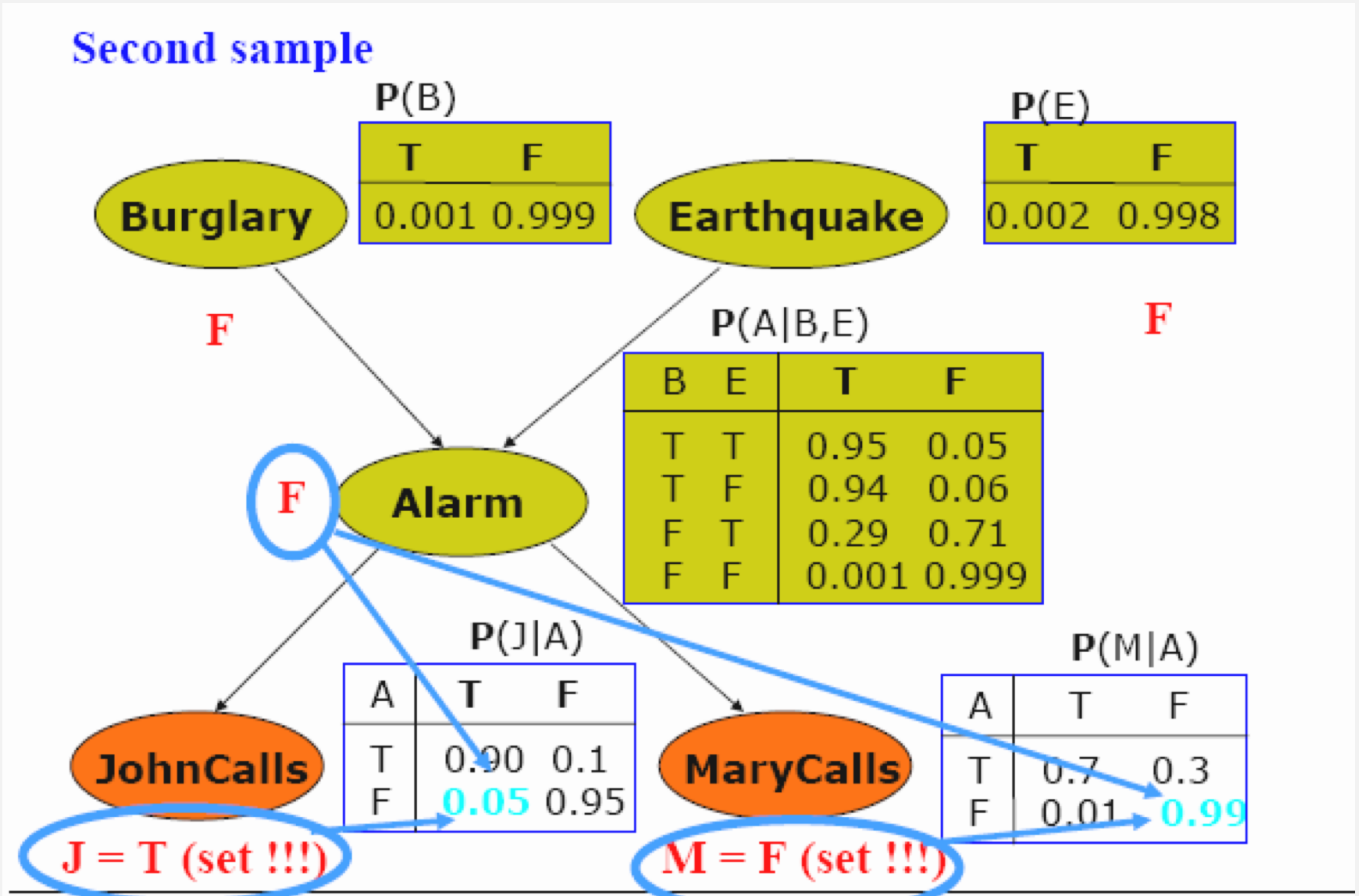
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

