CS 3710 Advanced Topics in AI

Lecture 2

Probabilistic graphical models

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square, x4-8845

http://www.cs.pitt.edu/~milos/courses/cs3710/

Motivation. Medical example.

We want to build a system for the **diagnosis of pneumonia**.

Problem description:

- **Disease:** pneumonia
- **Patient symptoms (findings, lab tests)**:
  - Fever, Cough, Paleness, WBC (white blood cells) count, Chest pain, etc.

Representation of a patient case:

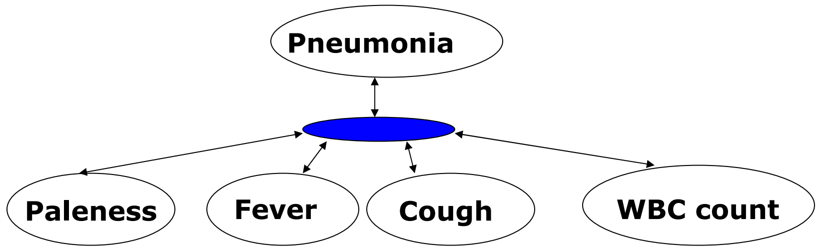
- Statements that hold (are true) for the patient.

  E.g: Fever =*True*
  Cough =*False*
  WBCcount=*High*

Diagnostic task: we want to decide whether the patient suffers from the pneumonia or not given the symptoms

Uncertainty

To make diagnostic inference possible we need to represent knowledge (axioms) that relate symptoms and diagnosis



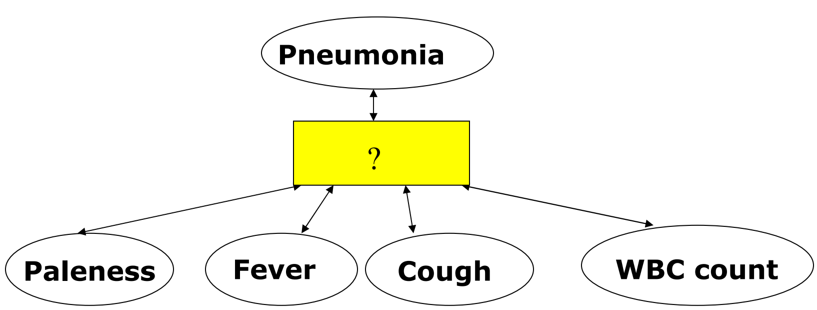
Problem: disease/symptoms relations are not deterministic

- **They are uncertain (or stochastic) and** vary from patient to patient

Modeling the uncertainty.

Key challenges:

- How to represent uncertain relations?
- How to manipulate such knowledge to make inferences?
  - **Humans can reason with uncertainty.**



Modeling uncertainty with probabilities

- **Random variables:**
  - **Binary** $Pneumonia$ is either $True, False$
    Random variable **Values**
  - **Multi-valued** $Pain$ is one of $\{Nopain, Mild, Moderate, Severe\}$
    Random variable **Values**
  - **Continuous** $HeartRate$ is a value in $< 0 ; 250 >$
    Random variable **Values**
- **A multivariate random variable or random vector** is a vector whose components are individual random variables
- **A patient state: an assignment of values to random variables. A value of a multivariate random var.**

  E.g. *Pneumonia =T , Fever =T, Paleness=F, WBCcount=medium, Cough=False*

Probabilities

Quantifies how likely is the outcome of a random variable

- **Unconditional probabilities (prior probabilities)**

$$P(Pneumonia = True) = 0.001$$

$$P(Pneumonia = False) = 0.999$$

$$P(WBCcount = high) = 0.005$$

Probability distribution

- Defines probabilities **for all possible value assignments to a random variable**
- Values are mutually exclusive

$Pneumonia$	$\mathbf{P}(Pneumonia)$
$True$	0.001
$False$	0.999

Probability distribution

Defines probability for **all possible value assignments**

Example 1:

$P(Pneumonia = True) = 0.001$
$P(Pneumonia = False) = 0.999$

$Pneumonia$	$\mathbf{P}(Pneumonia)$
$True$	0.001
$False$	0.999

$$P(Pneumonia = True) + P(Pneumonia = False) = 1$$

Probabilities sum to 1 !!!

