



CMU 15-781

Lecture 8:

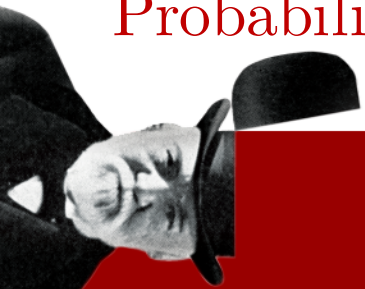
Bayesian Networks I

Teacher:

Gianni A. Di Caro

NEED FOR PROBABILISTIC MODELING

- So far, our rational problem-solving or planning agent enjoyed:
 - **Known environment:** It can access all the necessary information to model the world and to list the actions that can be taken in each state together with their effects
 - **Full observability:** Its “sensors” give access to the complete state of the environment
 - **Deterministic world:** The next state s_{t+1} of the environment is completely determined by (s_t, a_t)
- Unfortunately, ignorance, laziness, sensing limitations, make **uncertainty** the signature of real-world scenarios → need for **Probabilistic models** for knowledge representation & reasoning







PROBABILISTIC INFERENCE

- A probability model is completely determined by the **joint probability distribution** for all the random variables $\{X_1, X_2, \dots, X_n\} \rightarrow P(X_1, X_2, \dots, X_n)$
- The joint probability distribution is the “*knowledge base*” from which answers to all questions may be derived
- **Probabilistic inference:** Compute probability of a **query** variable (or variable set) taking on a value (or set of values) given some **evidence** on a subset of variables
- *Posterior probability* $\Pr[X_j = x_j \mid X_1=e_1, \dots, X_k=e_k]$?



RECALL: JOINT PROBABILITY DISTRIBUTION

X_1	Sedan	SUV	Coupe	Truck
X_2				
●	.05	.2	0	.1
●	.1	0	.1	0
●	0	.1	.05	.1
●	.1	0	.1	0

$X_1 \sim \text{Car} \{ \text{Sedan, SUV, Coupe, Truck} \}$

$X_2 \sim \text{Color} \{ \text{red, blue, yellow, black} \}$

X_i are discrete rv $\rightarrow P(X_1, X_2)$ can be given in tabular form: $4 \times 4 = 16$ parameter values need to be assigned

- Joint probability distribution:

$$P(\text{Car}=x \wedge \text{Color}=y) \quad \forall (x,y) \in X_1 \times X_2$$

- Joint probability:

$$P(\text{Car}=\text{Sedan} \wedge \text{Color}=\text{red}) = 0.05$$



RECALL: MARGINAL PROBABILITY DISTRIBUTION

$X_1 \backslash X_2$	Sedan	SUV	Coupe	Truck
●	.05	.2	0	.1
●	.1	0	.1	0
●	0	.1	.05	.1
●	.1	0	.1	0

- Marginal probability:
 $P(\text{Car}=\text{Sedan})?$ $P(\text{Color}=\text{red})?$

- Marginal probability distribution: the distribution of *one variable* ignoring all other variables (i.e., given all their possible outcomes)

$$P(X_j) = \sum_{\substack{X_i, i=1, \dots, n \\ i \neq j}} P(X_1, X_2, \dots, X_n)$$

$$P(X_1 = \text{SUV}) = \sum_{X_2=r,b,y,bl} P(X_1 = \text{SUV}, X_2) = 0.2 + 0 + 0.1 + 0 = 0.3$$



RECALL: MARGINAL PROBABILITY DISTRIBUTION

- Marginal distribution of a subset of variables
- V is a set of $n > 2$ variables, we want to compute the marginal probability distribution for a subset of two of them, X_1, X_2

$V = \{X_1, X_2, \mathbf{Y}\}$ where \mathbf{Y} is a set of additional variables

$$P(X_1, X_2) = P(\text{Marginal of } X_1, X_2) = \sum_{\mathbf{y} \in \mathbf{Y}} P(X_1, X_2, \mathbf{y})$$



RECALL: CONDITIONAL PROBABILITY DISTRIBUTION

$X_1 \backslash X_2$	Sedan	SUV	Coupe	Truck
● (Red)	.05	.2	0	.1
● (Blue)	.1	0	.1	0
● (Yellow)	0	.1	.05	.1
● (Black)	.1	0	.1	0

- Conditional probability:
 $P(\text{Car}=\text{Sedan} \mid \text{Color}=\text{red})?$
- Conditional probability distribution: the joint distribution of X_1 given the value of the other variable X_2

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)} = \frac{\text{Joint}(X_1, X_2)}{\text{Marginal}(X_2)}$$

$$P(X_1 = \text{SUV} \mid X_2 = \text{red}) = \frac{P(X_1 = \text{SUV}, X_2 = \text{red})}{P(X_2 = \text{red})} = \frac{0.2}{0.05 + 0.2 + 0 + 0.1} = 0.57$$



RECALL: CONDITIONAL PROBABILITY DISTRIBUTION

- **Conditional probability distribution:** the distribution of X_i *given* the value of the all other variables $X_j, j=1, \dots, n, j \neq i$

$$P(X_1 | X_2, X_3, \dots, X_n) = \frac{P(X_1, X_2, X_3, \dots, X_n)}{P(X_2, X_3, \dots, X_n)} = \frac{\text{Joint}(\text{all variables})}{\text{Marginal}(\text{given variables})}$$

- **Conditional probability distribution:** the distribution of a subset of X_i *given* the value of all the other variables

$$P(X_1, \dots, X_q | X_{q+1}, \dots, X_n) = \frac{P(X_1, \dots, X_n)}{P(X_{q+1}, \dots, X_n)} = \frac{\text{Joint}(\text{all variables})}{\text{Marginal}(\text{given variables})}$$

- For three variables: $P(X_1 | X_2, X_3) = \frac{P(X_1, X_2, X_3)}{P(X_2, X_3)}$



RECALL: NORMALIZATION CONSTANTS

- When using conditionals, marginals play the role of normalization constants

$$P(X_1 = \text{SUV} \mid X_2 = \text{blue}) = \frac{P(X_1 = \text{SUV}, X_2 = \text{blue})}{P(X_2 = \text{blue})}$$

$$P(X_1 = \text{Sedan} \mid X_2 = \text{blue}) = \frac{P(X_1 = \text{Sedan}, X_2 = \text{blue})}{P(X_2 = \text{blue})}$$

$$\Rightarrow P(X_1 = x \mid X_2 = \text{blue}) = \alpha P(X_1 = x, X_2 = \text{blue})$$

Is a **normalization** constant for the distribution $P(X_1 \mid X_2 = \text{blue})$



