CS 561 Artificial Intelligence
Lecture # 3 Bayesian Networks

Rashmi Dutta Baruah

Department of Computer Science & Engineering

IIT Guwahati

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी

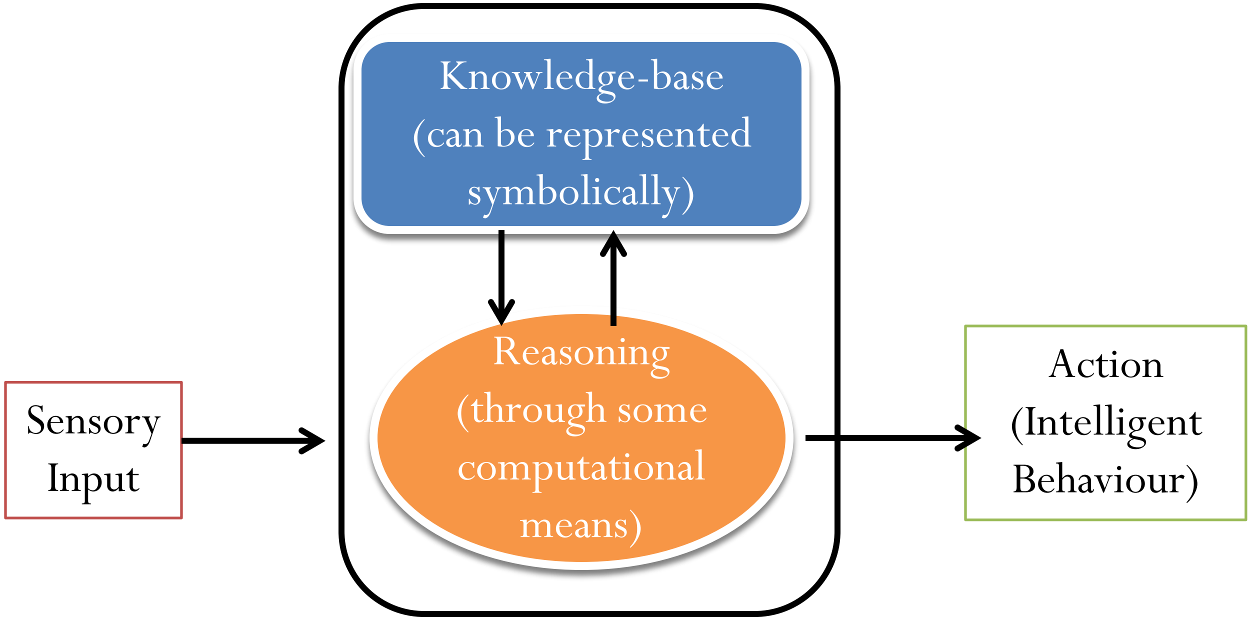
Indian Institute of Technology Guwahati

Guwahati - 781039, INDIA

Outline

- Background
  - Knowledge, Representation, Reasoning
  - Uncertainty
    - What is uncertainty?
    - Reasons of Uncertainty
- Reasoning under uncertainty: How probability theory can be used?
- Belief Networks
  - Structure
  - Notion of d-separation

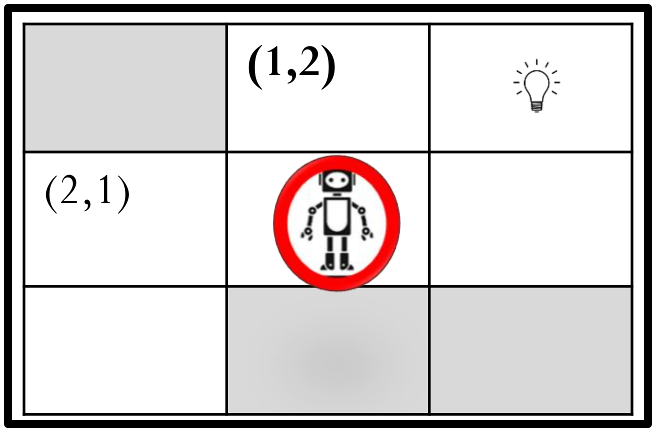
Background

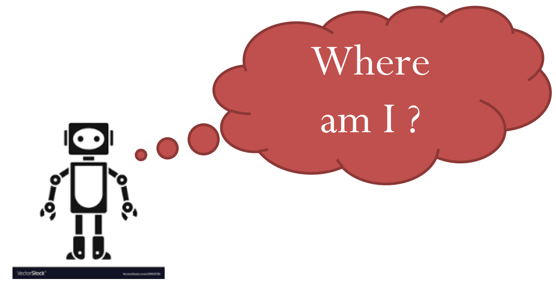


Knowledge-based Agent

- How an agent uses what it knows in deciding what to do?