Example 2:

$P(WBCcount = high) = 0.005$
$P(WBCcount = normal) = 0.993$
$P(WBCcount = high) = 0.002$

$WBCcount$	$\mathbf{P}(WBCcount)$
$high$	0.005
$normal$	0.993
low	0.002

Joint probability distribution

Joint probability distribution (for a set variables)

- Defines probabilities for **all possible assignments of values to variables in the set**

Example: variables *Pneumonia* and *WBCcount*

$\mathbf{P}(pneumonia, WBCcount)$

Is represented by 2×3 matrix

		$WBCcount$		
		$high$	$normal$	low
$Pneumonia$	$True$	0.0008	0.0001	0.0001
	$False$	0.0042	0.9929	0.0019

Joint probabilities

Marginalization

- reduces the dimension of the joint distribution
- Sums variables out

$\mathbf{P}(pneumonia, WBCcount)$ 2×3 matrix

$\mathbf{P}(Pneumonia)$

		WBCcount			
		high	*normal*	*low*	$\mathbf{P}(Pneumonia)$
Pneumonia	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	0.002	

Marginalization (here summing of columns or rows)

Full joint distribution

- **the joint distribution for all variables in the problem**
  - It defines the complete probability model for the problem

Example: pneumonia diagnosis

Variables: *Pneumonia, Fever, Paleness, WBCcount, Cough*

Full joint defines the probability for all possible assignments of values to *Pneumonia, Fever, Paleness, WBCcount, Cough*

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = T)$$
$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = F)$$
$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = F, Paleness = T)$$

… etc

Conditional probabilities

Conditional probability distribution

- Defines probabilities of outcomes of a variable, given a fixed assignment to some other variable values

  $P(Pneumonia = true \mid WBCcount = high)$

$\mathbf{P}(Pneumonia \mid WBCcount)$ 3 element vector of 2 elements

		WBCcount		
		high	*normal*	*low*
Pneumonia	*True*	0.08	0.0001	0.0001
	False	0.92	0.9999	0.9999
		1.0	1.0	1.0

$P(Pneumonia = true \mid WBCcount = high)$
$+ P(Pneumonia = false \mid WBCcount = high)$

Conditional probabilities

Conditional probability

- Is defined in terms of the joint probability:

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \text{ s.t. } P(B) \neq 0$$

- **Example:**

$$P(pneumonia = true \mid WBCcount = high) = \frac{P(pneumonia = true, WBCcount = high)}{P(WBCcount = high)}$$

$$P(pneumonia = false \mid WBCcount = high) = \frac{P(pneumonia = false, WBCcount = high)}{P(WBCcount = high)}$$

Conditional probabilities

- **Conditional probability distribution.**

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \quad \text{s.t. } P(B) \neq 0$$

- **Product rule.** Joint probability can be expressed in terms of conditional probabilities

$$P(A,B) = P(A \mid B)P(B)$$

- **Chain rule**. Any joint probability can be expressed as a product of conditionals

$$P(X_1, X_2, \ldots X_n) = P(X_n \mid X_{1,} \ldots X_{n-1})P(X_{1,} \ldots X_{n-1})$$
$$= P(X_n \mid X_{1,} \ldots X_{n-1})P(X_{n-1} \mid X_{1,} \ldots X_{n-2})P(X_{1,} \ldots X_{n-2})$$
$$= \prod_{i=1}^{n} P(X_i \mid X_{1,} \ldots X_{i-1})$$

Bayes rule

Conditional probability.

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad \longrightarrow \qquad P(A,B) = P(B \mid A)P(A)$$

Bayes rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

When is it useful?