RECALL: CHAIN/PRODUCT RULE

- $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(\text{Conditional})P(\text{Marginal}(\text{given}))$
- $P(X_1, X_2, X_3) = P(X_3 | X_2, X_1)P(X_2 | X_1)P(X_1)$
- $P(X_1, X_2, X_3, X_4) = P(X_4 | X_3, X_2, X_1)P(X_3 | X_2, X_1)P(X_2 | X_1)P(X_1)$
- ...
- $P(X_1, \dots, X_n) = P(X_n | X_{n-1}, \dots, X_1)P(X_{n-1} | X_{n-2}, \dots, X_1) \cdots P(X_2 | X_1)P(X_1)$

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \bigcap_{j=1}^{i-1} X_j)$$



RECALL: MARGINALS BY CONDITIONING

Conditioning: Using the product rule, marginals can be computed using conditional probabilities:

$$P(X_1) = \sum_{\substack{X_i \\ i=2, \dots, n}} P(X_1, X_2, \dots, X_n) = \sum_{\substack{X_i \\ i=2, \dots, n}} P(X_1 \mid X_2, \dots, X_n) P(X_2, \dots, X_n)$$

Sums have to be intended as summing up over all possible combinations of values of the set of variables X_2, \dots, X_n



RECALL: INDEPENDENCE

- Two RVs are independent if: $P(X_1, X_2) = P(X_1)P(X_2)$
- n variables are independent if: $P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2)\dots P(X_n)$

Dependence

x, y	$P(X=x, Y=y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

Joint

x	$P(X=x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y=y)$
on-time	0.45
late	0.55

Marginal

Independence

x, y	$P(X=x, Y=y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

Joint

x	$P(X=x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y=y)$
fly-United	0.9
fly-Northwest	0.1

Marginal

$$P(x)P(y) = P(x, y) \quad \forall (x, y) \in X \times Y$$



RECALL: CONDITIONAL INDEPENDENCE

- X_1 and X_2 are *conditionally independent given X_3* if:
$$P(X_1 | X_2, X_3) = P(X_1 | X_3)$$
- Once the value of X_3 is known, knowing X_2 doesn't tell anything about X_1
- Another way of saying the same thing is:
$$P(X_1, X_2 | X_3) = P(X_1 | X_3) P(X_2 | X_3)$$
- In a sense, the dependence between X_1 and X_2 “dissolves” once the knowledge about X_3 is made available



RECALL: CONDITIONAL INDEPENDENCE

- *Example:* John and Mary work in the same place and come to work by different transportation means. In spite of this, the fact that John is late at work might say something about the fact that also Mary is late (e.g, because of train strike that would affect all transportation means). Therefore, John being late and Mary being late are not independent events
- However, if we know that there is a train strike (*given evidence*), then knowing about John being late doesn't add anything about Mary being late
- → The two events are conditionally independent given the knowledge of train strike



RECALL: CONDITIONAL INDEPENDENCE

Flu	Fever	Vomit	P
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Are Fever and Vomit independent?

NO.

e.g. $P(\text{fever}, \text{vomit}) \neq P(\text{fever}) \times P(\text{vomit})$

RECALL: CONDITIONAL INDEPENDENCE

Flu	Fever	Vomit	P
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Are Fever and Vomit conditionally independent given Flu: YES.

$$P(\text{fever}, \text{vomit} \mid \text{flu}) = P(\text{fever} \mid \text{flu}) \times P(\text{vomit} \mid \text{flu})$$

$$P(\text{fever}, \text{vomit} \mid \neg \text{flu}) = P(\text{fever} \mid \neg \text{flu}) \times P(\text{vomit} \mid \neg \text{flu})$$

etc.



RECALL: BAYES RULE

$$P(X_1 | X_2) = \frac{P(X_2 | X_1)P(X_1)}{P(X_2)} = \frac{P(X_2 | X_1)P(X_1)}{\sum_{X_1} P(X_2 | X_1)P(X_1)}$$

- In some (many) cases it's easier to estimate $P(A|B)$ rather than $P(B|A)$
- For instance, A =symptom, B =cause

$$P(X_1 | X_2, X_3) = \frac{P(X_3 | X_2, X_1)P(X_2 | X_1)P(X_1)}{P(X_2)P(X_3 | X_2)}$$



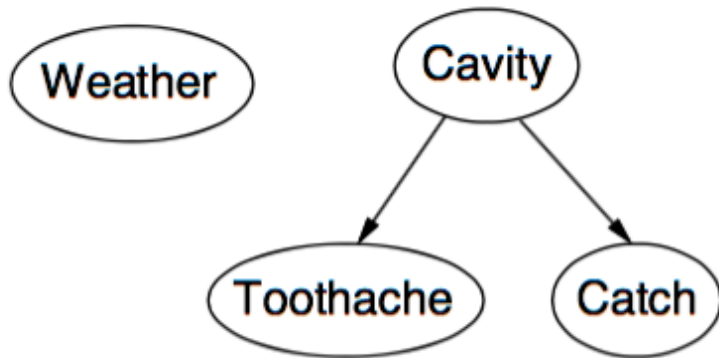
BAYESIAN NETWORKS

- Graphical data structure for conditional independence assertions → A compact specification of full joint distributions
- A set of **nodes**, one per variable
- A directed, acyclic **graph** (link \approx “directly influences”)
 - A node is a *parent of a child* (or successor) if there is an arc from the former to the latter
 - If there is a directed chain of nodes, a node is *ancestor* of another if it appears earlier in the chain, a *descendant* if it appears later



BAYESIAN NETWORKS

- A *conditional distribution* for each node *given* its parents: $P(X_i \mid \text{Parents}(X_i))$
- It can be a **Conditional Probability Table (CPT)** giving the distribution over X_i for each combination of the parent values