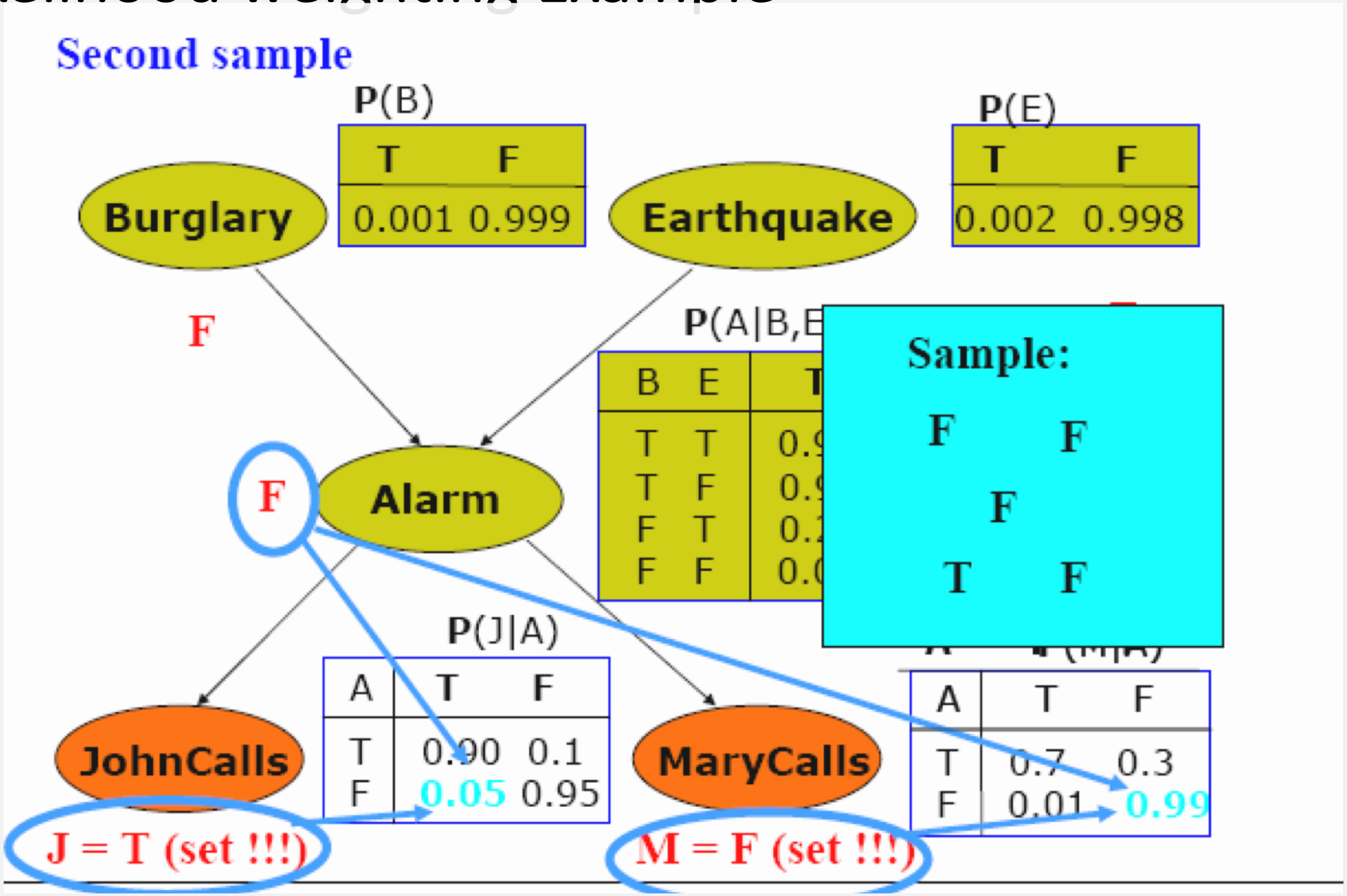
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

