CSC 665: Artificial Intelligence

Bayes Nets: Introduction

Instructor: Pooyan Fazli
San Francisco State University

Quick Warm-up

Random Variables A, B, C are all binary (can be true or false)

1. How many values are in each of these tables?
2. What do the values sum to (if it is possible to determine)?

	Size	Sum
P(A)	2	1
P(a)	1	?
P(A,B)	4	1
P(A,b)	2	?
P(A|b)	2	1
P(A|B)	4	|B|
P(b|A)	2	?

	Size	Sum
P(A,B,C)	8	1
P(A|B,C)	8	|B||C|
P(A|B,c)	4	|B|
P(A|b,c)	2	1
P(a,B|c)	2	?
P(A,B|c)	4	1
P(A,B|C)	8	|C|

Independence

Independence

- Two variables X and Y are *independent* if:

$$\forall x, y : P(x|y) = P(x)$$

  - Knowing the value of one does not tell you anything about the other
  - Example: Weather is independent of the result of a coin toss

  - Another form:

$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product of two simpler distributions

  - We write:

$$X \perp\!\!\!\perp Y$$

Conditional Independence

- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

  if and only if:

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

- learning that Y=y does not change your belief in X when we already know Z=z
- and this is true for all values y that Y could take and all values z that Z could take

- or, equivalently, if and only if

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

Conditional Independence and the Chain Rule

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$

- Trivial decomposition:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

- With assumption of conditional independence:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

- Conditional Independence allows us to write them compactly

Probability Recap

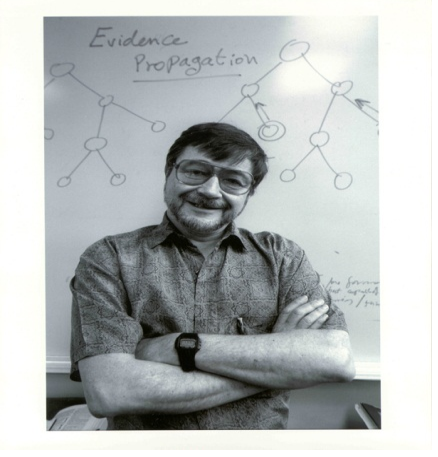
- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $$\forall x, y : P(x|y) = P(x)$$ $$\forall x, y : P(x,y) = P(x)P(y)$$
- X and Y are conditionally independent given Z if and only if: $$\forall x, y, z : P(x|z,y) = P(x|z)$$ $$\forall x, y, z : P(x,y|z) = P(x|z)P(y|z)$$

Bayes Nets

Bayes Nets: Intro

- We want a representation and reasoning system that is based on conditional probability
  - Compact yet expressive representation
  - Efficient reasoning procedures
- Bayes[ian] (Belief) Net[work]s are such a representation
  - Named after Thomas Bayes (ca. 1702 –1761)
  - Term coined in 1985 by Judea Pearl (1936 –)
  - Their invention changed the *primary* focus of AI from logic to probability!
  - In statistics, the term **graphical model** refers to a somewhat broader class that includes Bayes Nets.

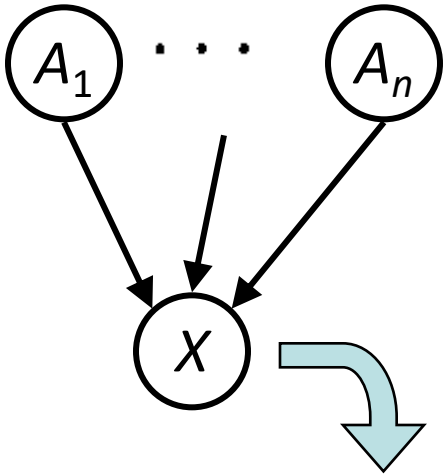
Thomas Bayes 

Judea Pearl 

Bayes Net Syntax and Semantics

Bayes Net Syntax

- A set of nodes, one per variable X
- A directed, acyclic graph (An arc from A to B indicates A ‘causes’ B or A ‘influences’ B)
- A conditional distribution for each node
  - CPT: conditional probability table