- When we are interested in computing the diagnostic query from the causal probability

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

- **Reason:** It is often easier to assess causal probability
  - E.g. Probability of pneumonia causing fever
    vs. probability of pneumonia given fever

Bayes rule

Assume a variable A with multiple values $a_1, a_2, \ldots a_k$

Bayes rule can be rewritten as:

$$P(A = a_j \mid B = b) = \frac{P(B = b \mid A = a_j) P(A = a_j)}{P(B = b)}$$

$$= \frac{P(B = b \mid A = a_j) P(A = a_j)}{\sum_{i=1}^{k} P(B = b \mid A = a_j) P(A = a_j)}$$

Used in practice when we want to compute:

$\mathbf{P}(A \mid B = b)$ for all values of $a_1, a_2, \ldots a_k$

Probabilistic inference

Various inference tasks:

- **Diagnostic task. (from effect to cause)**

  $$\mathbf{P}(Pneumonia \mid Fever = T)$$

- **Prediction task. (from cause to effect)**

  $$\mathbf{P}(Fever \mid Pneumonia = T)$$

- **Other probabilistic queries** (queries on joint distributions).

  $$\mathbf{P}(Fever)$$

  $$\mathbf{P}(Fever, ChestPain)$$

Inference

Any query can be computed from the full joint distribution !!!

- **Joint over a subset of variables** is obtained through marginalization

$$P(A=a, C=c) = \sum_i \sum_j P(A=a, B=b_i, C=c, D=d_j)$$

- **Conditional probability over set of variables**, given other variables' values is obtained through marginalization and definition of conditionals

$$P(D=d \mid A=a, C=c) = \frac{P(A=a, C=c, D=d)}{P(A=a, C=c)}$$

$$= \frac{\sum_i P(A=a, B=b_i, C=c, D=d)}{\sum_i \sum_j P(A=a, B=b_i, C=c, D=d_j)}$$

Inference.

Any query can be computed from the full joint distribution !!!

- Any joint probability can be expressed as a product of conditionals via the **chain rule**.

$$P(X_1, X_2, \ldots X_n) = P(X_n \mid X_{1,} \ldots X_{n-1}) P(X_{1,} \ldots X_{n-1})$$

$$= P(X_n \mid X_{1,} \ldots X_{n-1}) P(X_{n-1} \mid X_{1,} \ldots X_{n-2}) P(X_{1,} \ldots X_{n-2})$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{1,} \ldots X_{i-1})$$

- Sometimes it is easier to define the distribution in terms of conditional probabilities:
  - E.g. $\mathbf{P}(Fever \mid Pneumonia = T)$
    $\mathbf{P}(Fever \mid Pneumonia = F)$

Modeling uncertainty with probabilities

- Defining the **full joint distribution** makes it possible to represent and reason with uncertainty in a uniform way
- We are able to handle an arbitrary inference problem

Problems:

- **Space complexity.** To store a full joint distribution we need to remember $O(d^n)$ numbers.

  n – number of random variables, d – number of values
- **Inference (time) complexity.** To compute some queries requires $O(d^n)$ steps.
- **Acquisition problem.** Who is going to define all of the probability entries?

Medical diagnosis example.

- **Space complexity.**
  - Pneumonia (2 values: T,F), Fever (2: T,F), Cough (2: T,F), WBCcount (3: high, normal, low), paleness (2: T,F)
  - Number of assignments: 2*2*2*3*2=48
  - We need to define at least 47 probabilities.
- **Time complexity.**
  - Assume we need to compute the marginal of Pneumonia=T from the full joint

$$P(Pneumonia = T) =$$
$$= \sum_{i \in T,F} \sum_{j \in T,F} \sum_{k=h,n,l} \sum_{u \in T,F} P(Fever = i, Cough = j, WBCcount = k, Pale = u)$$

  - Sum over: 2*2*3*2=24 combinations

Graphical models

Aim: alleviate the representational and computational bottlenecks

Idea: Take advantage of the structure, in particular, **independences and conditional independences** that hold among random variables

Two classes of models:

- **Bayesian belief networks**
  - Modeling asymmetric (causal) effects and dependencies
- **Markov random fields**
  - Modeling symmetric effects and dependencies among random variables
  - Used often to model spatial dependences (image analysis)

Bayesian belief networks (BBNs)

Bayesian belief networks.

- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables

- **A and B are independent**

$$P(A,B) = P(A)P(B)$$

- **A and B are conditionally independent given C**

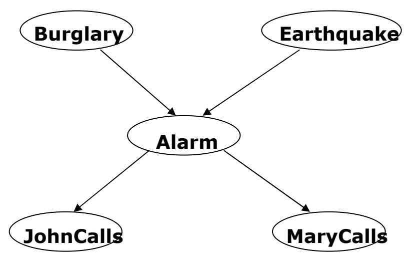
$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

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

Alarm system example.

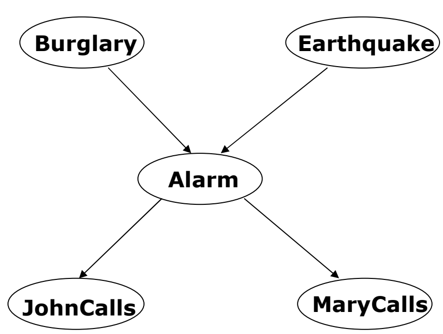
- Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events:
  - Burglary, Earthquake, Alarm, Mary calls and John calls

Causal relations



Bayesian belief network.

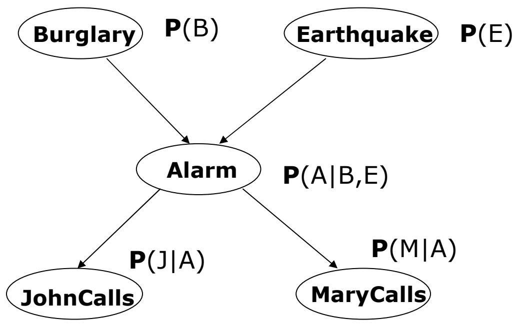
1. **Directed acyclic graph**
   - **Nodes** = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   - **Links** = direct (causal) dependencies between variables.
     The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm



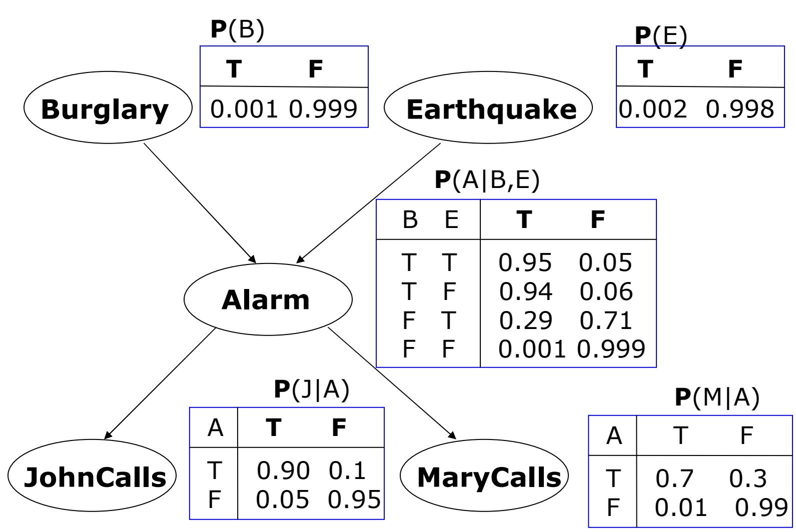
Bayesian belief network.

2. Local conditional distributions

- relate variables and their parents



Bayesian belief network.