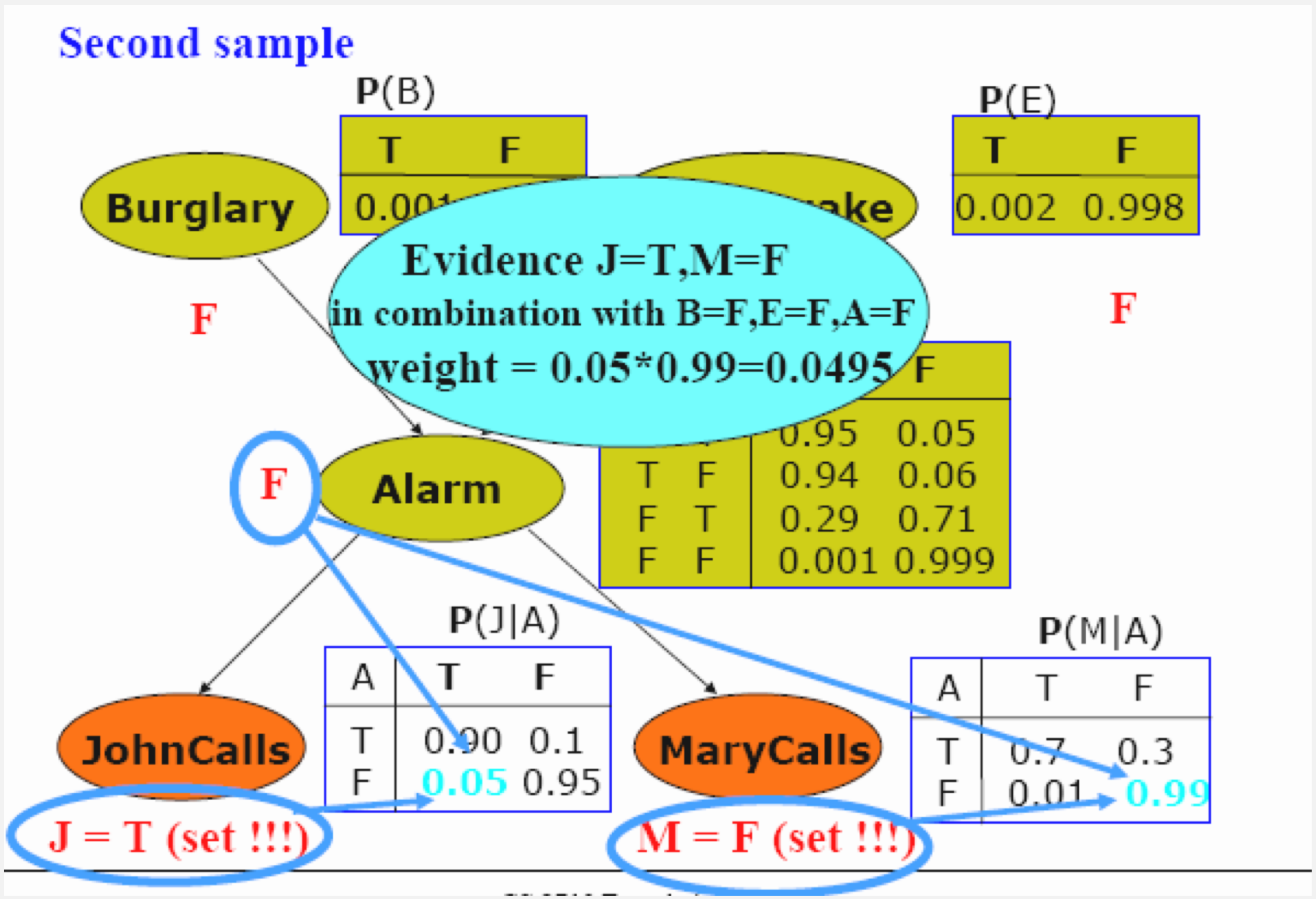
Likelihood weighting Example



Likelihood weighting Example

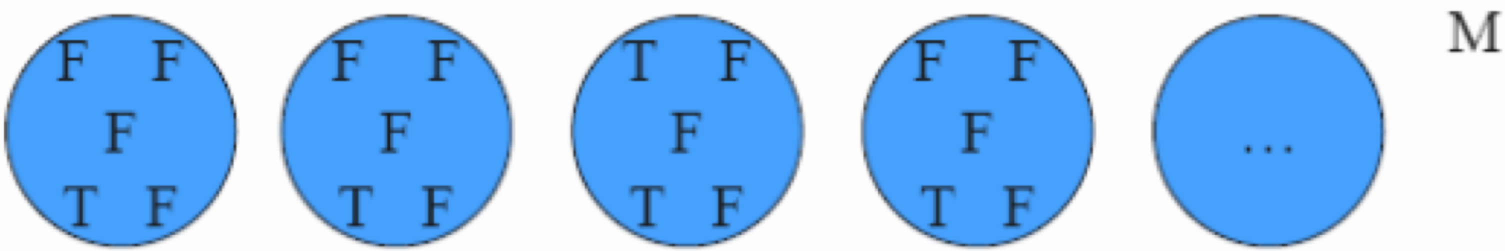


Likelihood weighting Example



Likelihood Sampling

- Assume we have generated the following M samples:

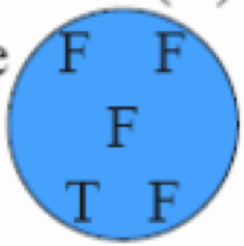


- If we calculate the estimate:

$$P(B=T \mid J=T, M=F) = \frac{\#sample\_with(B=T)}{\#total\_sample}$$

a less likely sample from P(X) may be generated more often.

- For example, sample

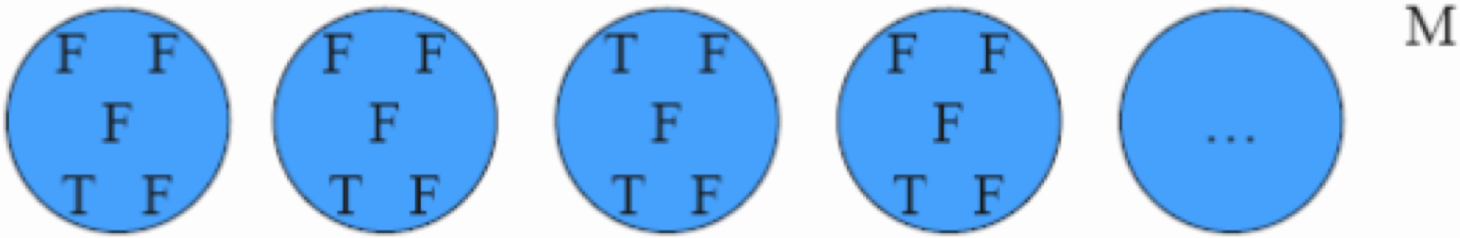


is generated more often than in P(X)

- So the samples are not consistent with P(X).

Likelihood Sampling

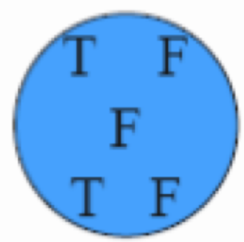
- Assume we have generated the following M samples:



How to make the samples consistent?

Weight each sample by probability with which it agrees with the conditioning evidence P(e).