Background





- Propositional Logic
  - $S_1 \colon \neg O_{1,2}, S_2 \colon O_{3,2}$
  - $S_3 \colon B_{2,2} \Leftrightarrow \mathrm{L}_{1,2} \vee L_{2,1}$ ….

- Robot Localization: robot needs to determine its current location
  - given a map of the world
  - four sonar sensors (NSWE) and one light sensor (L)
    - tells whether there is an obstacle (the outer wall or gray square in the figure), and also if a room is bright.
    - Current sensor value : [N S W E L] : [False True False False False]

Background

- Uncertainty
  - the state of being unsure of something (from dictionary)
- Uncertainty in data (facts)
  - Imprecise, inaccurate and unreliable data
  - Missing data
  - Example: Medical domain
    - Patient's weight is 45 kg vs. 45.25 kg
    - Patient's weight is 43.444 kg vs. 45 kg (former is more precise not accurate if a person's actual weight is 44.9 kg)
    - Patient's weight is 45.25 kg, measured again – it is 44.95 kg.
    - Medication requires two tests, however results of only one test is available

Background

- Uncertainty in knowledge (rules)
  - Vagueness in rules
    - Example: If the person is overweight then they usually have large waistline.
  - Not enough rules to cover the problem space (theoretical or practical ignorance, lack of available theory to describe a situation)
  - Rules may be contradictory (different evidences suggesting same diagnosis)
- Issues:
  - How to represent uncertain data and knowledge?
  - How to draw inference using uncertain data and knowledge?

Background

- Probability theory- deals with incompleteness (ignorance about the world)

- Probability provides a way of summarizing the uncertainty that comes from ignorance, quantifies the *degree of belief*

- Probability **:** a measure of belief (as opposed to being a frequency) – Bayesian Probability or Subjective Probability

Bayesian networks

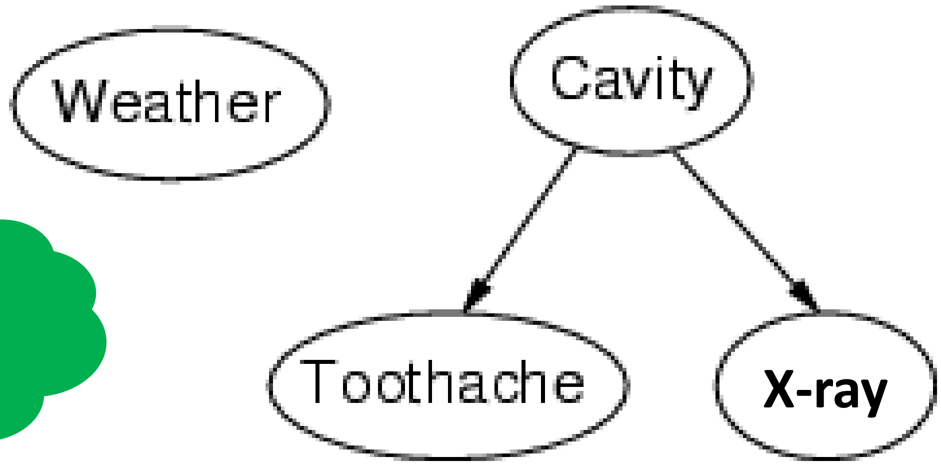
- Representing knowledge in uncertain domain
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.
- Syntax:
  - a set of nodes, one node per random variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}\,(X_i \mid \text{Parents}\,(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:



Causes should be parent of *Effect*

- *Weather* is independent of the other variables
- *Toothache* and *X-raySpot* are conditionally independent given *Cavity*