$$P(X|A_1 \ldots A_n)$$

A Bayes Net = Topology (Graph) + Conditional Probabilities

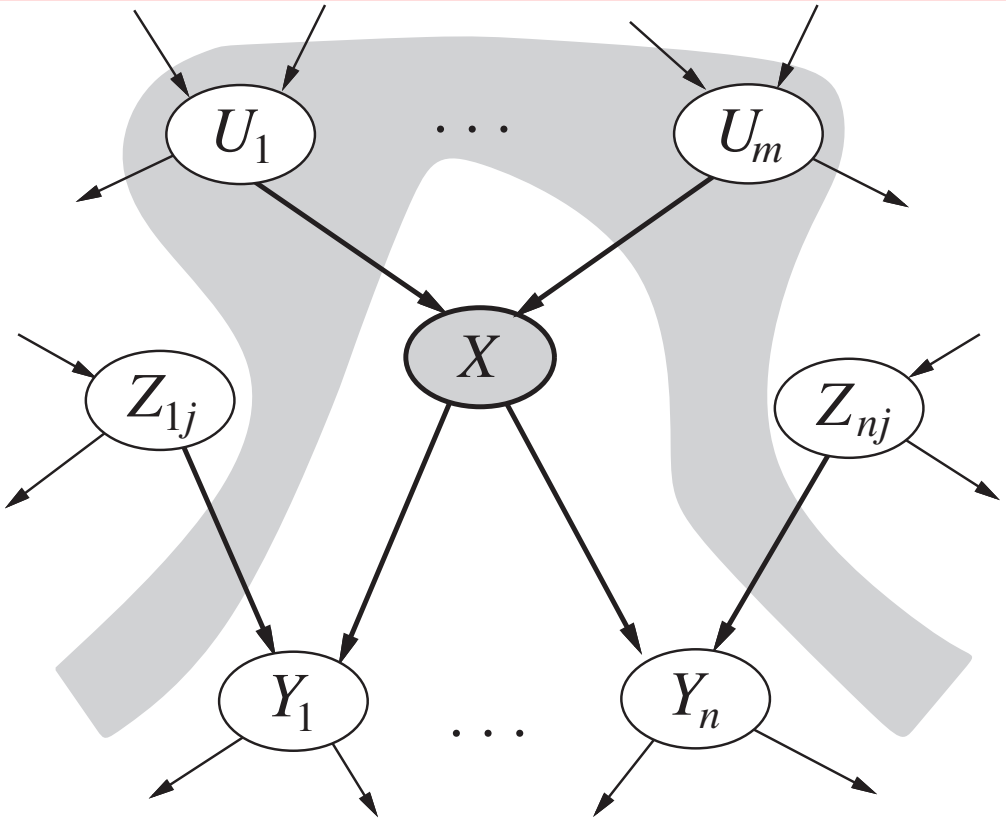
Bayes Net Semantics

- Bayes Nets implicitly encode joint distributions
  - As a product of conditional distributions

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$$

Bayes Net Semantics

Assumption: Every variable is conditionally independent of its non-descendants given its parents



A node X is conditionally independent of its non-descendants (e.g., the Z_{ij}s) given its parents (the U_is shown in the gray area).

Bayes Nets: Big Picture

- Two problems with using full joint distribution tables:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly. For n variables with domain size d, joint table has d^n entries - exponential in n.