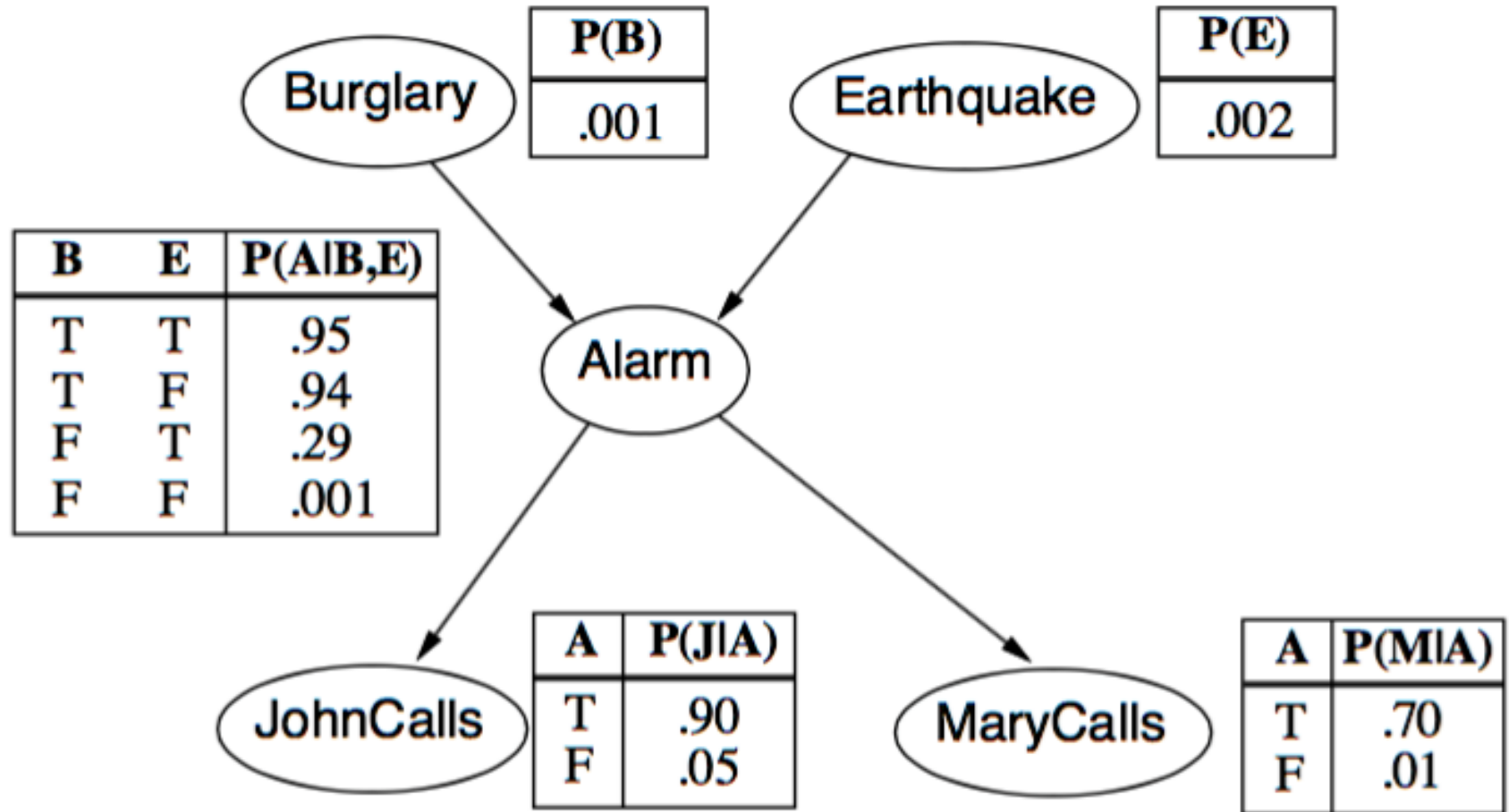
Weather is conditionally independent of other variables
Toothache and *Catch* are conditionally independent given *Cavity*

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: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects “causal” knowledge:
 - A burglar can set the alarm off
 - An earthquake can set the alarm off
 - The alarm can cause Mary to call
 - The alarm can cause John to call



EXAMPLE



A COMPACT REPRESENTATION

- A CPT for binary variables X_i each with k parents has 2^k rows for the combination of parent values
- If each variable has no more than k parents, the network requires $O(n2^k)$ parameter values to specify the CPTs
- \rightarrow The BN grows linearly with n , while the full joint distribution goes as $O(2^n)$
- E.g., in the burglary example: $1+1+4+2+2=10$ numbers, vs. $2^5-1=31$ for the full joint distribution



GLOBAL SEMANTICS

Global semantics: define the full joint distribution as the product of the local conditional distributions

$$\begin{aligned} P(x_1, x_2, \dots, x_n) &\equiv P(X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_n = x_n) \\ &= \prod_{i=1}^n P(x_i \mid \text{Parents}(X_i)) \quad * \end{aligned}$$

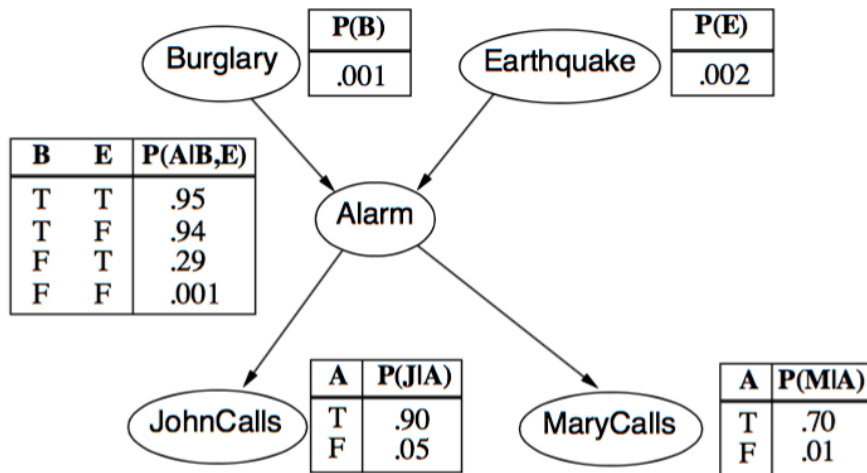
- This defines what a Bayes network means: a compact representation of a joint probability distribution in terms of conditional distribution



NUMERICAL EXAMPLE

- $J=T, M=T, A=T, B=T, E=T$ is indicated as j, m, a, b, e

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



$P(b|o)$?

It's not in the conditional distributions of the network. It needs to be *computed* by inference → next lecture!

GLOBAL SEMANTICS

- The derivation stems from the *chain rule*:

$$P(x_1, \dots, x_n) = P(x_n \mid x_{n-1}, \dots, x_1)P(x_{n-1} \mid x_{n-2}, \dots, x_1) \cdots P(x_2 \mid x_1)P(x_1)$$

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid x_{i-1}, \dots, x_1)$$

- This holds for any set of variables, in particular, by numbering the variables in a way that is consistent with the partial order implicit in the graph structure, it becomes equivalent to (*)



