Probabilistic Reasoning System (PRS)

Informatics Engineering Study Program

School of Electrical Engineering and Informatics

Institute of Technology Bandung

Contents

- Review
- Joint Probability Distribution
- Conditional Independence
- Bayesian Network
- Example of Bayesian Network
- Connection
- D - separation
- Bayesian (Belief) Network
- Inference in Bayesian Network
- Learning in Bayesian Network

Review

1. What is AI and Intelligent Agent
2. Deterministic
   - Problem Solving agent & Planning Agent
   - Knowledge Based agent
   - Learning agent (supervised, unsupervised, reinforcement)
3. Non Deterministic (Uncertainty)
   - Probabilistic & Bayes' Rule → Supervised Learning

Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: "A_{25} will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Probability

- Logic represents uncertainty by disjunction
- But, cannot tell us how likely the different conditions are
- Probability theory provides a quantitative way of encoding likelihood

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time | …) = 0.04

$P(A_{90}$ gets me there on time | …) = 0.70

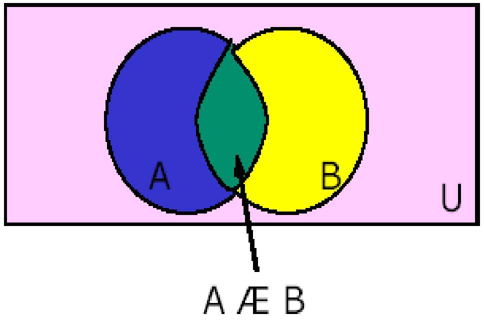
$P(A_{120}$ gets me there on time | …) = 0.95

$P(A_{1440}$ gets me there on time | …) = 0.9999

- Which action to choose?
- Depends on my preferences for missing flight vs. time spent waiting, etc.
  - Utility theory is used to represent and infer preferences
  - Decision theory = probability theory + utility theory

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- Kolmogorov's axioms about unconditional/prior probability:
  - P: events → [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P()
  - P(A U B) = P(A) + P(B) – P(A ∩ B)



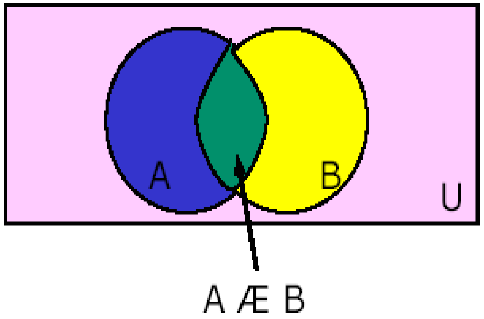
- Bayesian → Subjectivist
  - Probability is a model of your degree of belief

Examples of Human Probability Reasoning

Jane is from Berkeley. She was active in anti-war protests in the 60's. She lives in a commune.

- Which is more probable?

1. Jane is a bank teller
2. Jane is a feminist bank teller

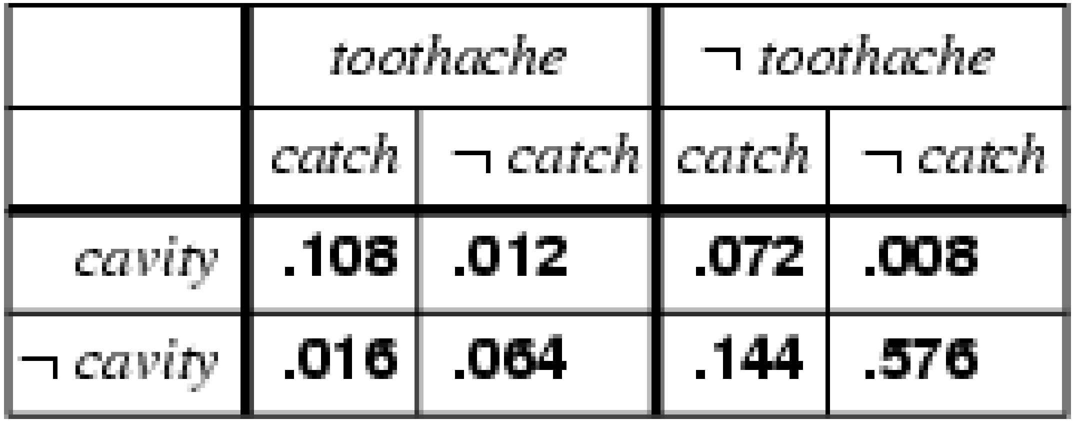


Random Variables

- Random variables
  - Function: discrete domain → [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable
    - Raining(true) = 0.2
      - □ P(Raining = true) = 0.2
    - Raining(false) = 0.8
      - □ P(Raining = false) = 0.8
- Joint distribution
  - Probability assignment to all combinations of values of random variables

Inference by enumeration

- Start with the joint probability distribution (as knowledge base):