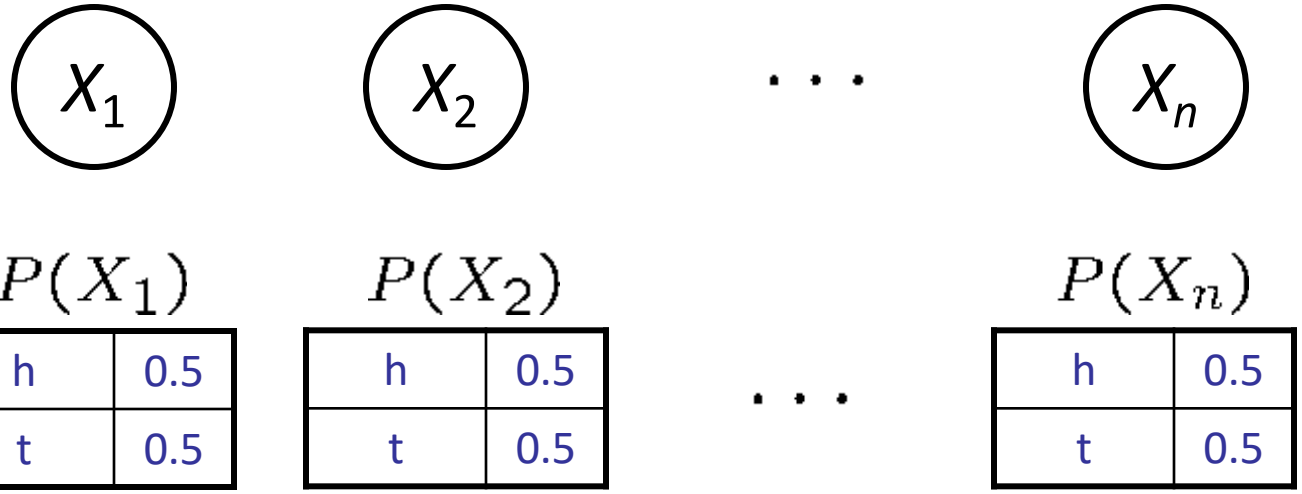
- Bayes nets: a technique for describing complex joint distributions using simple, local distributions (conditional probabilities)

Example: Coin Flips

- N independent coin flips

- No interactions between variables: absolute independence

Example: Coin Flips

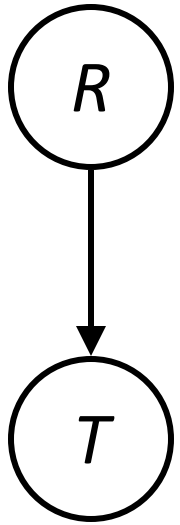


$$P(h, h, t, h) = \quad 0.5 * 0.5 * 0.5 * 0.5$$

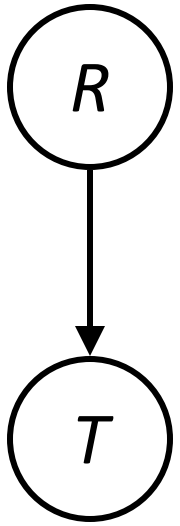
Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic

- Model: rain causes traffic



Example: Traffic



$P(R)$

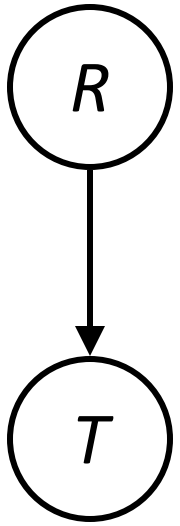
+r	1/4
-r	3/4

$P(T|R)$

T	R	P(T|R)
+t	+r	3/4
+t	-r	1/2
-t	+r	1/4
-t	-r	1/2

$$P(+r, -t) =$$

Example: Traffic



$P(R)$

+r	1/4
-r	3/4

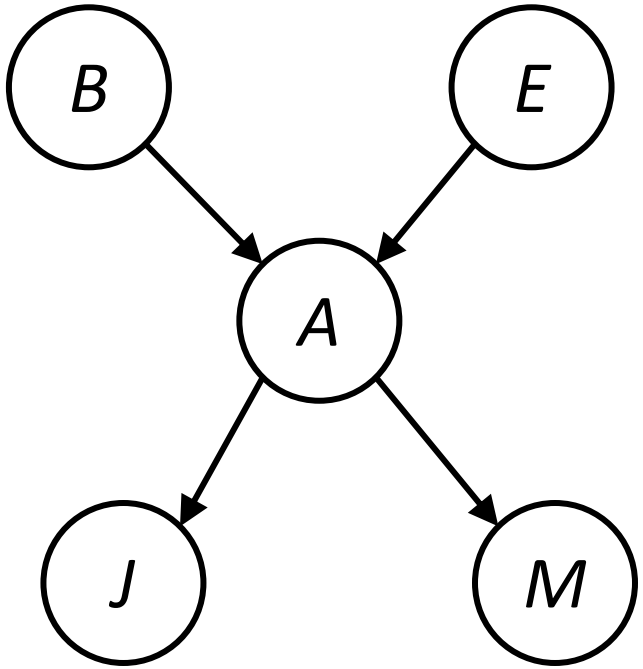
$P(T|R)$

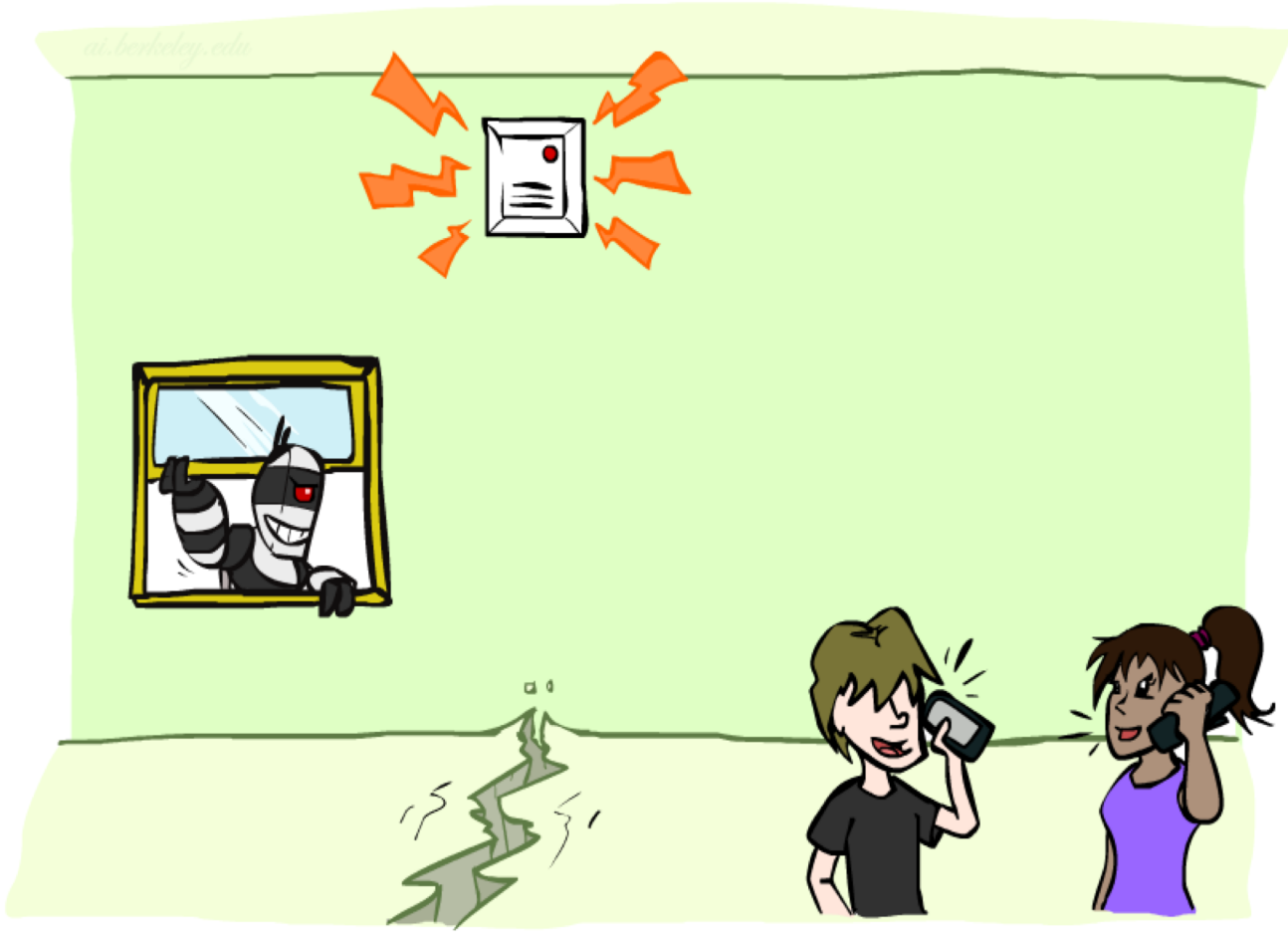
T	R	P(T|R)
+t	+r	3/4
+t	-r	1/2
-t	+r	1/4
-t	-r	1/2