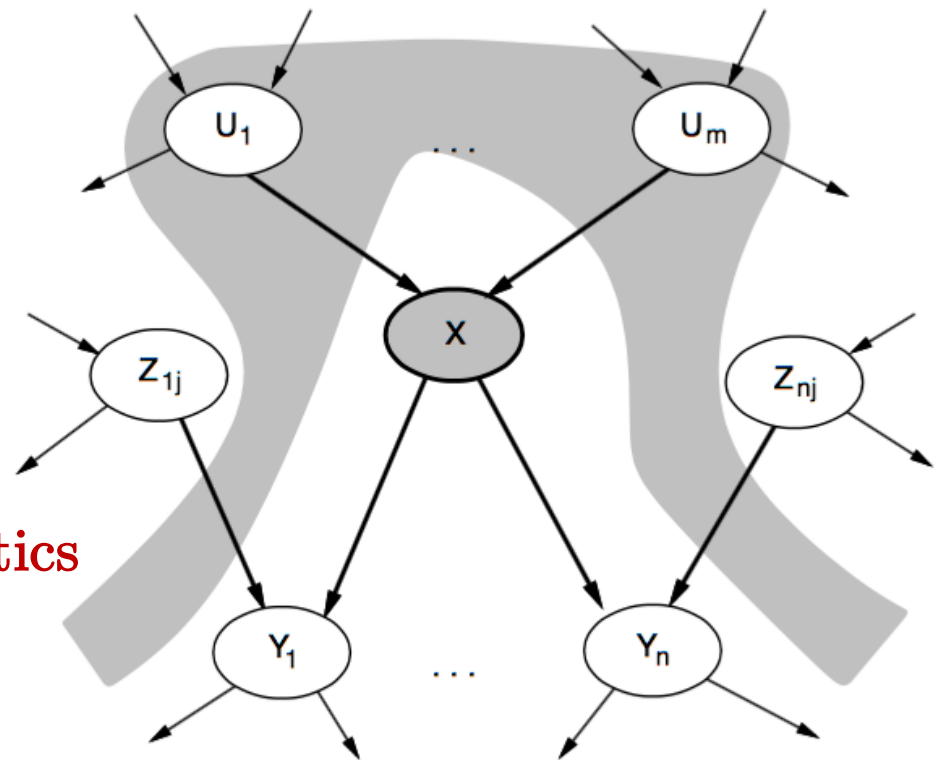
LOCAL SEMANTICS

Local semantics: each node is **conditionally independent** of its non-descendants *given* its parents

$U_1, U_m = \text{Parents}$

$Y_1, Y_n = \text{Descendants}$

$Z_{ij} = \text{Other nodes from which } X \text{ is conditionally independent}$

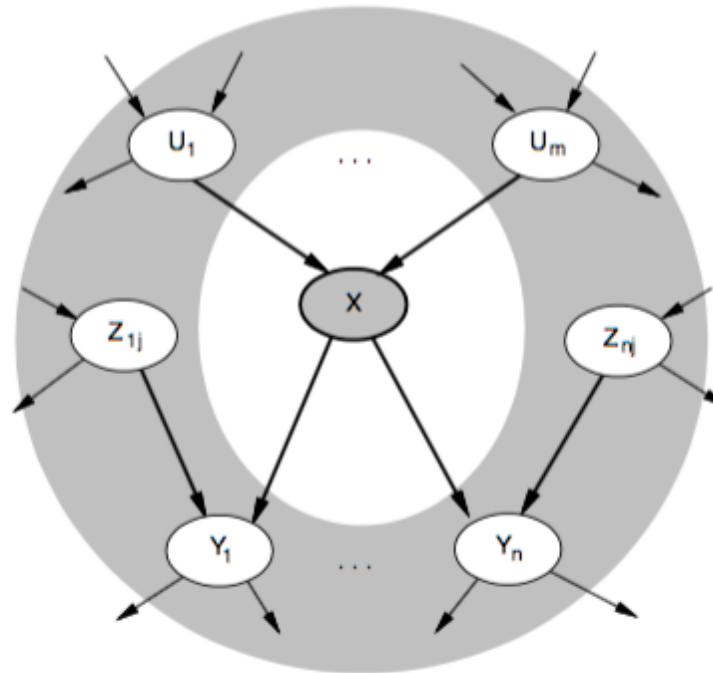


Local semantics \Leftrightarrow Global semantics



MARKOV BLANKET

Each node is **conditionally independent** of all others *given* its **Markov blanket**: parents + children + children's parents



CONSTRUCTION OF A BN

A method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables X_1, \dots, X_n
2. For $i = 1$ to n
add X_i to the network
select parents from X_1, \dots, X_{i-1} such that
$$\mathbf{P}(X_i | \text{Parents}(X_i)) = \mathbf{P}(X_i | X_1, \dots, X_{i-1})$$

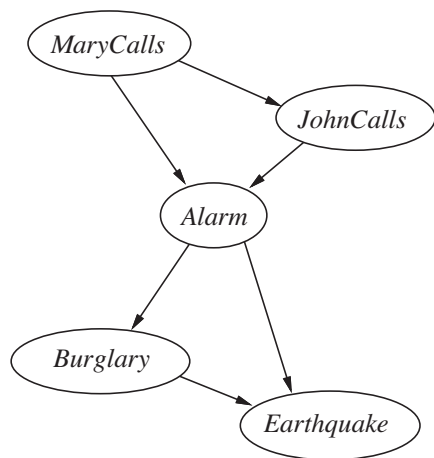
This choice of parents guarantees the global semantics:

$$\begin{aligned}\mathbf{P}(X_1, \dots, X_n) &= \prod_{i=1}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1}) \quad (\text{chain rule}) \\ &= \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i)) \quad (\text{by construction})\end{aligned}$$



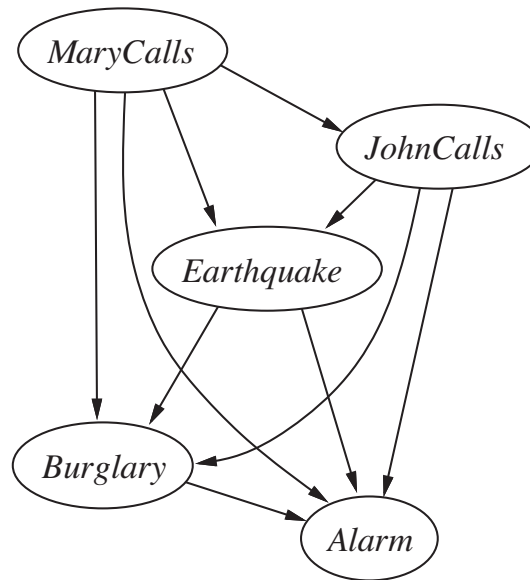
ORDER MATTERS (FOR EFFICIENCY)

- Parents of X_i should contain all those nodes in X_1, \dots, X_{i-1} that directly influence X_i : arcs should go from causes to effects, rather than the reverse, that would require to specify additional dependencies among otherwise independent causes

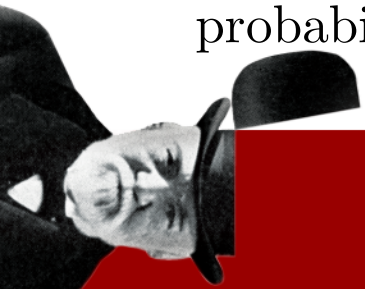


- Order: M, J, A, B, E*
- If $M=T$ then likely $Alarm=T$, which makes likely that $J=T \rightarrow M$ parent of J
- If both $M=T$ & $J=T$, then it's likely that $A=T \rightarrow M$ and J parents of A
- If $A=T$ then M and J do not give any information about $B \rightarrow A$ is the only parent of B
- If $A=T$ then E is likely, but if also $B=T$, then this would explain $A \rightarrow$ Both A and B are parents of E