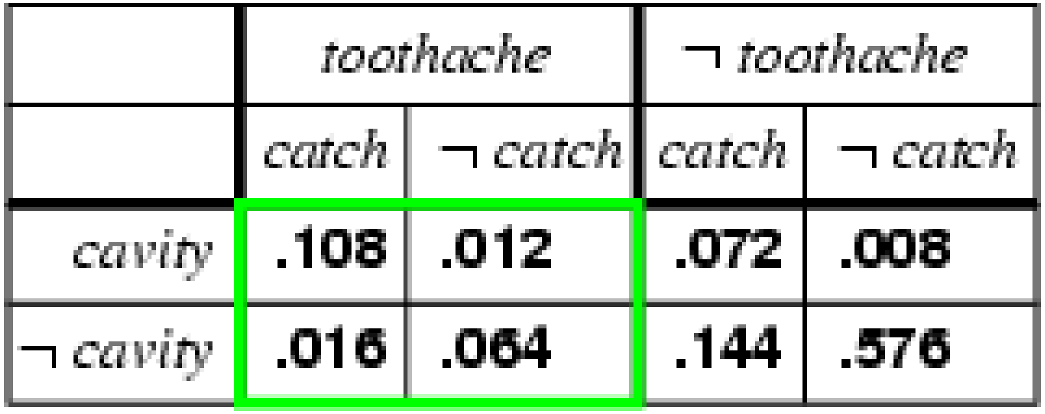


	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

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

Inference by enumeration

- Start with the joint probability distribution (as knowledge base):



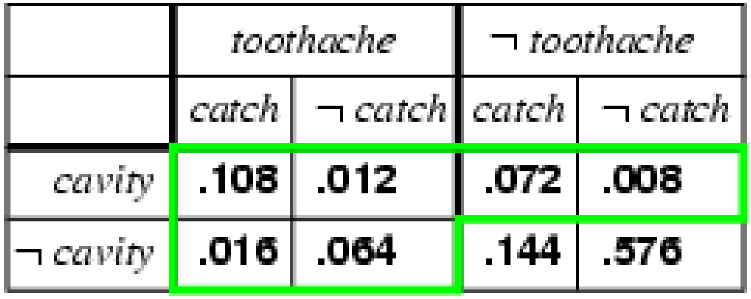
	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- 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:



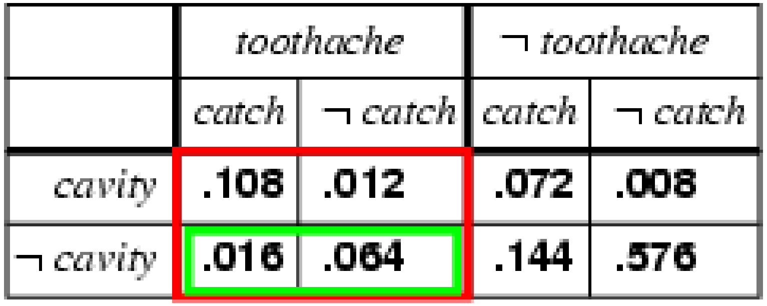
	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

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

- P(*cavity*) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2
- P(cavity U tootache) = ?

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:

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

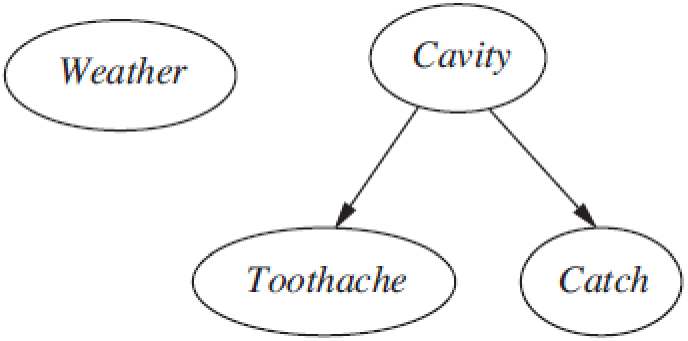
$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Bayes' Rule

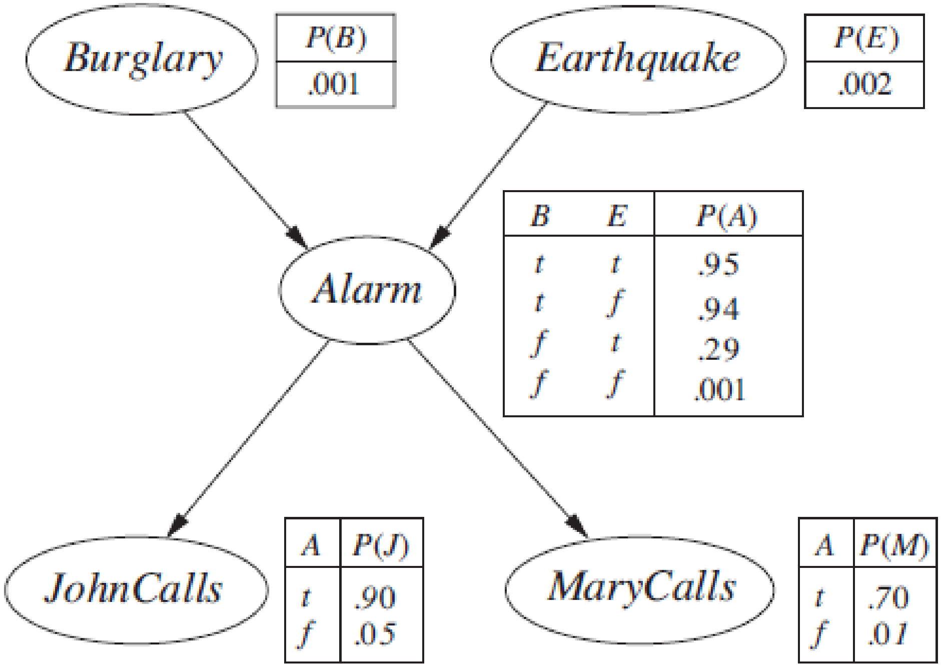
- Bayes' Rule
  - $P(A \mid B) = P(A \cap B) / P(B)$
    $= P(B \mid A) P(A) / P(B)$
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)
- Conditioning
  - $P(A) = P(A \mid B) P(B) + P(A \mid \neg B) P(\neg B)$
    $= P(A \cap B) + P(A \cap \neg B)$

Simple Bayesian Network



- Bayesian network is a directed graph in which each node is annotated with quantitative probability information.
- Weather is independent of the other three variables and Toothache and Catch are conditionally independent, given Cavity.