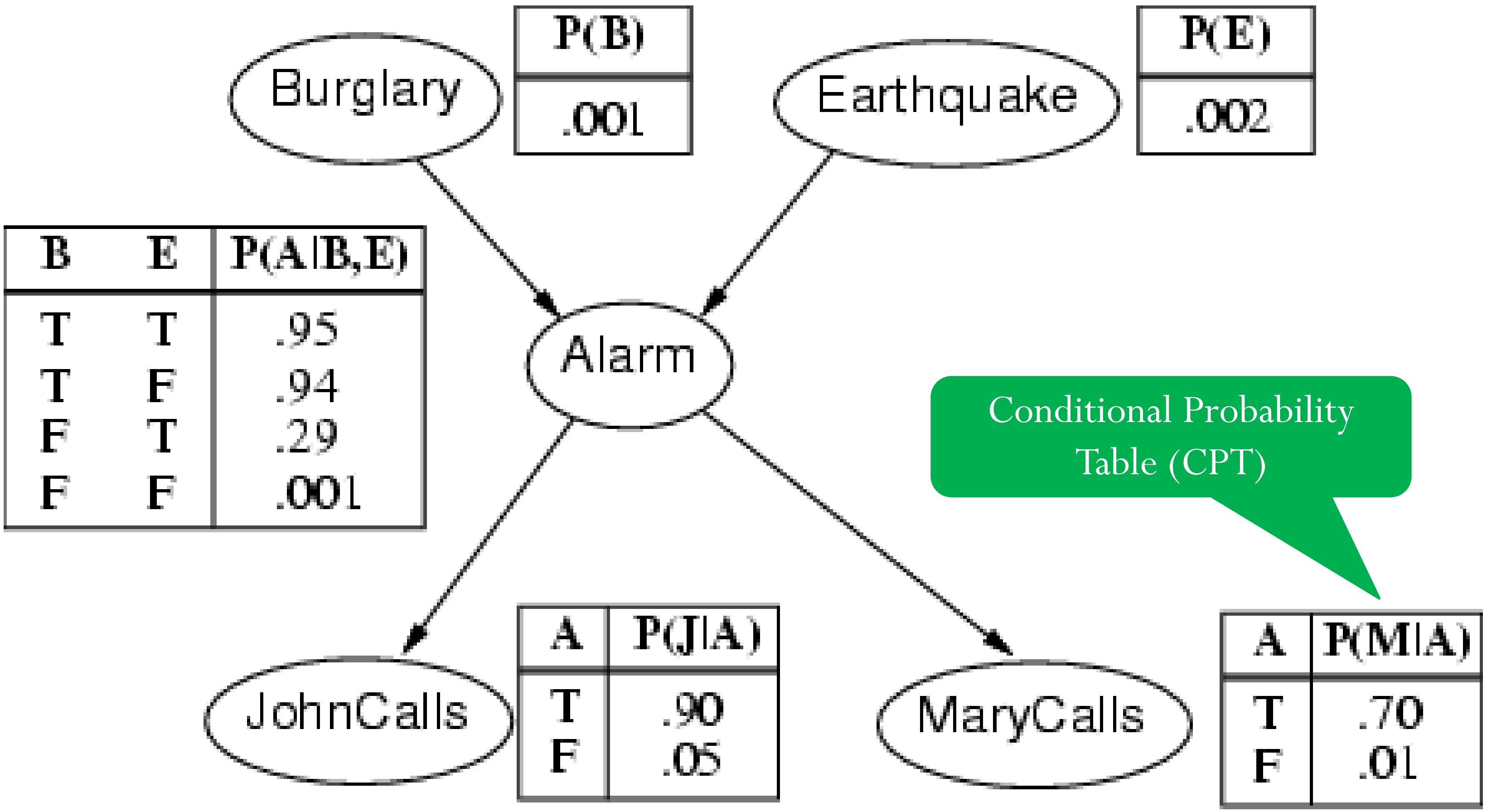
Example

- 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: ***Burglary***, ***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 contd.



P(B)
.001

P(E)
.002

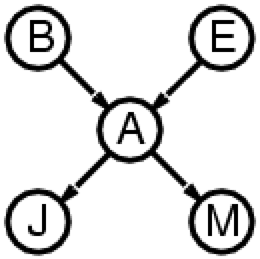
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)

- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

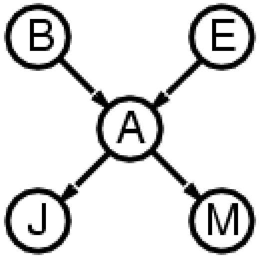
- i.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions that are associated with the nodes of the network:

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



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

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

$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$

≈ 0.00063

Constructing Bayesian networks

- The joint distribution $P(X_1=x_1, \ldots, X_n = x_n)$ can be given in terms of conditional probability using product rule:

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, ..x_1)\, P(x_{n-1}, \ldots, x_1)$$

repeating the process, reducing each conjunctive probability to a conditional probability and a smaller conjunction

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

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

Chain Rule

This specification of joint distribution is equivalent to

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

provided $Parents(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$

Take care of the node ordering while constructing the network.