ORDER MATTERS (FOR EFFICIENCY)

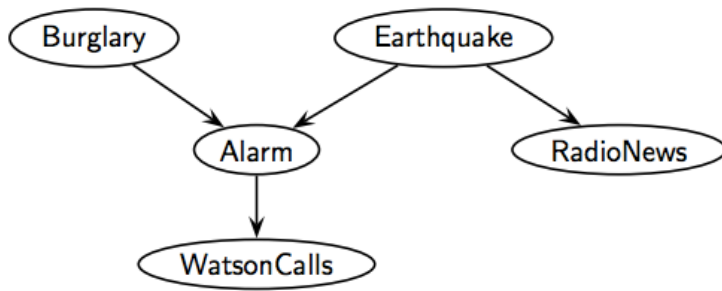


- *Order: M, J, A, B, E*
- 31 parameter values are needed to specify the CPTs!
- A number of “unnatural” probability judgments need to be quantified
- Independently from the order, *any* network represents the same joint probability distribution



INDEPENDENCE & INFORMATION PROPAGATION

- **Question:** Given a BN structure what **dependence and independence relationships** are represented?
- How **information propagates** over the network?

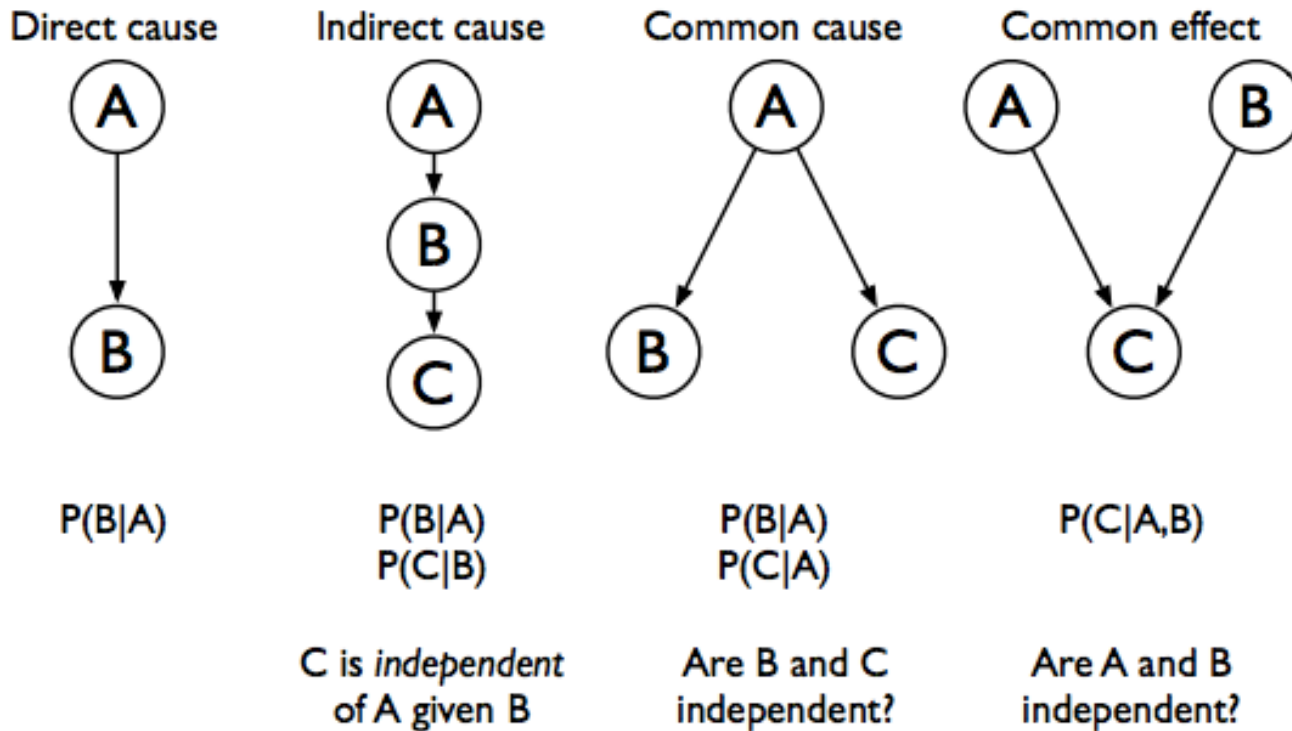


A	B	C	A and B are independent given C
Burglary	Earthquake	WatsonCalls	No
Burglary	Earthquake	Alarm	No
Burglary	WatsonCalls		No
Burglary	RadioNews	WatsonCalls	No
Burglary	RadioNews	Alarm	No
Earthquake	WatsonCalls		No
Alarm	RadioNews		No
RadioNews	WatsonCalls		No
Burglary	Earthquake		Yes
Burglary	WatsonCalls	Alarm	Yes
Burglary	RadioNews		Yes
Earthquake	WatsonCalls	Alarm	Yes
Alarm	RadioNews	Earthquake	Yes
RadioNews	WatsonCalls	Earthquake	Yes
RadioNews	WatsonCalls	Alarm	Yes

15 of the total of 53 dependence and independence statements encoded in the BN

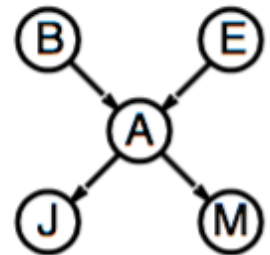
INDEPENDENCE & INFORMATION PROPAGATION

Types of probabilistic relationships



INDEPENDENCE & INFORMATION PROPAGATION: DIRECT CONNECTION

- If X and Y are connected by an edge, then X and Y are dependent
- Information can be transmitted over one edge
- Burglary and Alarm are dependent:
 - Knowing that a burglary has taken place increases the belief that the alarm went off
 - Knowing that the alarm went off increases the belief that there has been a burglary



DIRECT CONNECTION

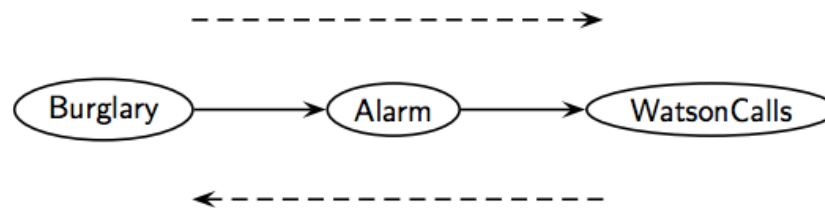
- If X and Y are connected by an edge, then X and Y are dependent. Information can be transmitted over one edge



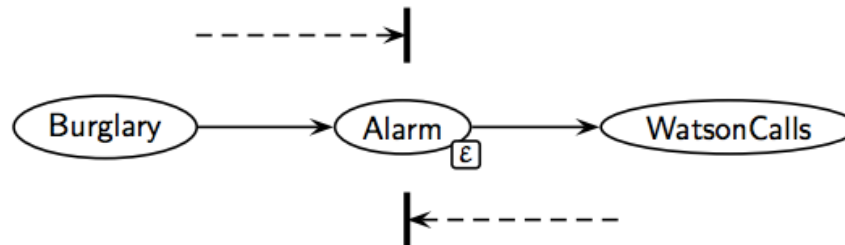
- Burglary and Alarm are dependent:
 - Knowing that a burglary has taken place increases the belief that the alarm went off
 - Knowing that the alarm went off increases the belief that there has been a burglary.