Typical Bayesian Network



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

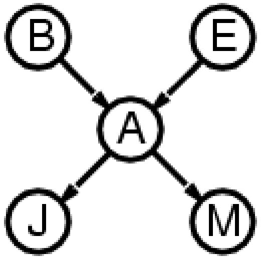
Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

John always calls when he hears the alarm, but sometimes confuses the telephone ringing with alarm and calls then, too.
Mary likes rather loud music and sometimes misses the alarm altogether.
We can estimater probability of burglary.

Chain Rule

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

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

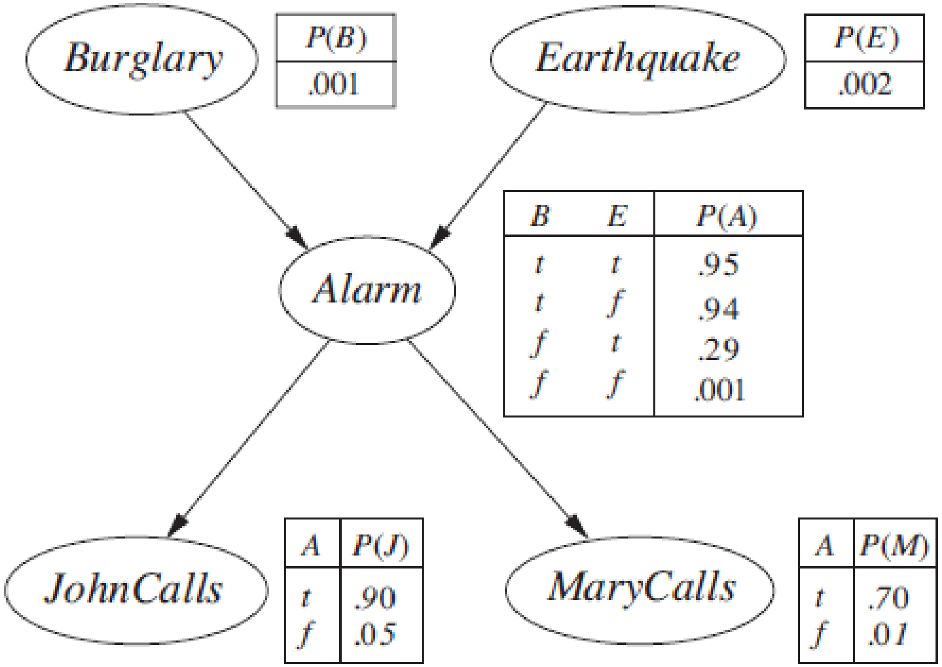


e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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

Probability that the alarm has sounded, but neither burglary nor an earthquake has occured, and both John and Mary call.

Chain Rule (lanj)



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

$\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e) = \mathbf{P}(j \mid a)\ \mathbf{P}(m \mid a)\ \mathbf{P}(a \mid \neg b, \neg e)\ \mathbf{P}(\neg b)\ \mathbf{P}(\neg e)$
$= 0.9 * 0.7 * 0.001 * 0.999 * 0.998 = 0.00062$

Independence

- A and B are independent iff
  - $P(A \cap B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Independence is essential for efficient probabilistic reasoning
- A and B are conditionally independent given C iff
  - $P(A \mid B, C) = P(A \mid C)$
  - $P(B \mid A, C) = P(B \mid C)$
  - $P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$

Example of Conditional Independence

- X is late (X)
- Traffic Jam (T)
- Y is late (Y)
- None of these propositions are independent of one other
- X and Y are conditionally independent given T

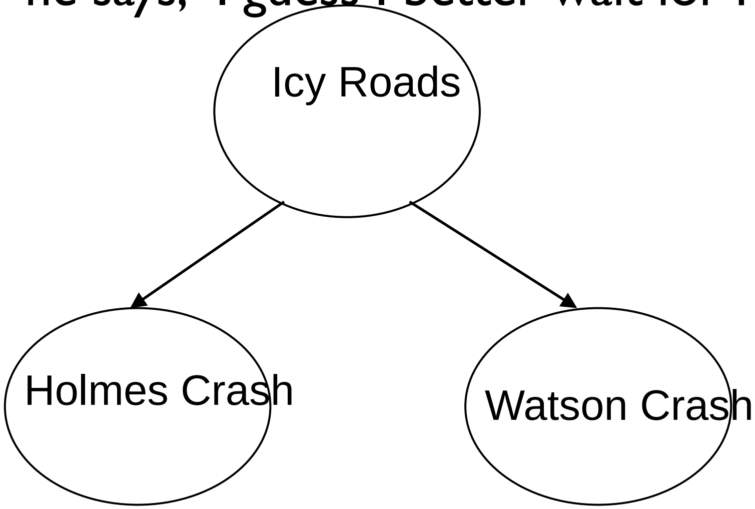
Bayesian Network

- To do probabilistic reasoning, you need to know the joint probability distribution
- But, in a domain with N binary propositional variables (2 possibilities value), one needs 2^N numbers to specify the joint probability distribution
- We want to exploit independences in the domain
- Two components: structure and numerical parameters