Constructing Bayesian networks

- Determine the set of variables, choose an ordering of variables $X_1, \ldots, X_n$ (if causes precede effects, this will result in compact network)
- For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

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

this choice of parents guarantees:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \text{ (chain rule)}$$

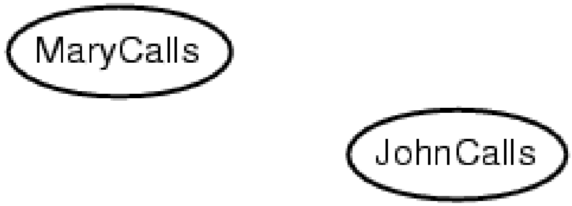
$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \text{ (by construction)}$$

  - for each parent insert a link from parent to X_i
  - CPTs: write down the conditional probability table, ${}_1\mathbf{P}(X_i \mid Parents(X_i))$

Example

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

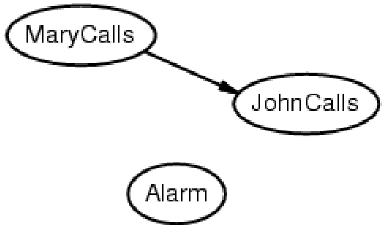
Wrong ordering??



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

Example

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



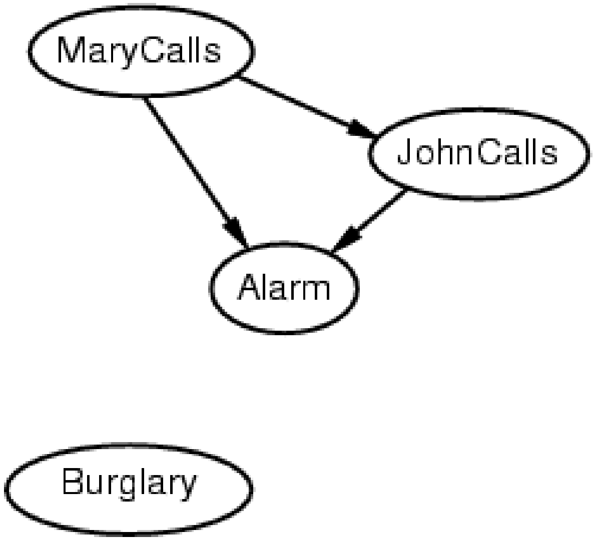
$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

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

Example

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



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

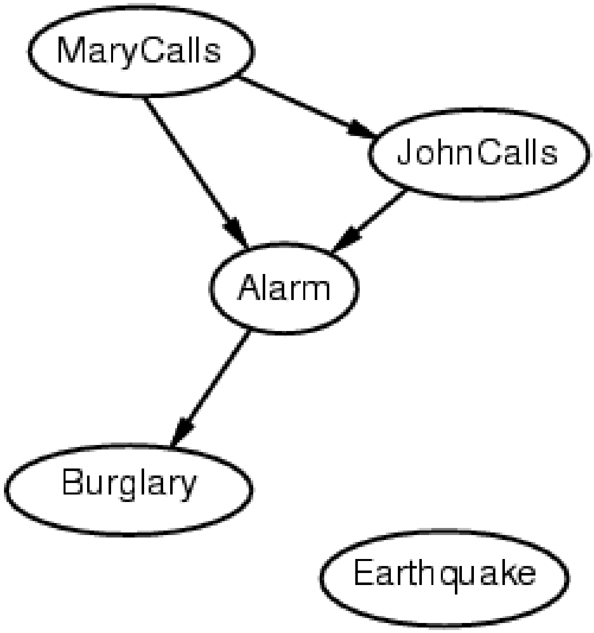
$\mathbf{P}(A \mid J, M) = \mathbf{P}(A)?$ **No**

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

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

Example

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



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

No

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

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

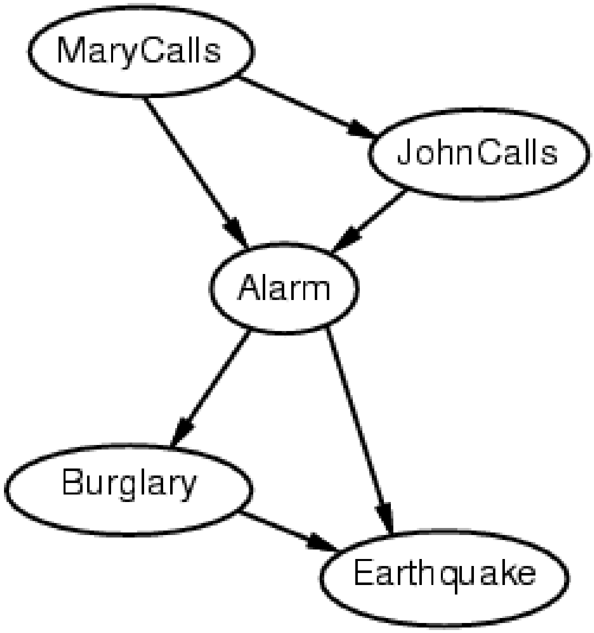
$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B)$? **No**