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

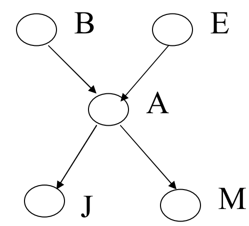
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Bayesian belief networks (general)

Two components: $B = (S, \Theta_S)$

- **Directed acyclic graph**
  - Nodes correspond to random variables
  - (Missing) links encode independences



- **Parameters**
  - Local conditional probability distributions for every variable-parent configuration

$$\mathbf{P}(X_i \mid pa(X_i))$$

Where:

$pa(X_i)$ - stand for parents of X_i

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

Full joint distribution in BBNs

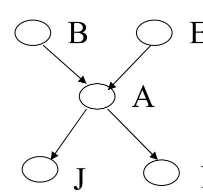
Full joint distribution is defined in terms of local conditional distributions (obtained via the chain rule):

$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

Example:

Assume the following assignment of values to random variables

$B = T, E = T, A = T, J = T, M = F$



Then its probability is:

$P(B = T, E = T, A = T, J = T, M = F) =$

$$P(B = T)P(E = T)P(A = T \mid B = T, E = T)P(J = T \mid A = T)P(M = F \mid A = T)$$

Bayesian belief networks (BBNs)

Bayesian belief networks

- Represent the full joint distribution over the variables more compactly using the product of local conditionals.
- **But how did we get to local parameterizations?**

Answer:

- **Graphical structure** encodes **conditional and marginal independences** among random variables
- **A and B are independent** $P(A,B) = P(A)P(B)$
- **A and B are conditionally independent given C**

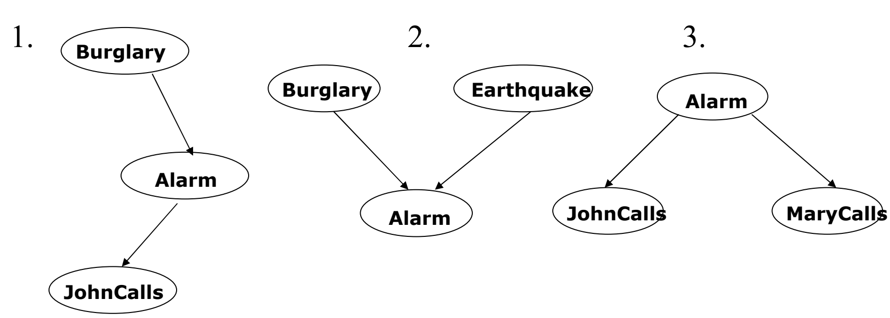
$$P(A \mid C,B) = P(A \mid C)$$

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

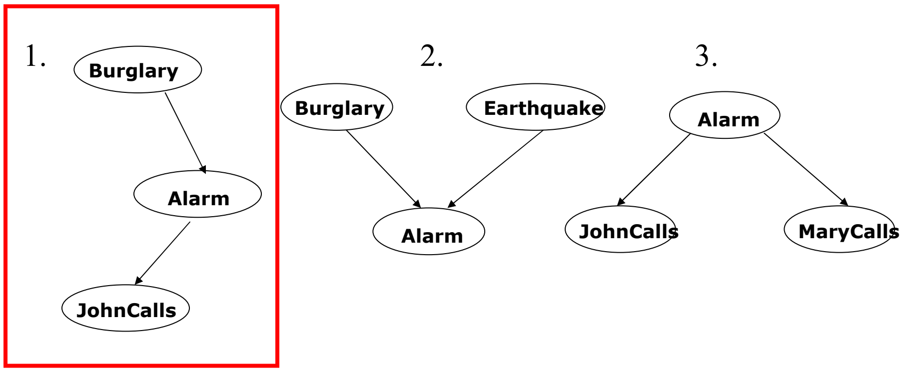
- **The graph structure implies the decomposition !!!**

Independences in BBNs

3 basic independence structures:



Independences in BBNs

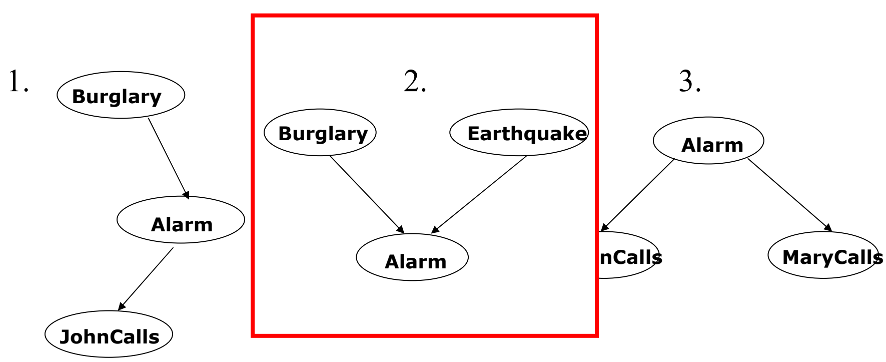


1. JohnCalls **is independent** of Burglary given Alarm

$$P(J \mid A, B) = P(J \mid A)$$

$$P(J, B \mid A) = P(J \mid A) P(B \mid A)$$

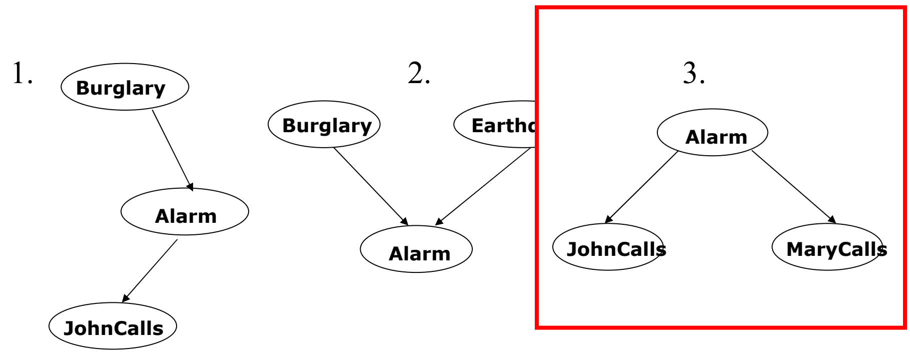
Independences in BBNs



2. Burglary **is independent** of Earthquake (not knowing Alarm)
   Burglary and Earthquake **become dependent** given Alarm !!

$$P(B, E) = P(B) P(E)$$

Independences in BBNs



3. MaryCalls **is independent** of JohnCalls given Alarm

$$P(J \mid A, M) = P(J \mid A)$$

$$P(J, M \mid A) = P(J \mid A) P(M \mid A)$$

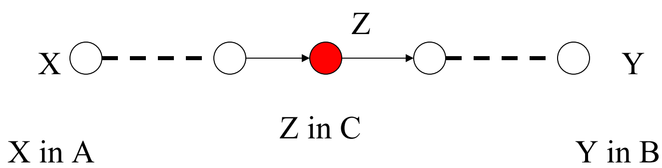
Independences in BBN

- BBN distribution models many conditional independence relations among distant variables and sets of variables
- These are defined in terms of the graphical criterion called d-separation
- **D-separation and independence**
  - Let X,Y and Z be three sets of nodes
  - If X and Y are d-separated by Z, then X and Y are conditionally independent given Z
- **D-separation :**
  - A is d-separated from B given C if every undirected path between them is **blocked with C**
- **Path blocking**
  - 3 cases that expand on three basic independence structures

Undirected path blocking

A is d-separated from B given C if every undirected path between them is **blocked**

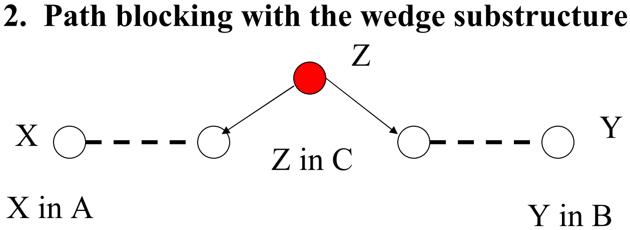
- **1. Path blocking with a linear substructure**