SERIAL CONNECTION

- If A is not observed, B and WC are dependent
- Information can be transmitted between B and WC through A if A is not observed

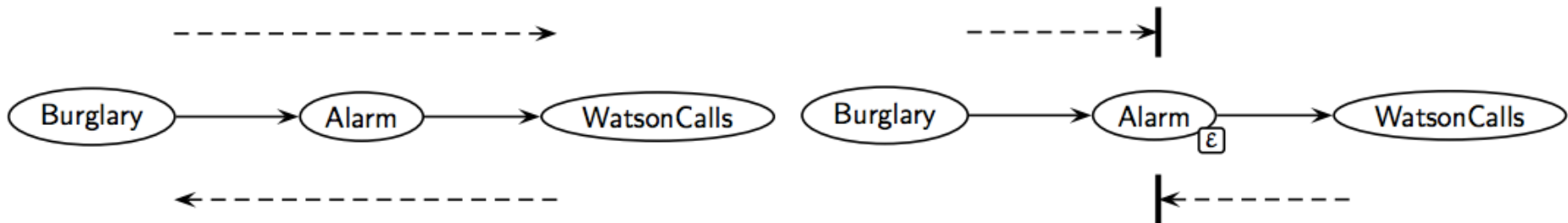


- If A is observed, B and WC are independent
- Information cannot be transmitted between B and WC through A. Observing A *blocks the information path*



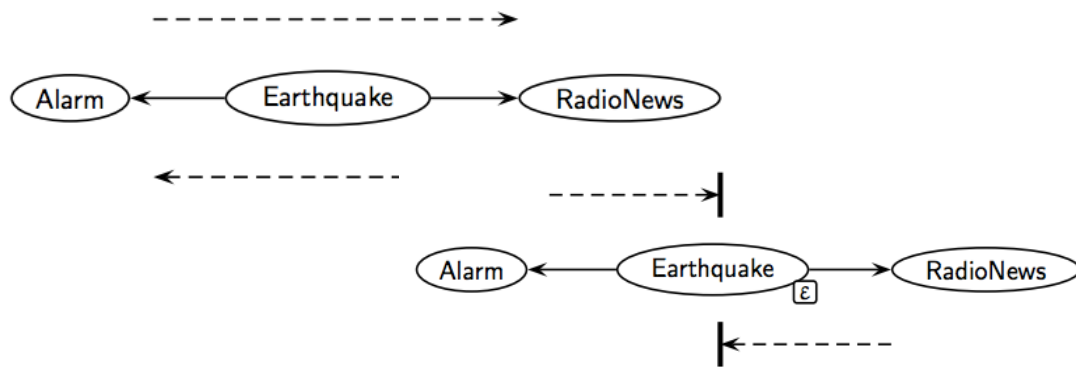
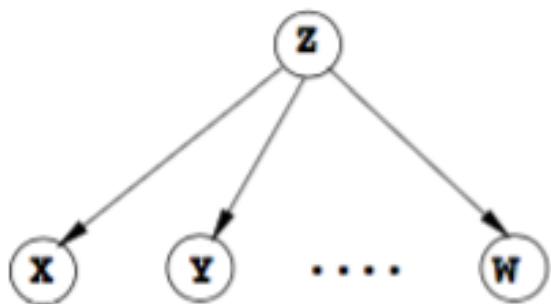
SERIAL CONNECTION

- *In absence of hard evidence on the middle variable Alarm, evidence on Burglary updates our belief on Alarm, and in turn this affects our belief about the state of WatsonCalls. The opposite is also true.*
- *If we have hard evidence on Alarm, any information about the state of Burglary will not make us change our belief about WatsonCalls (and vice versa)*



DIVERGING CONNECTION (COMMON CAUSE)

- If Z is not observed, X, Y, \dots, W are dependent
- Information can be transmitted through Z among its children
- If Z is observed, X, Y, \dots, W are independent
- Information cannot be transmitted through Z among its children.
Observing Z *blocks the information path*



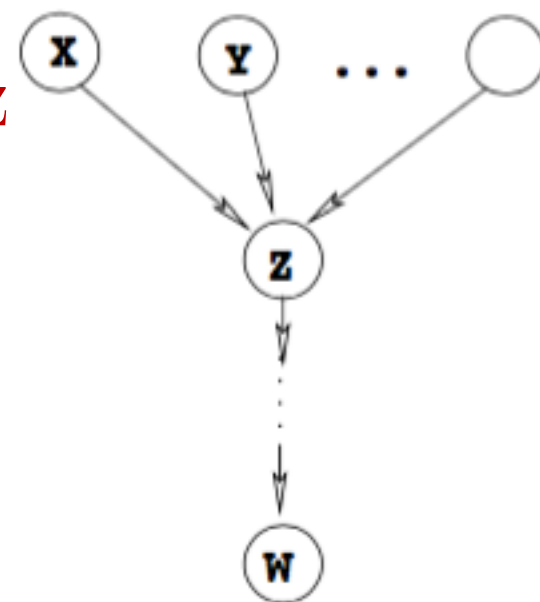
DIVERGING CONNECTION (COMMON CAUSE)

- If we have no hard evidence on Earthquake, then receiving information about Alarm influences our belief about Earthquake, as earthquake is a possible explanation for alarm. The updated belief about Earthquake in turn make us update our belief about the state of RadioNews. Similar arguments hold for the opposite case
- If the state of Earthquake is known, if information is received about the state of Alarm, this information is not going to change our belief about the state of Earthquake, and consequently we are not going to update our belief about RadioNews.