$P(+r, -t) =$ 1/16

Example: Alarm Network

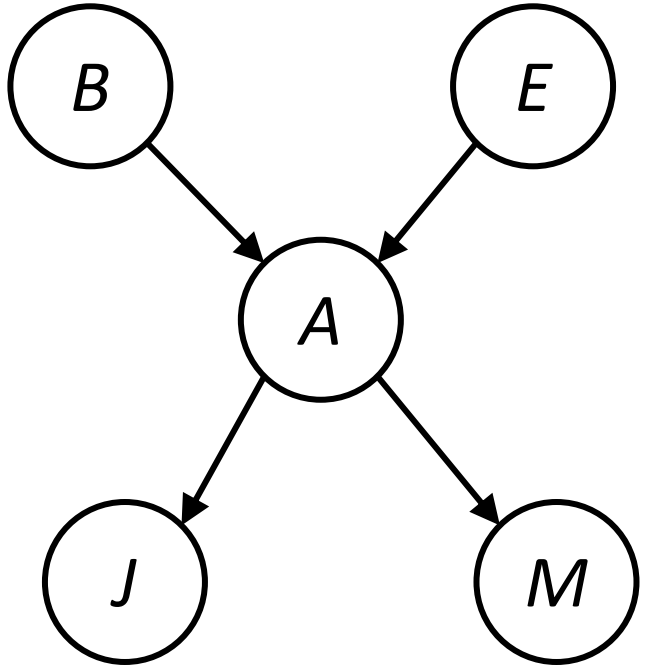
- Variables
  - B: Burglary
  - E: Earthquake
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls





Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

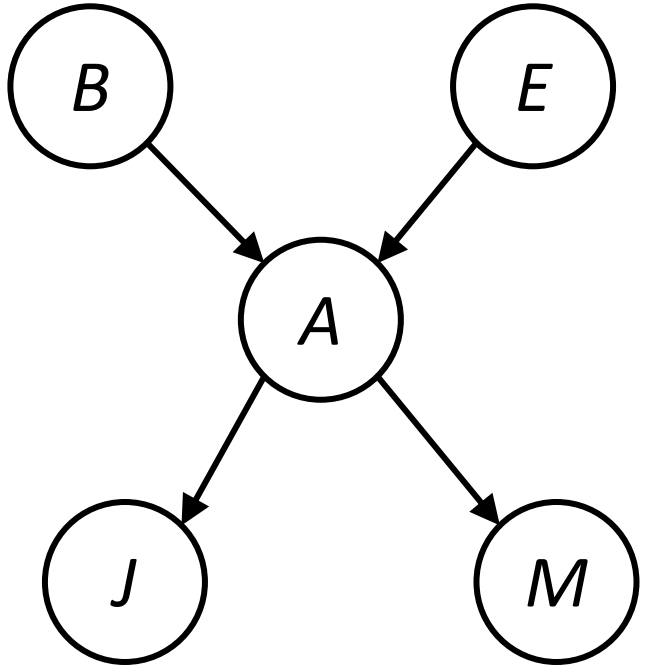
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