Undirected path blocking

A is d-separated from B given C if every undirected path between them is **blocked**

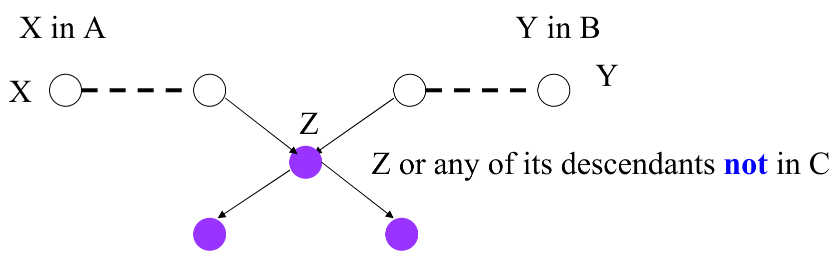
- **2. Path blocking with the wedge substructure**



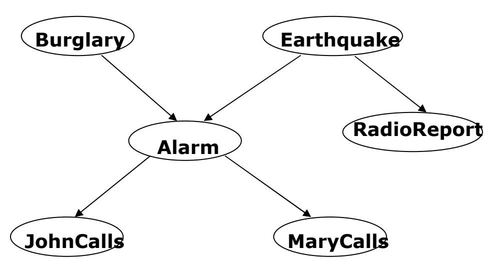
Undirected path blocking

A is d-separated from B given C if every undirected path between them is **blocked**

- **3. Path blocking with the vee substructure**

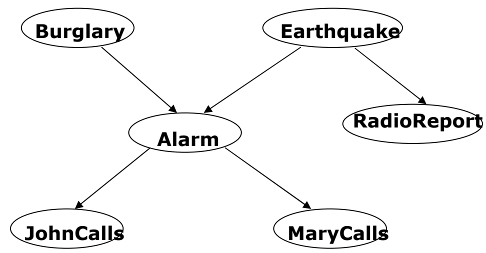


Independences in BBNs



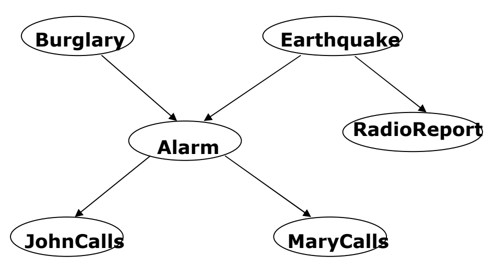
- Earthquake and Burglary are independent given MaryCalls **?**

Independences in BBNs



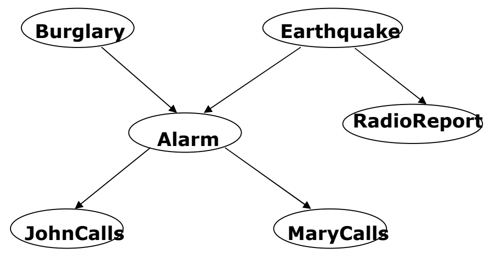
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **?**

Independences in BBNs



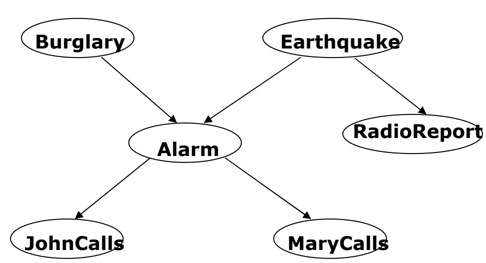
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **?**

Independences in BBNs



- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **T**
- Burglary and RadioReport are independent given MaryCalls **?**

Independences in BBNs



- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **T**
- Burglary and RadioReport are independent given MaryCalls **F**

Bayesian belief networks (BBNs)