← **Weight 0.0495**



← **Weight 0.27**

Likelihood Weighting

- How to compute weights for the sample?
- Assume the query $P(B=T \mid J=T, M=F)$
- Likelihood weighting:
  - **With every sample keep a weight with which it should count towards the estimate**

$$\widetilde{P}(B=T \mid J=T, M=F) = \frac{\sum_{i=1}^{M} 1\{B^{(i)}=T\} w^{(i)}}{\sum_{i=1}^{M} w^{(i)}}$$

$$\widetilde{P}(B=T \mid J=T, M=F) = \frac{\sum_{\text{samples with } B=T \text{ and } J=T, M=F} w_{B=T}}{\sum_{\text{samples with any value of } B \text{ and } J=T, M=F} w_{B=x}}$$

First order probability models

- Can we combine probability with the expressive power of first order logic (FOL) representation?
- Problem: The set of possible worlds represented by an FOL sentence can be infinite
- Relational probability models (RPM) ‘solve’ this problem by replacing standard FOL semantics by database semantics
  - Unique names assumption (e.g., each customer has a unique ID)
  - Domain closure assumption (there are no more objects beyond the ones that have been named)

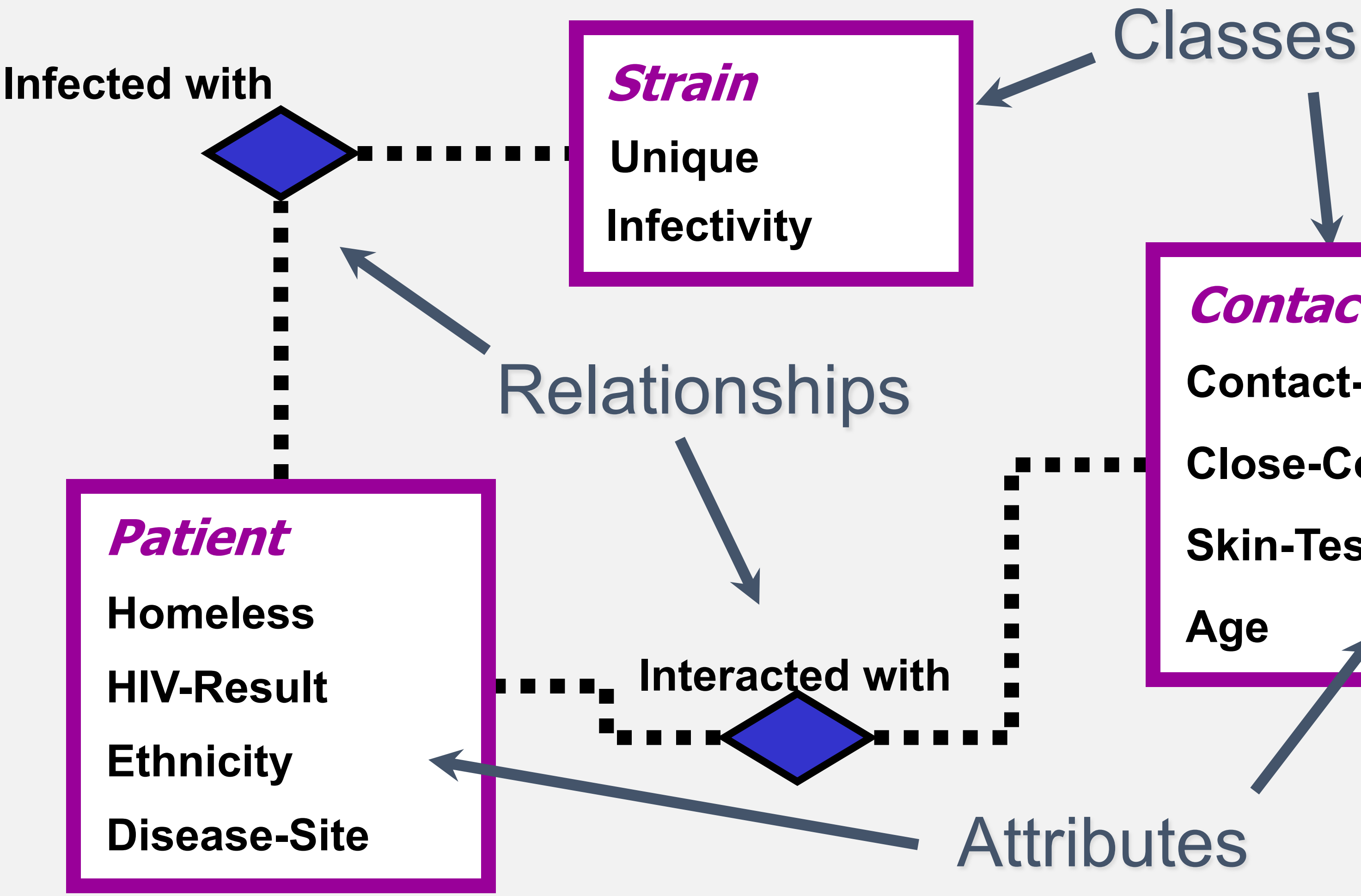
Koller, Pfeffer, Getoor et al. 1999-2007

Probabilistic Relational Models

- Combine advantages of relational logic & Bayesian networks:
  - natural domain modeling: objects, properties, relations;
  - generalization over a variety of situations;
  - compact, natural probability models.