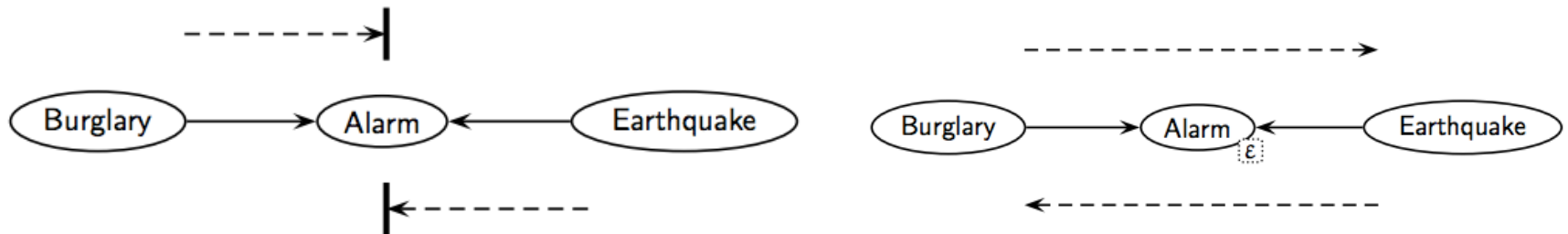
CONVERGING CONNECTION (COMMON EFFECT)

- If neither **Z** nor any of its descendants are observed, **X** and **Y** are independent.
- Information cannot be transmitted through **Z** among parents of **Z**.
- If **Z** or any of its descendants are observed, **X** and **Y** are dependent.
- Information can be transmitted through **Z** among parents of **Z** if **Z** or any of its descendants are observed. Observing **Z** or its descendants *opens the information path*.



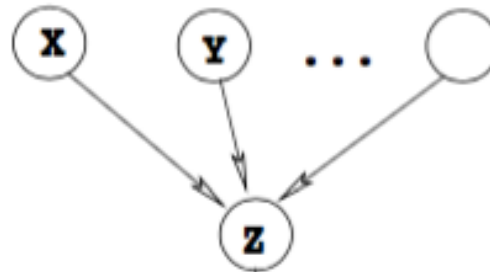
CONVERGING CONNECTION (COMMON EFFECT)

- *If no evidence is available on Alarm, then information on Burglary will not provide any derived information about Earthquake: burglary is not an indicator of earthquake, and vice versa.*
- *If evidence is available on Alarm, then information on Burglary will provide an explanation for the evidence that was received about the state of Alarm, and thus either confirm or dismiss Earthquake as the cause of the evidence received for Alarm. The opposite also holds true.*



EXPLAINING AWAY

- The property of converging connections, $X \rightarrow Z \leftarrow Y$, that information about the state of X (or Y) provides an explanation for an *observed effect* on Z, and hence confirms or dismisses Y (or X) as the cause of the effect, is often referred to as **explaining away** or as “*intercausal inference*”.

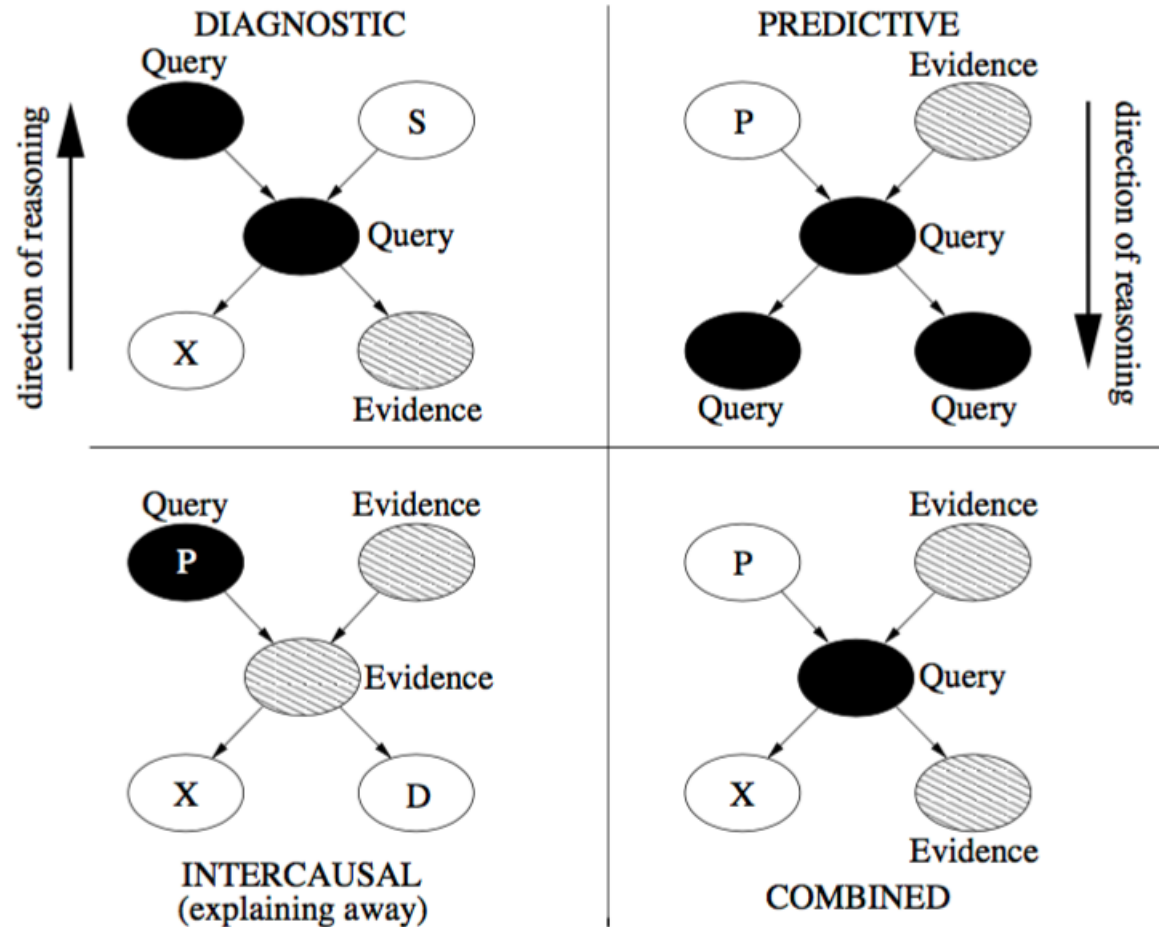


Getting a radio report on earthquake provides strong evidence that the earthquake is responsible for a burglar alarm, and hence *explaining away* a burglary as the cause of the alarm.