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

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

Example

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



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

No

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

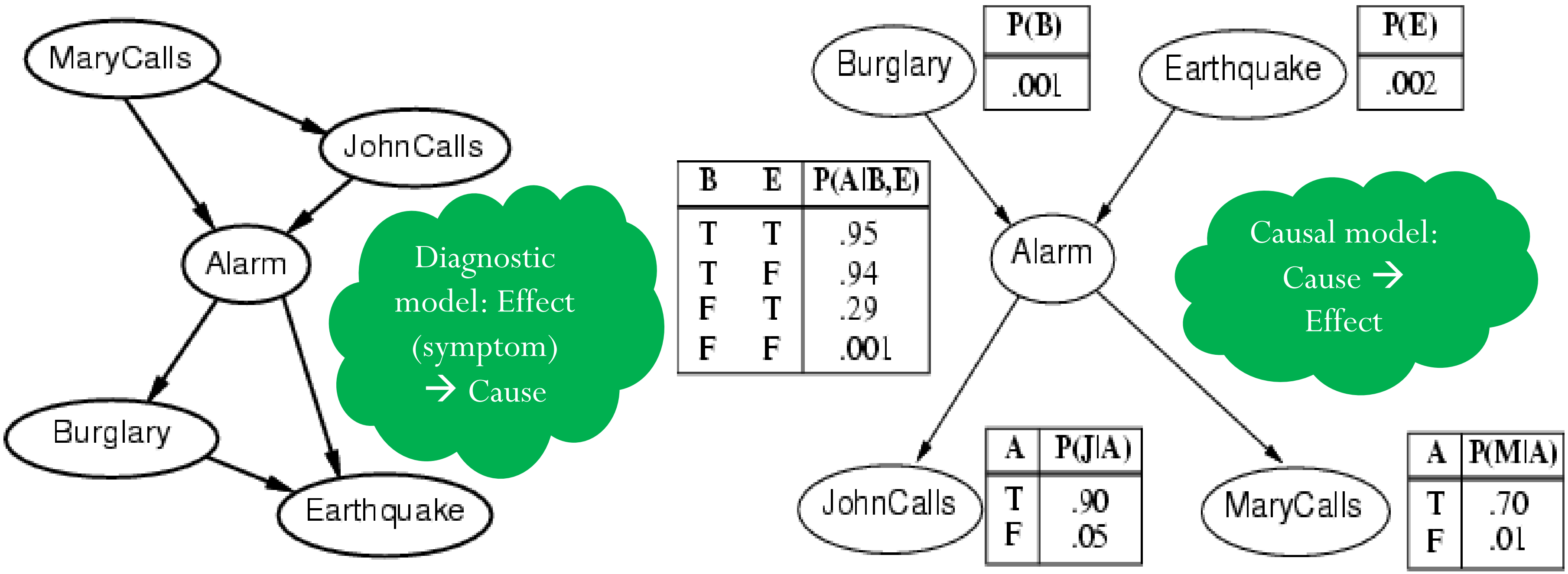
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**

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

Example contd.

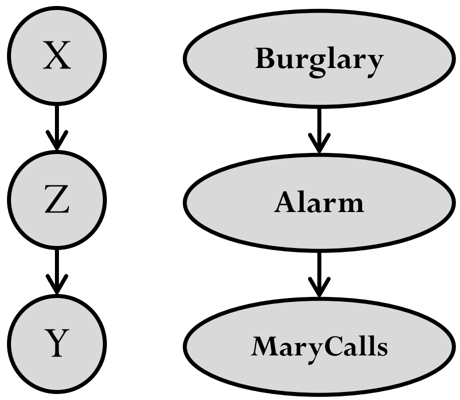


- Resulting network has two more links, requires three more probabilities to be specified: Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed
- Deciding conditional independence is hard in noncausal directions

Conditional Independence

- Can we find all the independences of a BN by inspecting its structure (from the graph)?
- Let us first see a three-node network where variables X and Y are connected via third variable Z in four different ways and we will try to understand when an observation regarding a variable X can possibly change our beliefs about Y, in the presence of evidence variable Z.