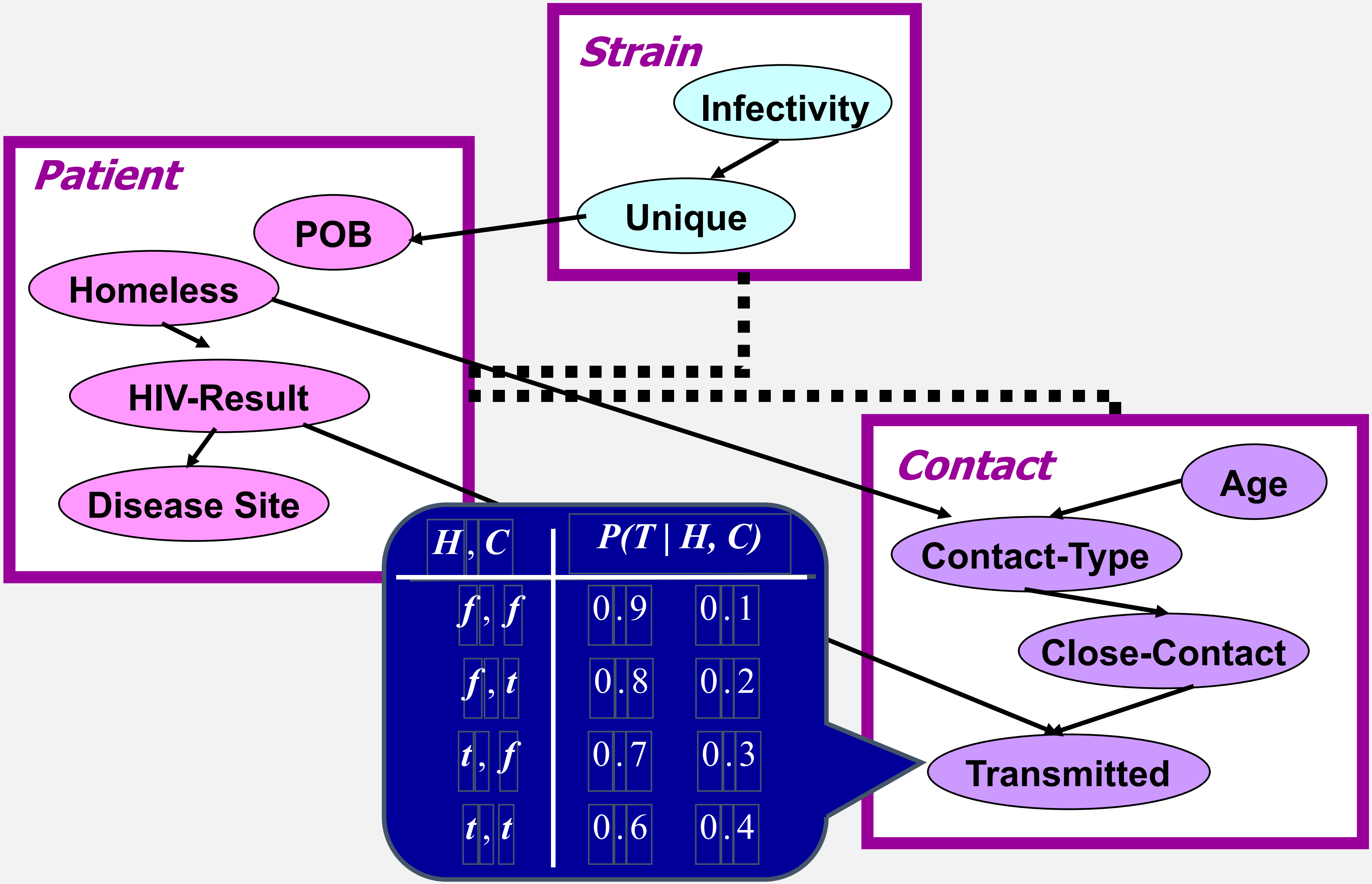
- Integrate uncertainty with relational model:
  - properties of entities can depend on properties of related entities;
  - uncertainty over relational structure of domain.