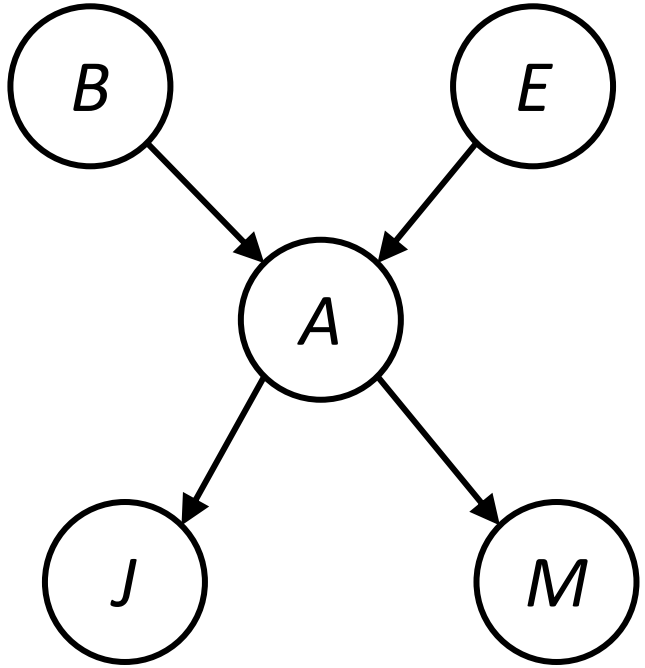
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

$$P(+b, -e, +a, -j, +m) =$$

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

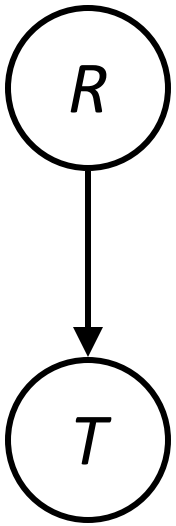
$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Example: Traffic

- Causal direction



$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

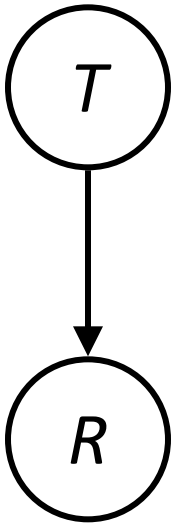
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

$$P(x_i|x_1, \ldots x_{i-1}) = P(x_i|parents(X_i))$$

Probabilities in BNs

- Why are we guaranteed that setting

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$$

results in a proper joint distribution?

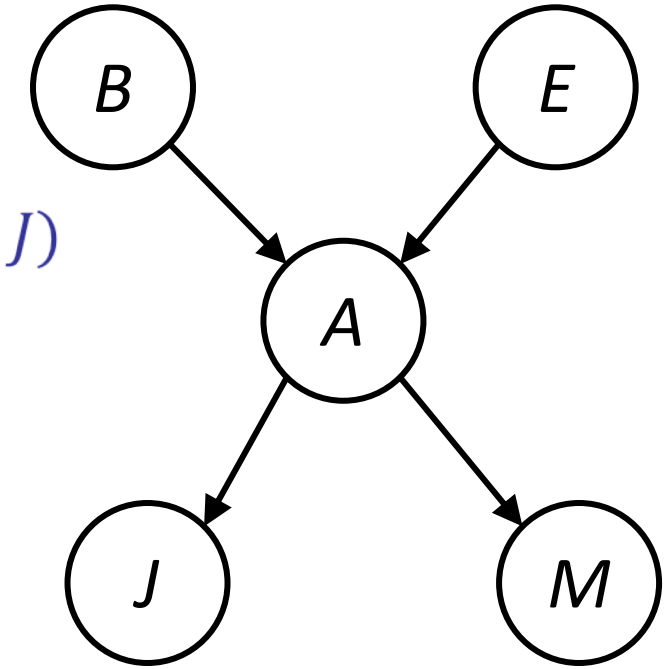
- Chain rule (valid for all distributions): $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, X_2, ..., X_{i-1})$
- Assume conditional independences: $P(X_i | X_1, X_2..., X_{i-1}) = P(X_i | parents(X_i))$