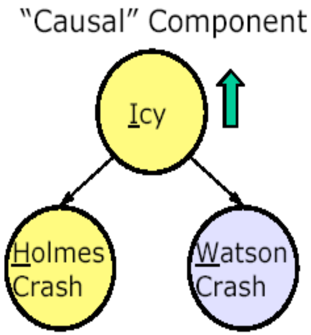
Example of Bayesian Network

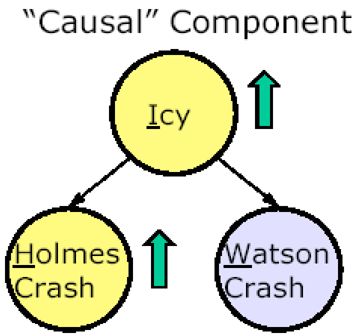
- Icy Roads

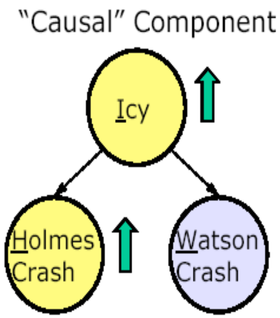
Inspector Smith is waiting for Holmes and Watson, who are driving (separately) to meet him. It is winter. His secretary tells him that Watson has had an accident. He says, "It must be that the roads are icy. I bet that Holmes will have an accident too. I should go to lunch." But, his secretary says, "No, the roads are not icy, look at the window." So, he says, "I guess I better wait for Holmes."



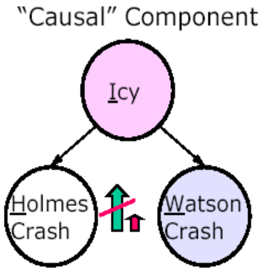
Icy Roads (con’t)



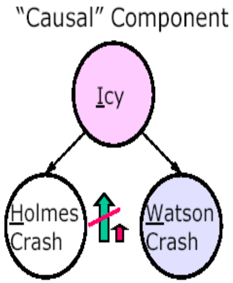




H and W are dependent,



H and W are dependent,



H and W are dependent, but conditionally independent given I

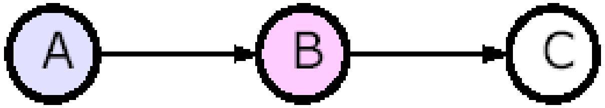
Lecture

Connections

A = battery dead

B = car won’t start

C = car won’t move



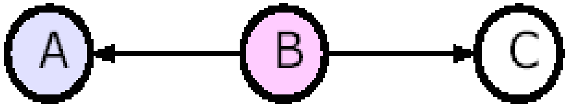
- Forward Serial Connection
  - Knowing about A will tell us something about C
  - But if we know B then knowing about A will not tell us anything about C
- Backward Serial Connection
  - Knowing about C will tell us something about A
  - But if we know B then knowing about C will not tell us anything about A

Connections (con't)

A = Watson Crash

B = Icy

C = Holmes Crash



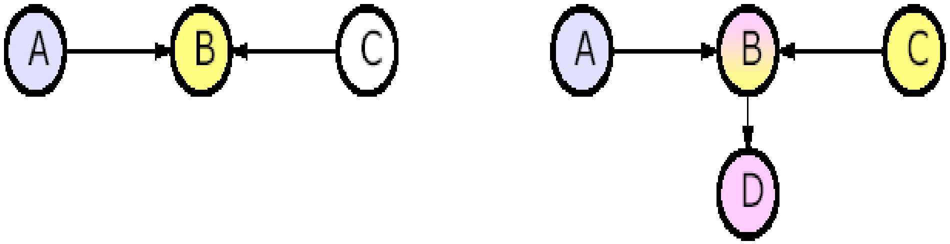
- Diverging Connection
  - Knowing about A will tell us something about C
  - Knowing about C will tell us something about A
  - But if we know B then knowing about A will not tell us anything new about C, and vice versa

Connections (con’t)