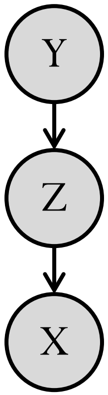
  - Forward serial connection (Causal trail - active iff Z is not observed)



- When Z is not instantiated (its truth value is not known variable is not observed) X can influence Y via Z (having observed X will tell something about Y).
- When Z is instantiated then X cannot influence Y (if we observe Z then knowing about X will not tell anything new about Y).

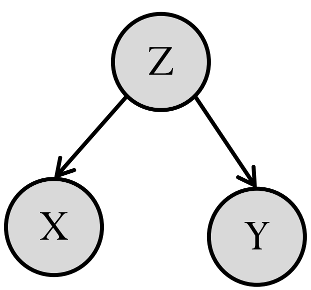
Conditional Independence

- Backward serial connection (evidential trail- active iff Z is not observed)

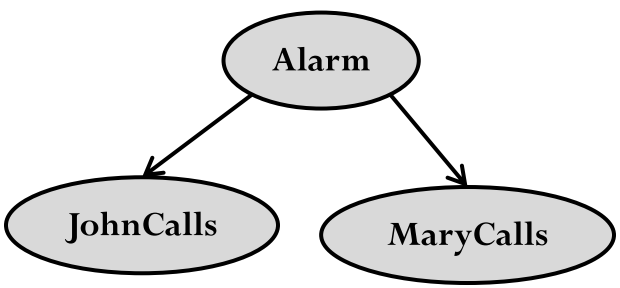


- When Z is not instantiated Y can influence X via Z (knowing about Y will tell something about X).
- When Z is instantiated then Y cannot influence X (if we observe Z then knowing about Y will not tell anything new about X).

- Diverging connection (Common cause- active iff Z is not observed)

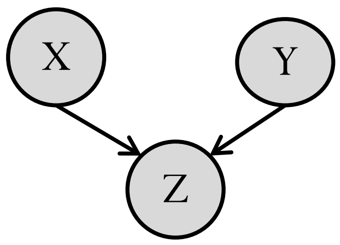


- Similar to previous two cases: X can influence Y via Z if and only if Z is not observed.
- In other words, if we know Z (or observe Z), then knowing about X will not give us any additional information about

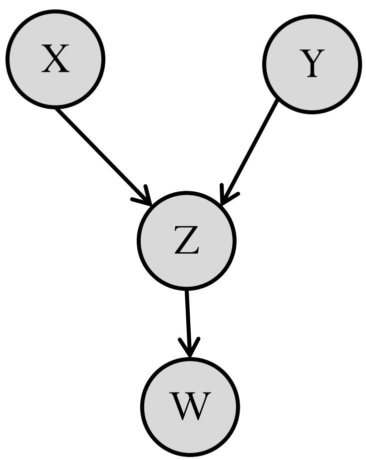


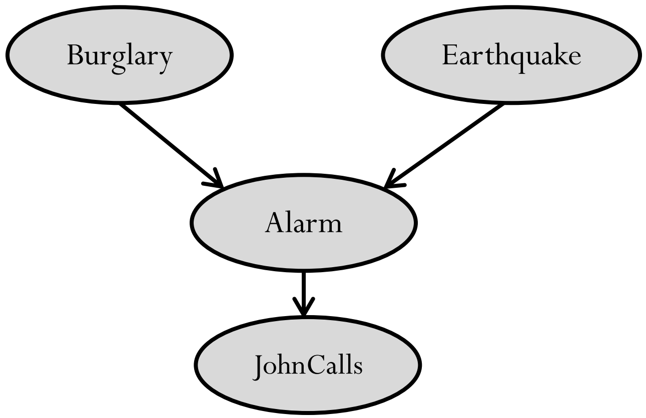
Conditional Independence

- Converging connection (Common effect- active iff either Z or one of Z's descendants is observed)