Relational Schema



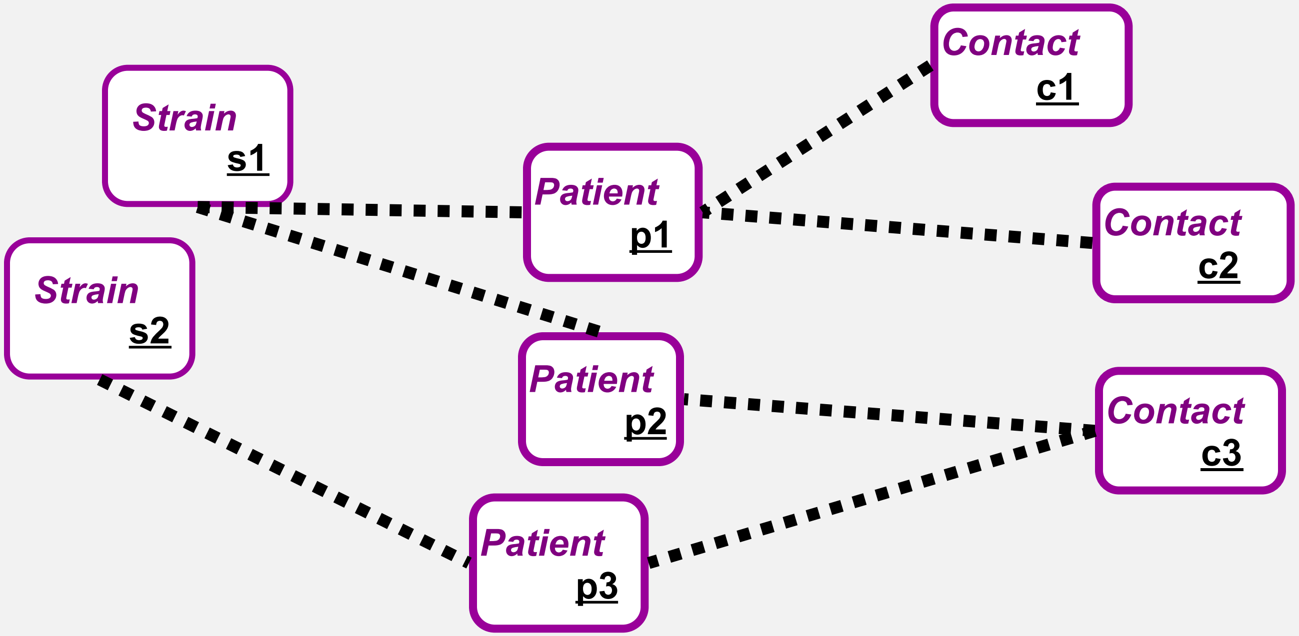
- Describes the types of objects and relations in the database