A = Bacterial Infection

B = Sore Throat

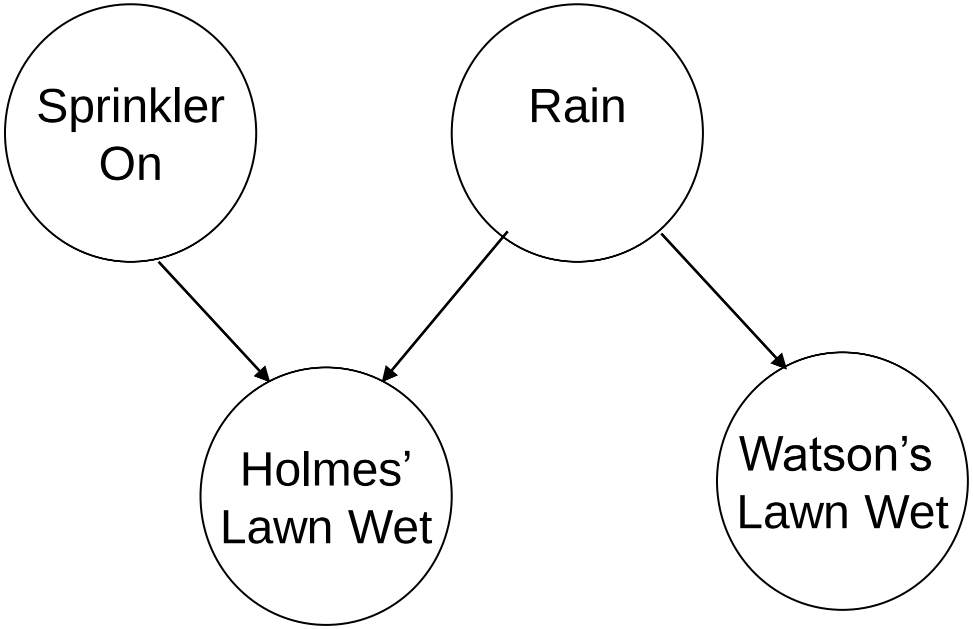
C = Viral Infection



- Converging Connection
  - Without knowing B finding A does not tell us something about C
  - If we see evidence for B, then A and C becomes dependent (potential for “explaining away”). If we find bacteria in patient with a sore throat, then viral infection is less likely.

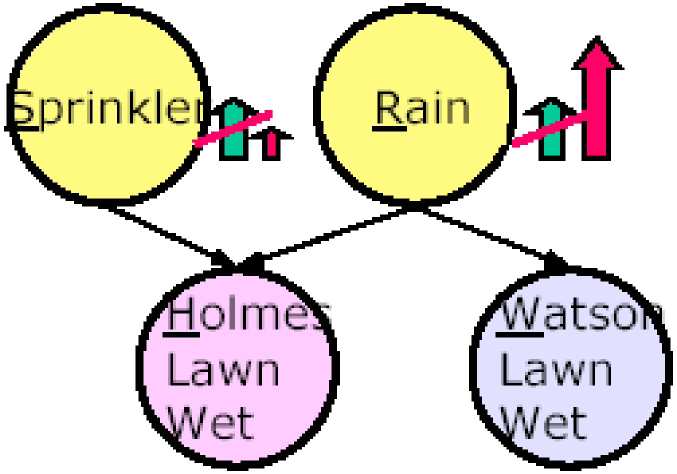
Connections (con't)

Holmes and Watson have moved to LA. He wakes up to find his lawn wet. He wonders if it has rained or if he left his sprinkler on. He looks at his neighbor Watson's lawn and he sees it is wet too. So, he concludes it must have rained.



Connections (con't)

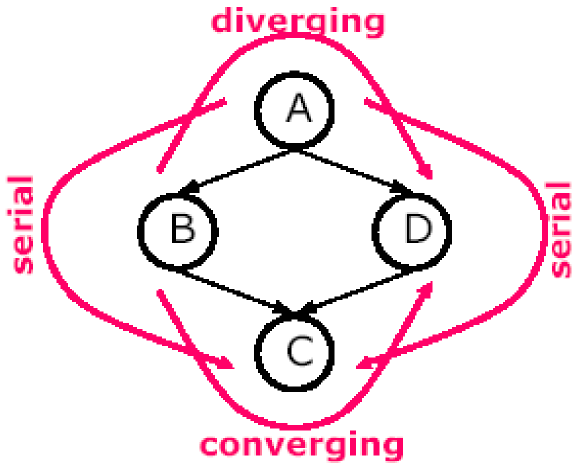
Holmes and Watson have moved to LA. He wakes up to find his lawn wet. He wonders if it has rained or if he left his sprinkler on. He looks at his neighbor Watson's lawn and he sees it is wet too. So, he concludes it must have rained.



Given W, P(R) goes up and P(S) goes down – "explaining away"

D Separation

- Two variables A and B are d-separated iff for every path between them, there is an intermediate variable V such that either
  - The connection is serial or diverging and V is known
  - The connection is converging and neither V nor any descendant is instantiated
  - Two variables are d-connected iff they are not d-separated



- A-B-C: serial, blocked when B is known, connected otherwise
- A-D-C: serial, blocked when D is known, connected otherwise
- B-A-D: diverging, blocked when A is known, connected otherwise
- B-C-D: converging, blocked when C has no evidence, connected otherwise

Bayesian (Belief) Network

- Set of variables, each has a finite set of values
- Set of directed arcs between them forming acyclic graph
- Every node A, with parents B1, …, Bn, has P(A |B1,…,Bn) specified