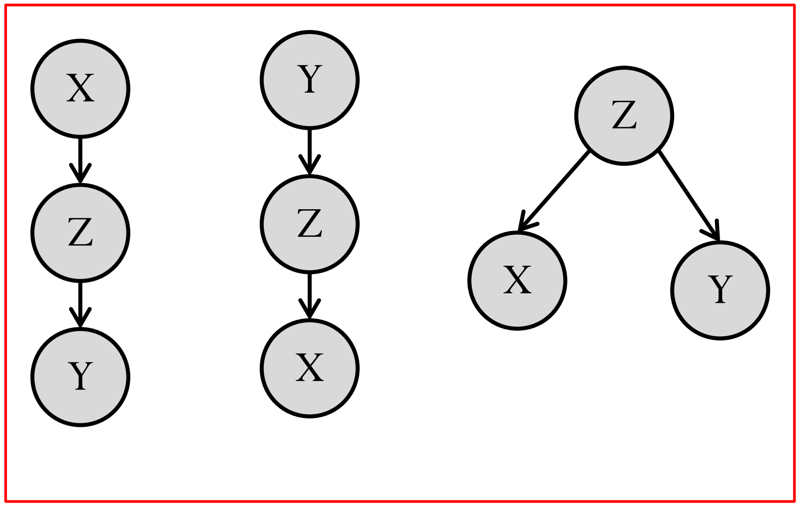
- X can influence Y only if Z or descendant of Z is instantiated.
- Without observing Z, knowing X does not tell anything about Y.
- When either node Z is instantiated, or one of its descendants is, then we know something about whether Z, and in that case information does propagate through from X to Y.

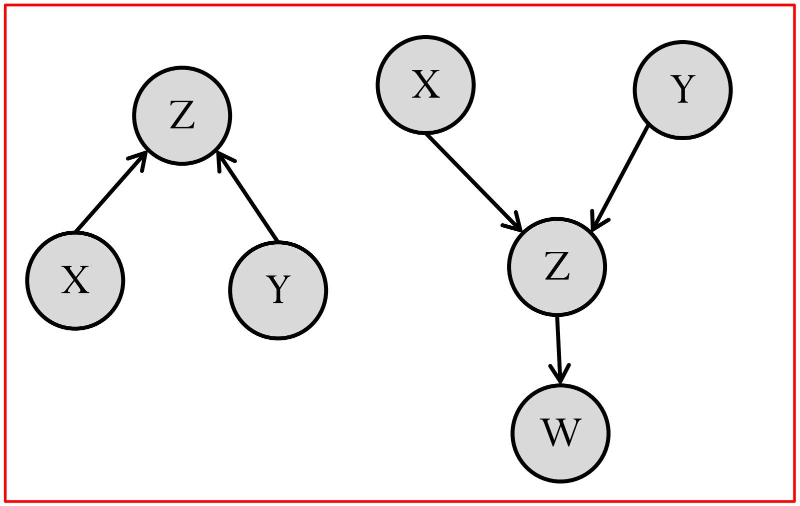




Conditional Independence

- Serial connections and diverging connections are essentially the same.





- **General case:** Considering longer trail $X_1 \rightleftharpoons \ldots \rightleftharpoons X_n$, for influence to "flow" from X_1 to X_n, it needs to flow through every single node on the trail.
- When multiple trails are there between two nodes then one node can influence another if there is any trail along which influence can flow.

Conditional Independence

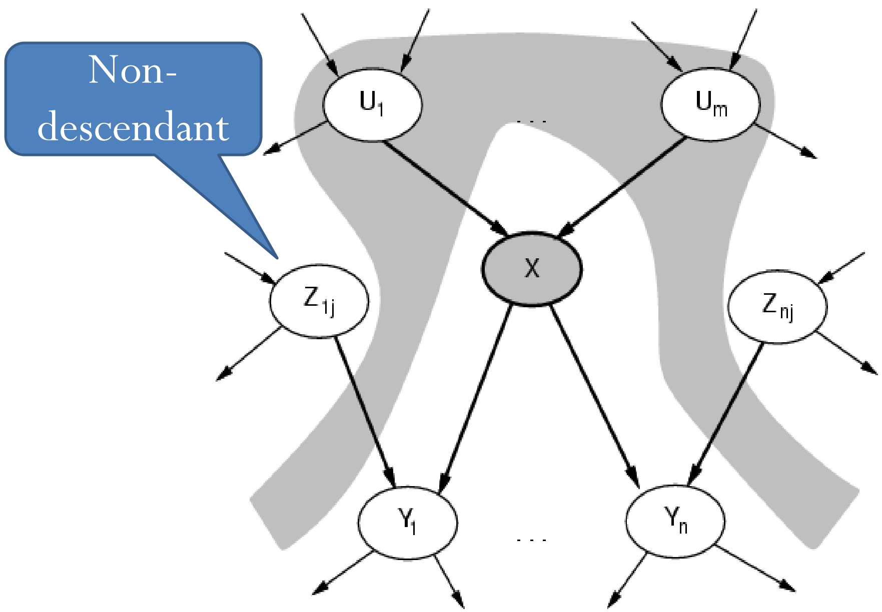
- d-separation (directed separation): provides a notion of separation between nodes in a directed graph.
- Variables X and Y are d-separated iff for every trail between them, there is an intermediate variable Z such that either
  - Z is in a serial or diverging connection and Z is known (observed).
  - Z is in converging connection and neither Z not any of Z's descendants are known.
- Two variable X and Y are d-connected if they are not d-separated.
- If variables X and Y are d-separated by Z then, X and Y are conditionally independent given Z.