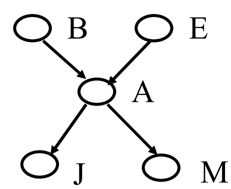
Bayesian belief networks

- Represents the full joint distribution over the variables more compactly using the product of local conditionals.
- **So how did we get to local parameterizations?**

$$\mathbf{P}(X_1, X_2,.., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

- **The decomposition is implied by the set of independences encoded in the belief network.**

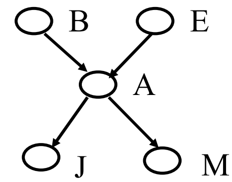
Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

$$P(B=T, E=T, A=T, J=T, M=F) =$$

Full joint distribution in BBNs

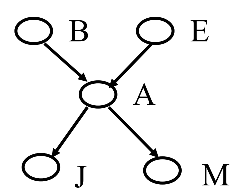


Rewrite the full joint probability using the product rule:

$$P(B=T,E=T,A=T,J=T,M=F)=$$

$$=P(J=T\,|\,B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$
$$=\underline{P(J=T\,|\,A=T)}\underline{P(B=T,E=T,A=T,M=F)}$$

Full joint distribution in BBNs

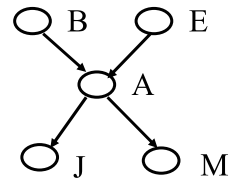


Rewrite the full joint probability using the product rule:

$$P(B=T,E=T,A=T,J=T,M=F)=$$

$$=P(J=T\,|\,B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$
$$=\underline{P(J=T\,|\,A=T)}\underline{P(B=T,E=T,A=T,M=F)}$$
$$P(M=F\,|\,B=T,E=T,A=T)P(B=T,E=T,A=T)$$
$$\underline{P(M=F\,|\,A=T)}\underline{P(B=T,E=T,A=T)}$$

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

$$P(B=T,E=T,A=T,J=T,M=F)=$$

$$=P(J=T\mid B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$

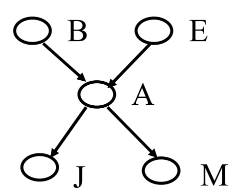
$$=\underline{P(J=T\mid A=T)}\underline{P(B=T,E=T,A=T,M=F)}$$

$$P(M=F\mid B=T,E=T,A=T)P(B=T,E=T,A=T)$$

$$\underline{P(M=F\mid A=T)}\underline{P(B=T,E=T,A=T)}$$

$$\underline{P(A=T\mid B=T,E=T)}\underline{P(B=T,E=T)}$$

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

$$P(B=T,E=T,A=T,J=T,M=F)=$$

$$=P(J=T\mid B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$

$$=\underline{P(J=T\mid A=T)}\underline{P(B=T,E=T,A=T,M=F)}$$

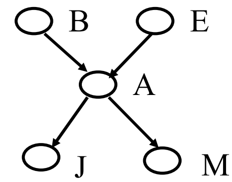
$$P(M=F\mid B=T,E=T,A=T)P(B=T,E=T,A=T)$$

$$\underline{P(M=F\mid A=T)}\underline{P(B=T,E=T,A=T)}$$

$$\underline{P(A=T\mid B=T,E=T)}\underline{P(B=T,E=T)}$$

$$P(B=T)P(E=T)$$

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

$$P(B=T,E=T,A=T,J=T,M=F)=$$

$$=P(J=T\mid B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$

$$=P(J=T\mid A=T)P(B=T,E=T,A=T,M=F)$$

$$P(M=F\mid B=T,E=T,A=T)P(B=T,E=T,A=T)$$

$$P(M=F\mid A=T)P(B=T,E=T,A=T)$$

$$P(A=T\mid B=T,E=T)P(B=T,E=T)$$

$$P(B=T)P(E=T)$$

$$=P(J=T\mid A=T)P(M=F\mid A=T)P(A=T\mid B=T,E=T)P(B=T)P(E=T)$$