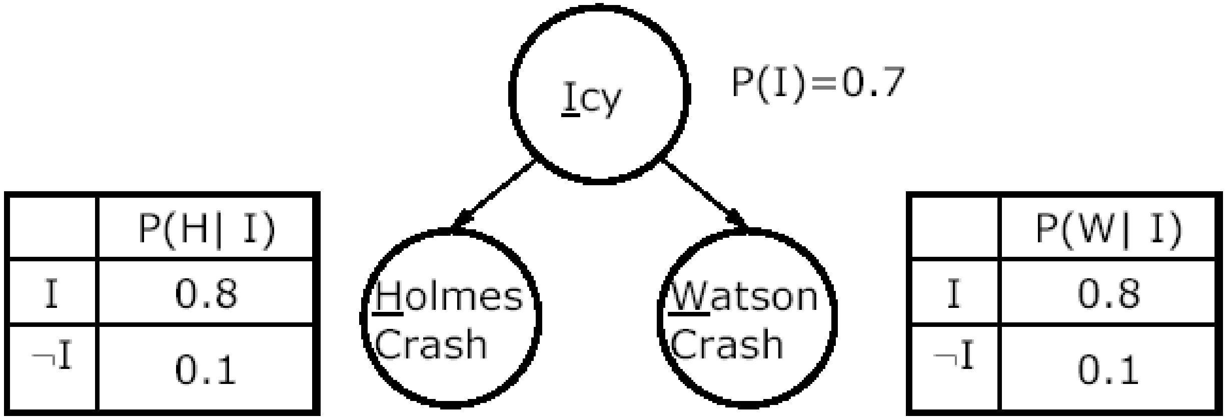
Theorem: If A and B are d-separated given evidence e, then P(A | e) = P(A | B, e)

Inference in Bayesian Network

- Exact inference
- Approximate inference
- Given a Bayesian Network, what questions might we want to ask?
  - Conditional probability query: P(x | e)
  - Maximum a posteriori probability:
    What value of x maximizes P(x|e) ?

- General question: What's the whole probability distribution over variable X given evidence e, P(X | e)?

Icy Roads with Numbers



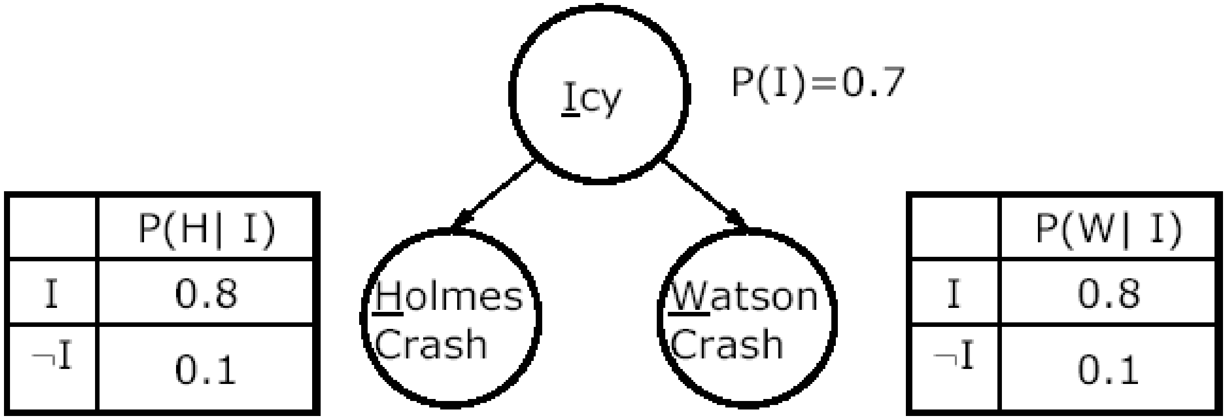
	P(H| I)
I	0.8
¬I	0.1

	P(W| I)
I	0.8
¬I	0.1

Probability that Watson Crashes:

$$
\begin{aligned}
P(W) &= P(W|\ I)\ P(I) + P(W|\ \neg I)\ P(\neg I) \\
&= 0.8 \cdot 0.7 + 0.1 \cdot 0.3 \\
&= 0.56 + 0.03 \\
&= 0.59
\end{aligned}
$$

Icy Roads with Numbers (con’t)



	P(H| I)
I	0.8
¬I	0.1

	P(W| I)
I	0.8
¬I	0.1

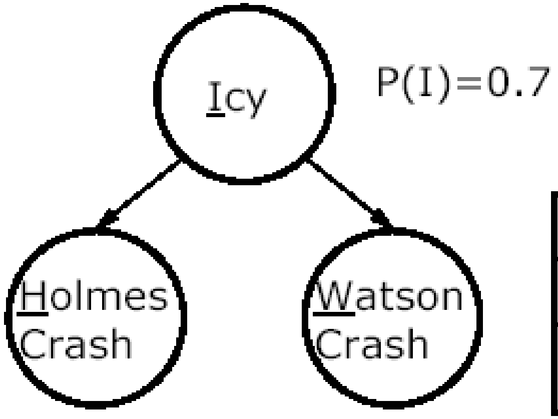
Probability of Icy given Watson (Bayes’ Rule):

$$
\begin{aligned}
P(I \mid W) &= P(W \mid I)\, P(I) / P(W) \\
&= 0.8 \cdot 0.7 / 0.59 \\
&= 0.95
\end{aligned}
$$

We started with P(I) = 0.7; knowing that Watson crashed raised the probability to 0.95

Icy Roads with Numbers (con't)

	P(H| I)
I	0.8
¬I	0.1



	P(W| I)
I	0.8
¬I	0.1

Probability of Holmes given Watson :

P(H|W) = P(H|W,I)P(I|W) + P(H|W,¬I) P(¬I| W)

= P(H|I)P(I|W) + P(H|¬I) P(¬I| W)