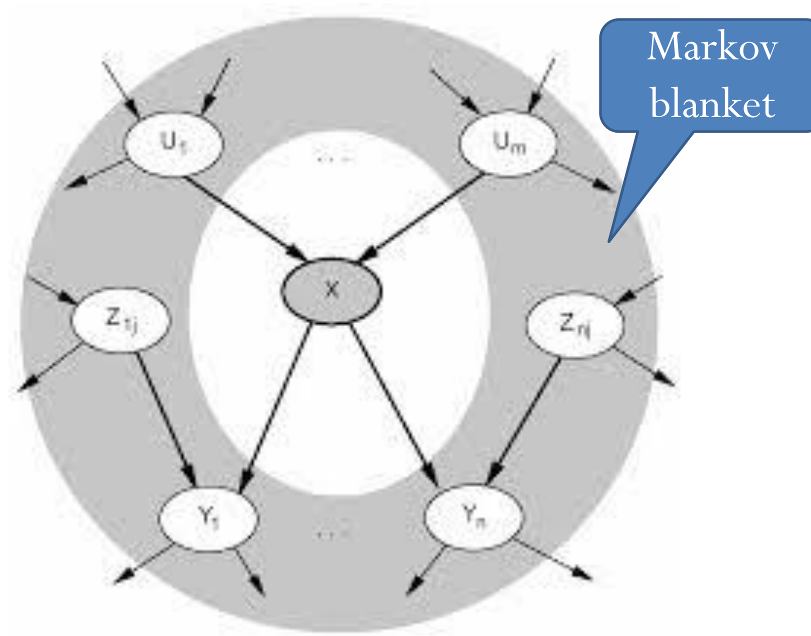
- **Definition:** Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be three sets of nodes in G (BN structure). We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated given $\boldsymbol{Z}$, denoted $dsep(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, if there is no active trail between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$.
- Let $I(G)$ denote the set of independencies that correspond to d-separation:

$$I(G) = \{\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}): dsep(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\}$$

This set is also called the set of global **Markov independencies**.

Conditional Independence relations





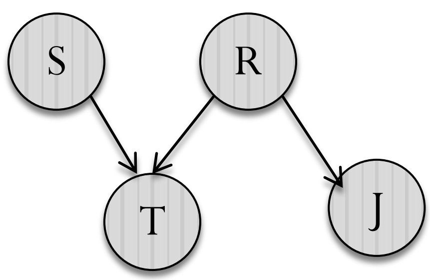
- Each node is conditionally independent of its non-descendants, given its parents.
- Each node is conditionally independent of all others given its Markov blanket: parents+children+children's parents

Conditional Independence

- **Example:** One morning Tracey leaves her house and realise that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night?
- Next she notices that the grass of her neighbour, Jack, is also wet.
- This **explains away** to some extent the possibility that her sprinkler was left on, and she concludes therefore that it is probably been raining (it decreases her belief that the sprinkler is on).
- Using the following four propositional random variables, construct the BN and determine if S is d-separated from J when T is known.
  - R: Rain $\in \{0,1\}$ (Rain = 1 means that it has been raining, and 0 otherwise)
  - S: Sprinkler $\in \{0,1\}$
  - J: Jack's grass wet $\in \{0,1\}$
  - T: Tracy's Grass wet

Conditional Independence

- Four propositional random variables are:
  - R: Rain ∈ {0,1} (Rain = 1 means that it has been raining, and 0 otherwise)
  - S: Sprinkler ∈ {0,1}
  - J: Jack's grass wet ∈ {0,1}
  - T: Tracy's Grass wet



- The trail between S and J: S-T-R-J
- S-T-R converging connection and T is known so influence flows from S to R.
- T-R-J diverging connection and R is not known so influence flows from T to J.
- So, S and J are not d-separated given T

What did we discuss in L3?

- What is knowledge, representation, and reasoning?
- What is uncertainty and reasoning under uncertainty?
- Under what situations does logic fail and how Probability theory can be useful in such situations?
- What kind of uncertainty is handled by Bayesian Probability?
- How Bayesian Networks can be used to represent knowledge under uncertainty?
- How to construct a Bayesian network?
- How to identify conditional independence relations from the structure of BNs?