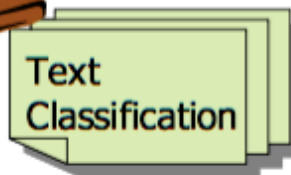
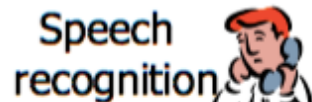
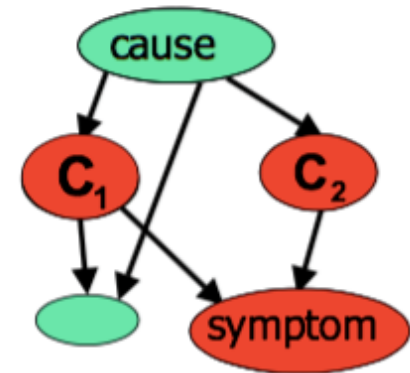
TYPES OF REASONING

Different choices for Query and Evidence variables



USE OF BAYESIAN NETWORKS

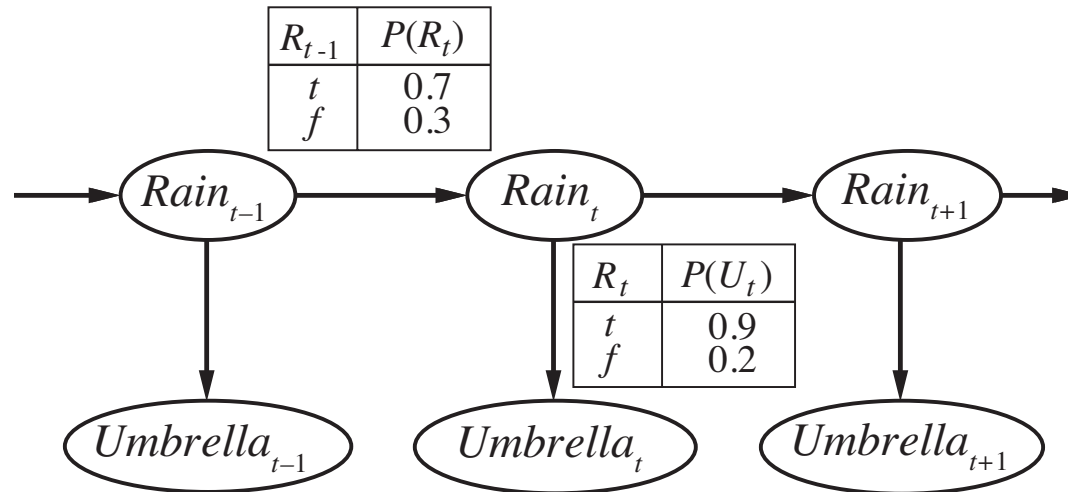
- Diagnosis: $P(\text{cause} \mid \text{symptom})$?
- Prediction: $P(\text{symptom} \mid \text{cause})$?
- Classification: $\max_{\text{class}} P(\text{class} \mid \text{data})$
- Decision-making (given a cost function)



USE OF BAYESIAN NETWORKS

Temporal models with hidden variables: Hidden Markov Models (HMM)

- The variable representing the state of the environment at time t , $Rain_t$ (T/F), is not directly observable (*hidden*) but defines causal dynamics
- The variable $Umbrella_t$ (T/F) is the *evidence* variable at time t
- **Filtering / State estimation:** Probability of $Rain_t$ given evidence in 1:t
- **Prediction:** Probability of $Rain_{t+1}$ given evidence in 1:t



PATIENT VISIT BAYES NET

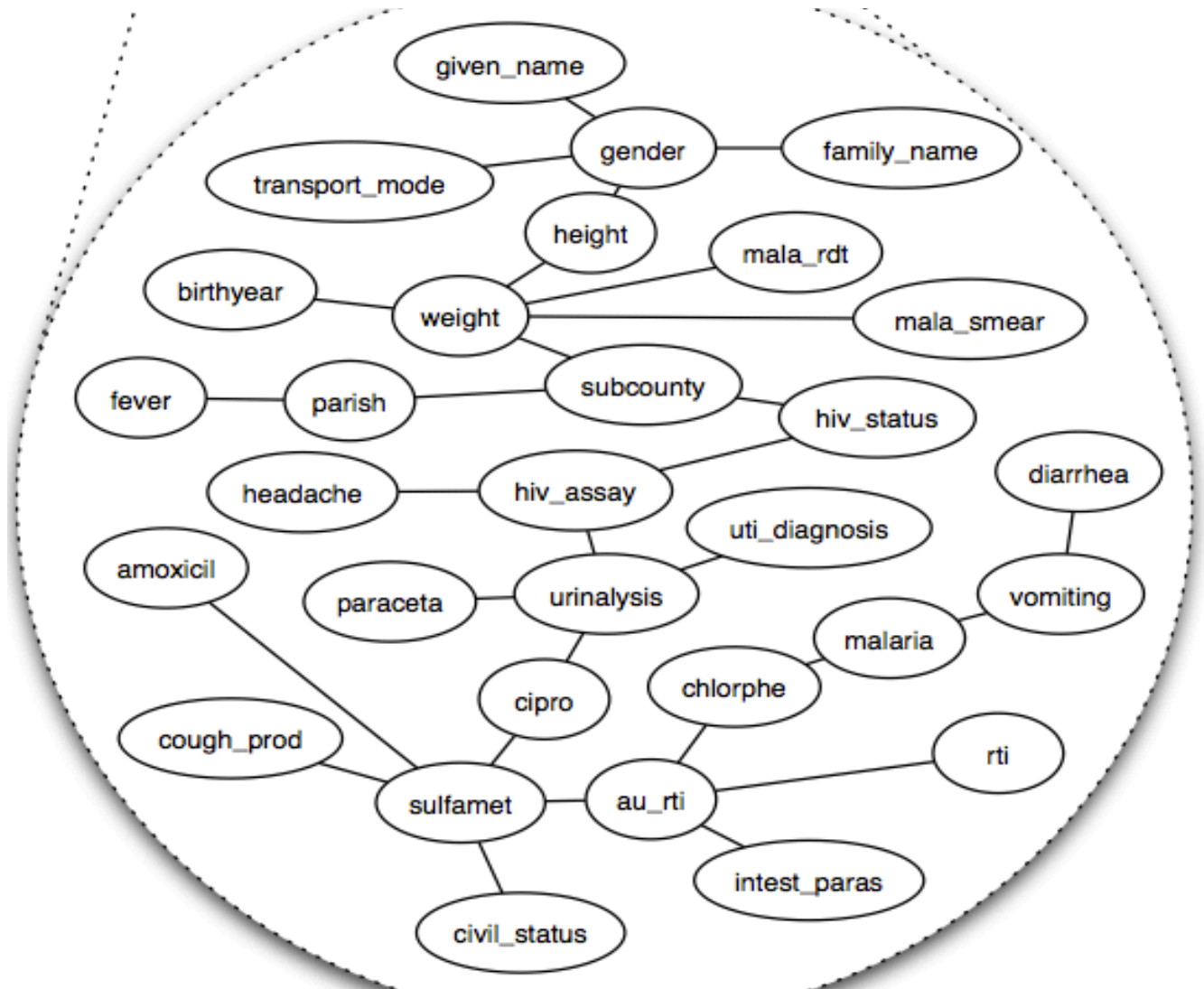


Figure: Chen,
Hellerstein,
Parikh, UIST
2010