Probabilistic Relational Model



H, C	P(T | H, C)	
f, f	0.9	0.1
f, t	0.8	0.2
t, f	0.7	0.3
t, t	0.6	0.4

Relational Skeleton

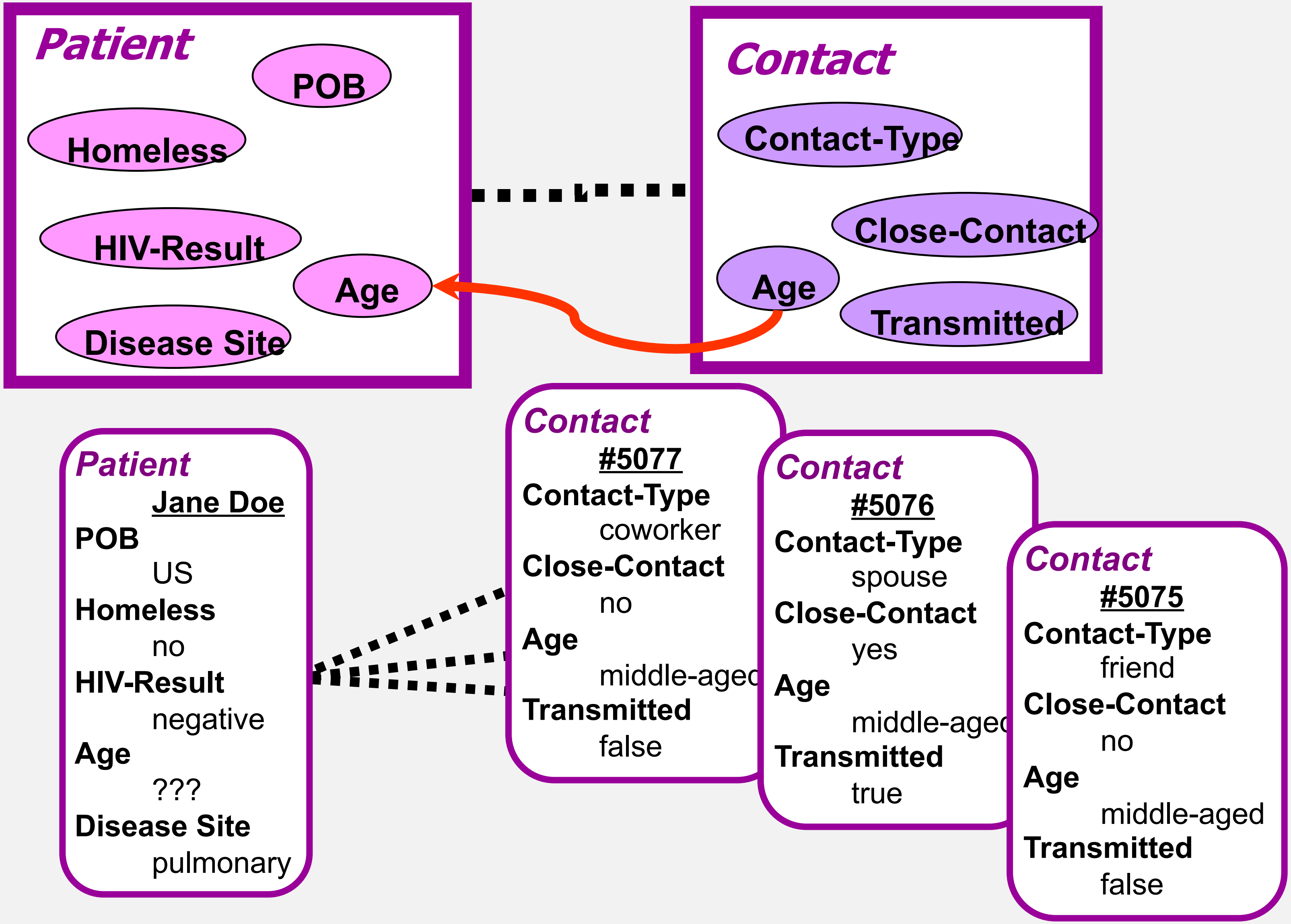


Fixed relational skeleton σ
- set of objects in each class
- relations between them

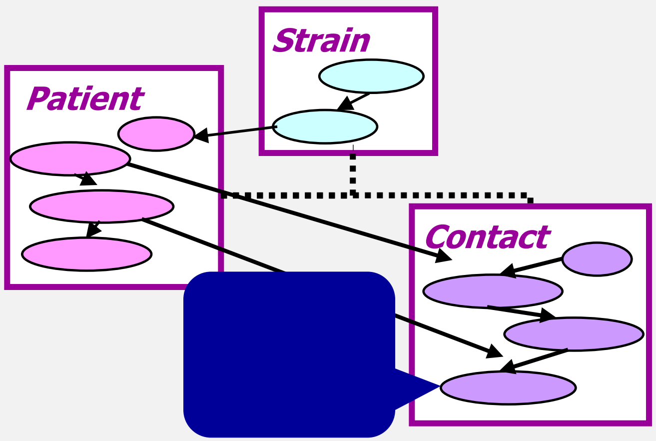
Uncertainty over assignment of values to attributes

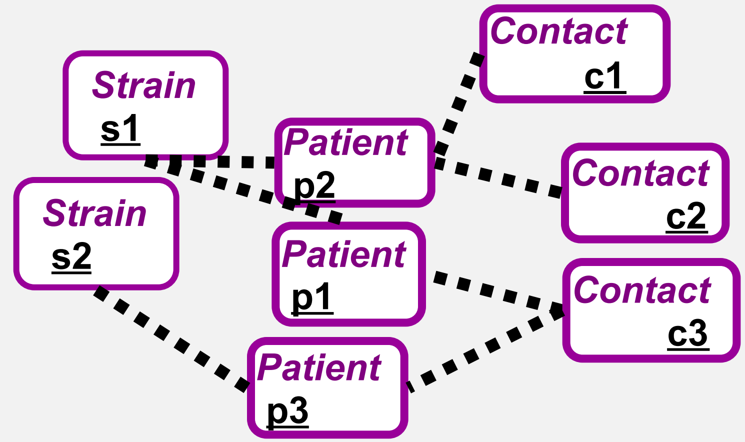
PRM defines distribution over instantiations of attributes

PRM: Aggregate Dependencies



PRM with AU Semantics





PRM + relational skeleton σ =

probability distribution over completions I:

$$P(\mathcal{I} \mid \sigma, S, \Theta) = \prod_{x \in \sigma} \prod_{x.A} P(x.A \mid parents_{S,\sigma}(x.A))$$

Objects Attributes

Open universe probability models

- Unique names assumption and domain closure assumption do not hold in the presence of uncertainty about existence and identity of objects
- Open universe probability models (OUPMs) extend Bayes networks and RPMs by adding
  - generative steps that add objects to the possible world under construction
  - where the number and type of objects added may depend on the objects that are already present