→ Consequence: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$

Assumption: Every variable is conditionally independent of its non-descendants given its parent

Example: Alarm Network

- Generic Chain Rule: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, X_2, ..., X_{i-1})$



$$P(B, E, A, J, M) = P(B)\, P(E|B)\, P(A|B, E)\, P(J|B, E, A)\, P(M|B, E, A, J)$$

$$P(B, E, A, J, M) = P(B)\, P(E) \quad P(A|B, E)\, P(J|A) \quad P(M|A)$$

- Bayes Nets: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$

Size of a Bayes Net

- How big is a joint distribution over N Boolean variables?

  2^N

- How big is an N-node bays net if nodes have up to k parents?

  $O(N * 2^{k+1})$

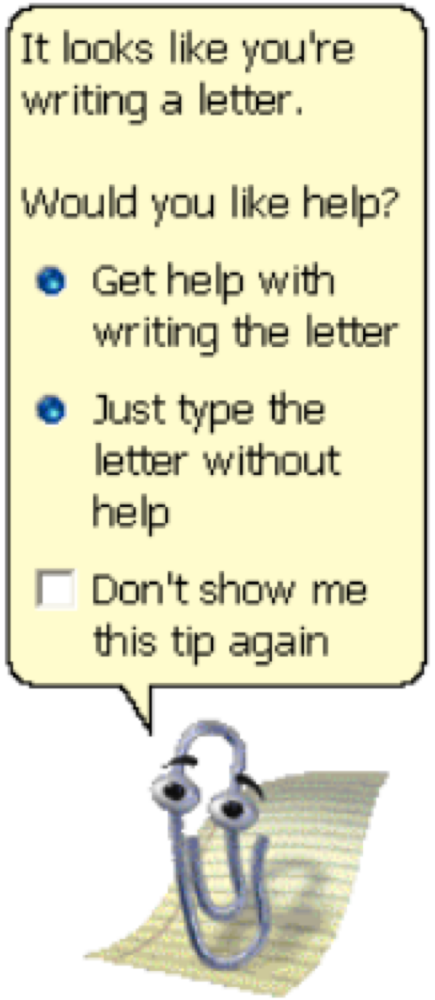
- Both give you the power to calculate

$$P(X_1, X_2, \ldots X_n)$$

- BNs: Huge space savings!

Example Bayes Net: Office Clip

- Older versions of MS Office used Bayes Nets to run the "Intelligent Assistant" program, including animated paper-clip, "Clippy"
- Tracked user behavior to see if it should suggest help, and to determine what sort of help the user might need
- Probably the least popular Bayes Net in the history of mankind!