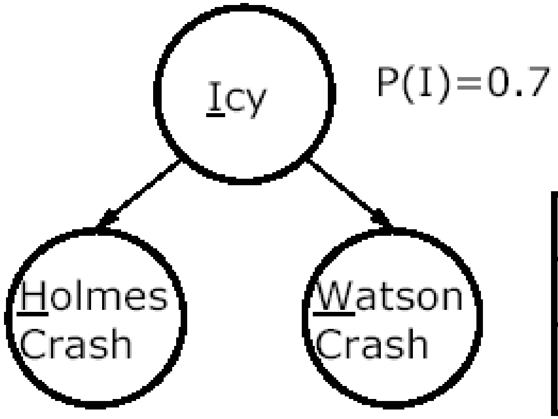
= 0.8 · 0.95 + 0.1 · 0.05

= 0.765

We started with P(H) = 0.59; knowing that Watson crashed raised the probability to 0.765

Icy Roads with Numbers (con’t)

	P(H| I)
I	0.8
¬I	0.1



	P(W| I)
I	0.8
¬I	0.1

Probability of Holmes given Icy and Watson :
P(H|W, ¬I) = P(H|¬I) = 0.1

H and W are d-separated given I, so H and W are conditionally independent given I

Where do Bayesian Networks Come From?

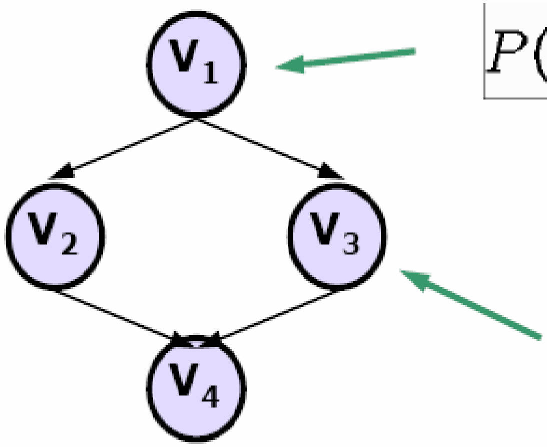
- Human Expert
  - Encoding rules obtained from expert
  - Very difficult in getting reliable probability estimates
- Learning From Data
  - Try to estimate the joint probability distribution
  - Looking for models that encode conditional independencies in data
  - Four cases →
    - Structure known or unknown
    - All variables are observable or some observable
- Combination of Both

Case 1: Structure is given

- Given nodes and arcs of a Bayesian network with m nodes
- Given a data set D = $\{<v_1^1,\ldots,v_m^1>,\ldots, <v_1^k,\ldots,v_m^k>\}$
- Elements of D are assumed to be independent given M
- Find the model M that maximizes Pr(D|M)
- Known as the maximum likelihood model
- Humans are good at providing structure, data is good at providing numbers

Case 1: Estimates the Conditional Probability

- Use counts and definition of conditional probability



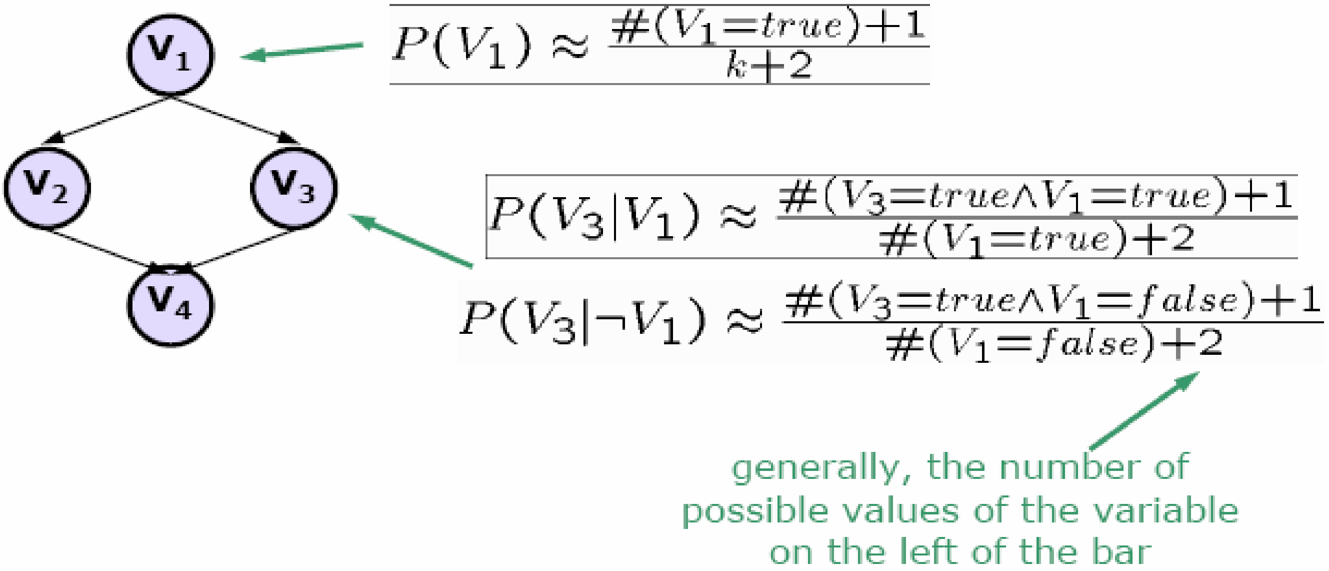
$$P(V_1) \approx \frac{\#(V_1 = true)}{k}$$

$$P(V_3|V_1) \approx \frac{\#(V_3 = true \wedge V_1 = true)}{\#(V_1 = true)}$$

$$P(V_3|\neg V_1) \approx \frac{\#(V_3 = true \wedge V_1 = false)}{\#(V_1 = false)}$$