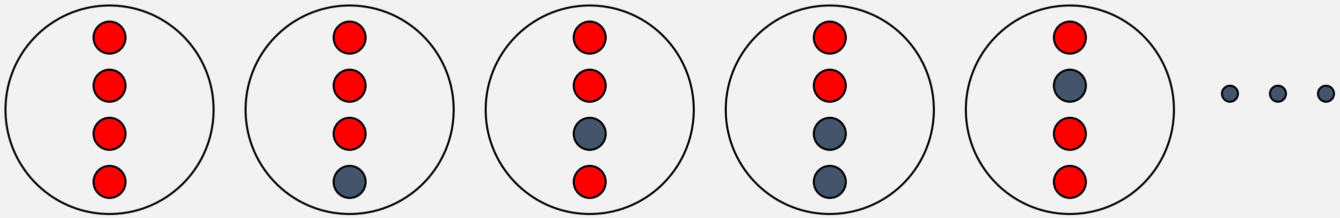
Milch et al., 2007

Herbrand vs full first-order semantics

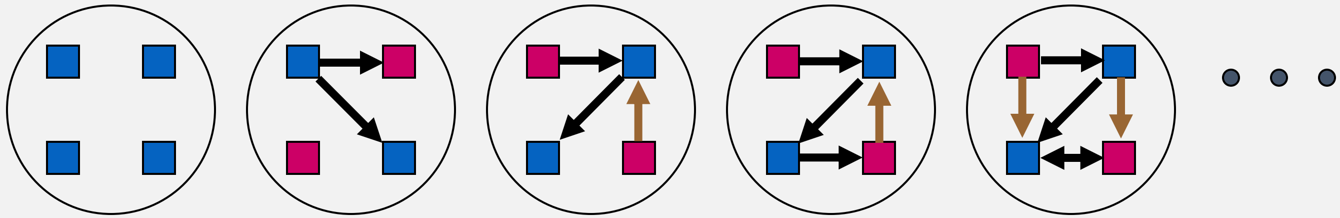
- Given: Father(Bill,William) and Father(Bill,Junior)
- How many children does Bill have?
  - Database (Herbrand) semantics: 2
  - First-order open world logical semantics:
    - Between 2 and ∞ (under the unique names assumption)
    - Between 1 and ∞ (in the absence of the unique names assumption)

Possible worlds

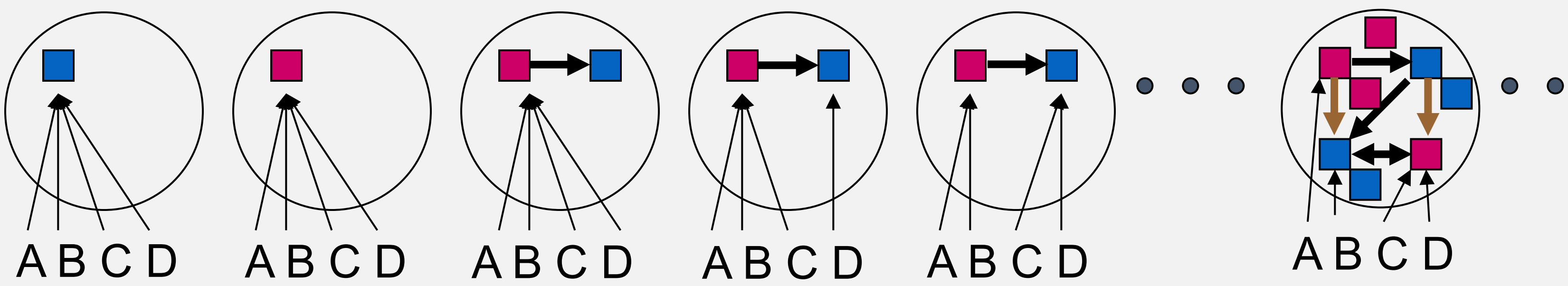
- Propositional (Boolean logic, Bayes nets)



- First-order closed-universe (DB, RPM)



- First-order open-universe: uncertainty about existence of objects and the relations



Open-universe models in BLOG

- Construct worlds using two kinds of steps, proceeding in topological order:
  - Dependency statements: Set the value of a function or relation on a tuple of (quantified) arguments, conditioned on parent values

Open-universe models in BLOG

- Construct worlds using two kinds of steps, proceeding in topological order:
  - Dependency statements: Set the value of a function or relation on a tuple of (quantified) arguments, conditioned on parent values
  - Number statements: Add some objects to the world, conditioned on what objects and relations exist so far

Technical basics

Theorem: Every well-formed* BLOG model specifies a unique proper probability distribution over open-universe possible worlds; equivalent to an infinite contingent Bayes net

Theorem: BLOG inference algorithms (rejection sampling, importance sampling, MCMC) converge to correct posteriors for any well-formed* model, for any first-order query

Example: cyber-security sibyl defense

```
#Person ~ LogNormal[6.9, 2.3]();
Honest(x) ~ Boolean[0.9]();
#Login(Owner = x) ~
   if Honest(x) then 1 else LogNormal[4.6,2.3]();
Transaction(x,y) ~
   if Owner(x) = Owner(y) then SibylPrior()
   else TransactionPrior(Honest(Owner(x)),
                         Honest(Owner(y)));
Recommends(x,y) ~
   if Transaction(x,y) then
      if Owner(x) = Owner(y) then Boolean[0.99]()
      else RecPrior(Honest(Owner(x)),
                    Honest(Owner(y)));
```

Evidence: lots of transactions and recommendations

Query: `Honest(x)`

Probabilistic Programming Languages

- Logic based
  - PRISM, Problog – logic programming + probability distributions over facts [Sato and Kameya, 2001; De Raedt, Kimmig, and Toivonen, 2007]
  - BLOG – a language based on open universe probability models [Milch et al., 2007]
- Functional programming based
  - Church, Venture – extend Scheme with probabilistic semantics for specifying recursively defined generative processes [Goodman, Mansinghka, Roy, Bonawitz and Tenenbaum, 2008]
  - IBAL – a stochastic functional programming language [Pfeffer, 2007]
- Object-oriented
  - Figaro – an expressive language with support for directed and undirected probabilistic graphical models, OUPMs, models defined over complex data structures. [Pfeffer, 2009]