Example Bayes Net: Office Clip

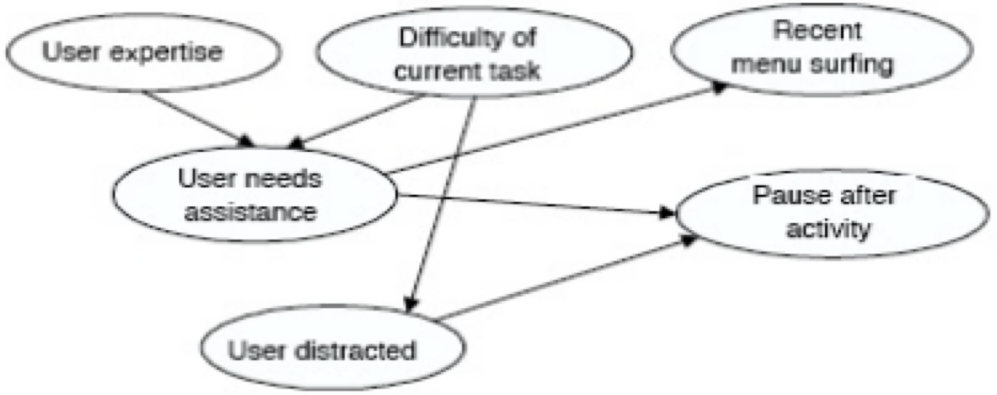
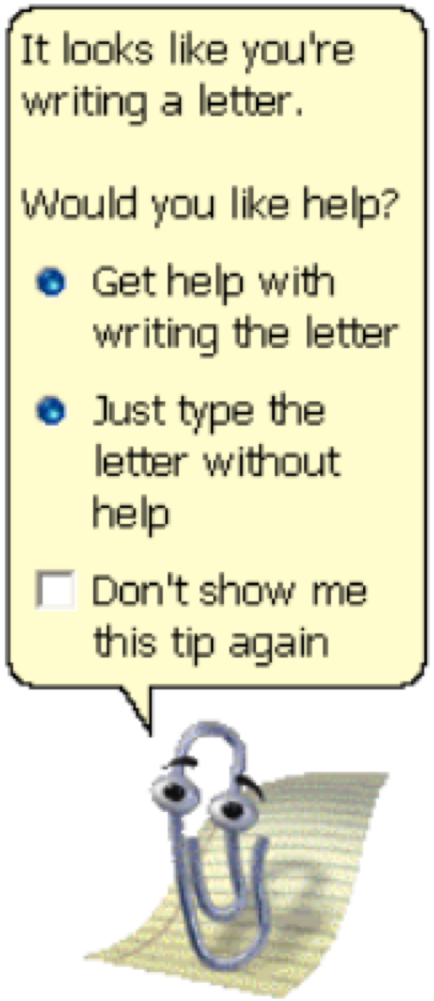


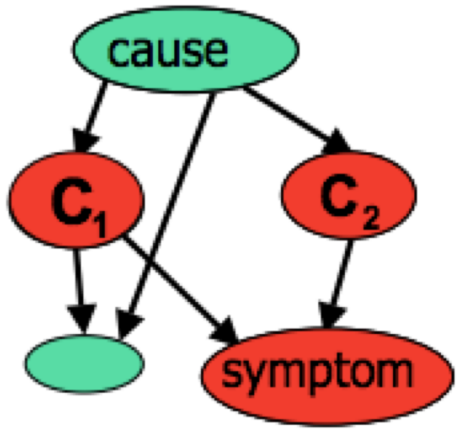
Figure 2: A portion of a Bayesian user model for inferring the likelihood that a user needs assistance, considering profile information as well as observations of recent activity.

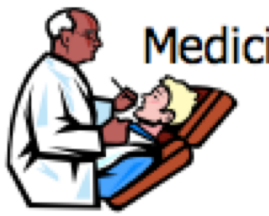


- The BN was used to predict the "Needs Assistance" variable
- Reasoning based on prior distributions over how hard certain things were to do in Office, and on how expert users were likely to be
- Also used evidence, taken from things like menu use, clicking, waiting, re-doing or un-doing things…

Other Applications

- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: $\max_{class}$ P(class|data)
- Decision-making (given a cost function)

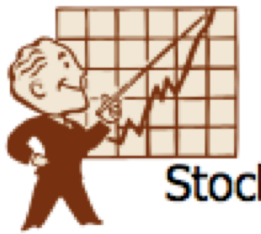




Medicine

Speech recognition

Bio-informatics



Stock market

Text Classification

Computer troubleshooting

Reading

- Read Sections 14.1, 14.2, and 14.4 in the AIMA textbook (Third Edition)