Case 1: Estimates the Conditional Probability

- Use counts and definition of conditional probability
- Initializing all counters to 1 avoids 0 probabilities and converges on the maximum likelihood estimate



$$P(V_1) \approx \frac{\#(V_1{=}true)+1}{k+2}$$

$$P(V_3|V_1) \approx \frac{\#(V_3{=}true \wedge V_1{=}true)+1}{\#(V_1{=}true)+2}$$

$$P(V_3|\neg V_1) \approx \frac{\#(V_3{=}true \wedge V_1{=}false)+1}{\#(V_1{=}false)+2}$$

generally, the number of possible values of the variable on the left of the bar

Constructing Bayesian/ Belief Network

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

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

This choice of parents guarantees:

$$\mathbf{P}(X_1, \dots, X_n) = \pi_{i=1} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$$

(chain rule)

$$= \pi_{i=1} \mathbf{P}(X_i \mid Parents(X_i))$$

(by construction)

Example

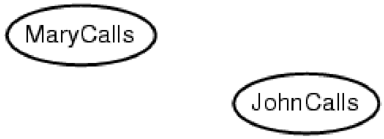
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

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

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

(Incorrect) Example

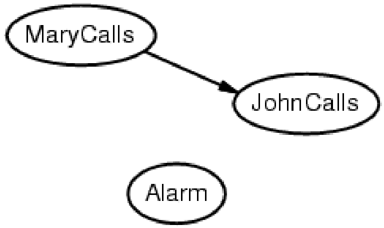
- Suppose we choose the ordering *M, J, A, B, E*



P*(J | M) =* ***P****(J)?*

(Incorrect) Example

- Suppose we choose the ordering *M, J, A, B, E*



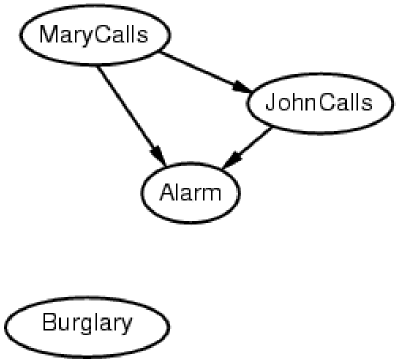
$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$?

(Incorrect) Example

- Suppose we choose the ordering *M, J, A, B, E*



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

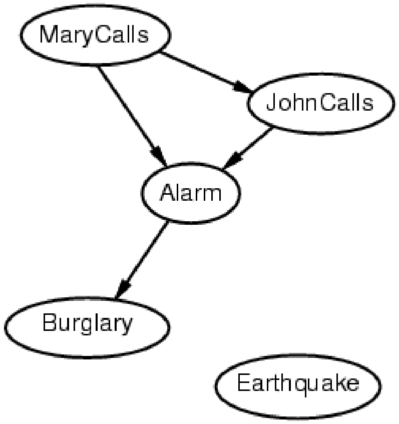
$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$?

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$?

(Incorrect) Example

- Suppose we choose the ordering M, J, A, B, E



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**

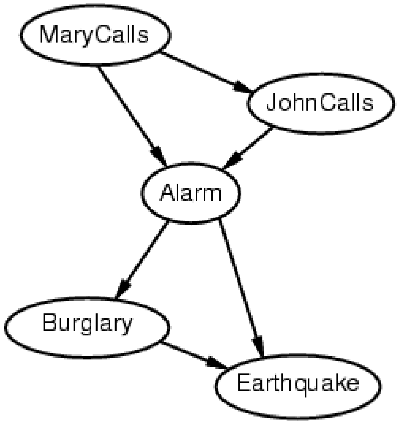
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

$\mathbf{P}(E \mid B, A ,J, M) = \mathbf{P}(E \mid A)$?

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A, B)$?

(Incorrect) Example

- Suppose we choose the ordering M, J, A, B, E



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

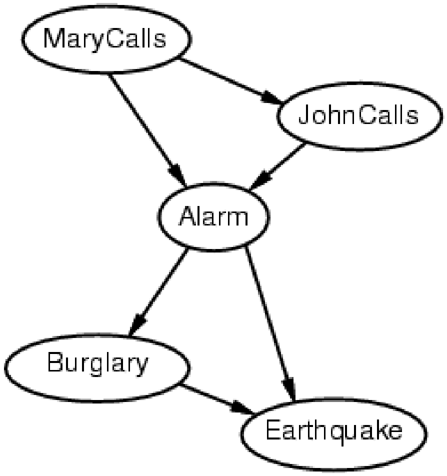
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

$\mathbf{P}(E \mid B, A ,J, M) = \mathbf{P}(E \mid A)$? **No**

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A, B)$? **Yes**

(Incorrect) Example contd.



- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed
- For 'correct' network, only requires 1 + 1 + 4 + 2 + 2 = 10 numbers

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct

THANK YOU