CS 2740 Knowledge Representation
Lecture 17

# Bayesian belief networks

**Milos Hauskrecht**
milos@cs.pitt.edu
5329 Sennott Square



## 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

## Modeling uncertainty with probabilities

- **Knowledge based system era (70s – early 80's)**
  - **Extensional non-probabilistic models**
  - Solve the space, time and acquisition bottlenecks in probability-based models
  - froze the development and advancement of KB systems and contributed to the slow-down of AI in 80s in general
- Breakthrough (late 80s, beginning of 90s)
  - **Bayesian belief networks**
    - Give solutions to the space, acquisition bottlenecks
    - Partial solutions for time complexities
- Bayesian belief network



## 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**

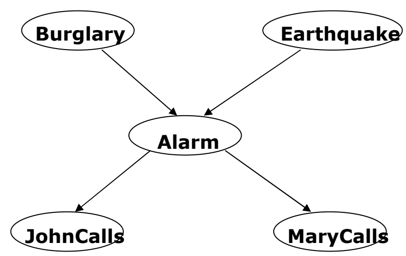
$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

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

# 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**

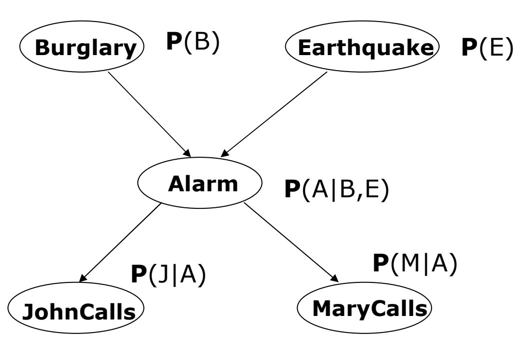





# Bayesian belief network.

1. **Directed acyclic graph**
   - **Nodes** = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   - **Links** = direct (causal) dependencies between variables.

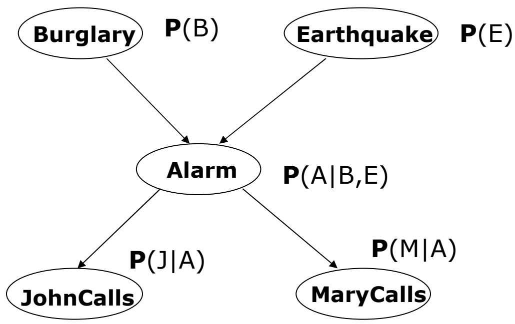
     The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm



# Bayesian belief network.

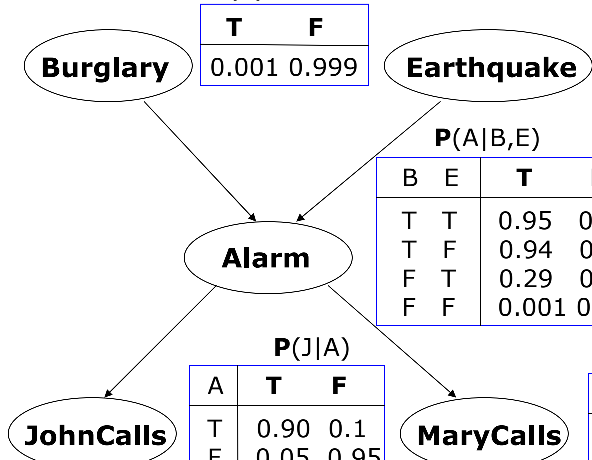
**2. Local conditional distributions**

- 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 |

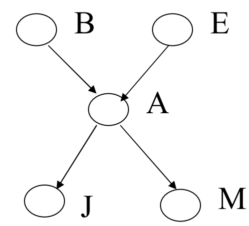
**P**(M|A)

| 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 |

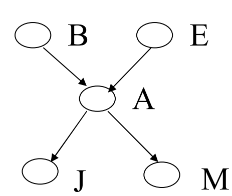


# Full joint distribution in BBNs

**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**

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

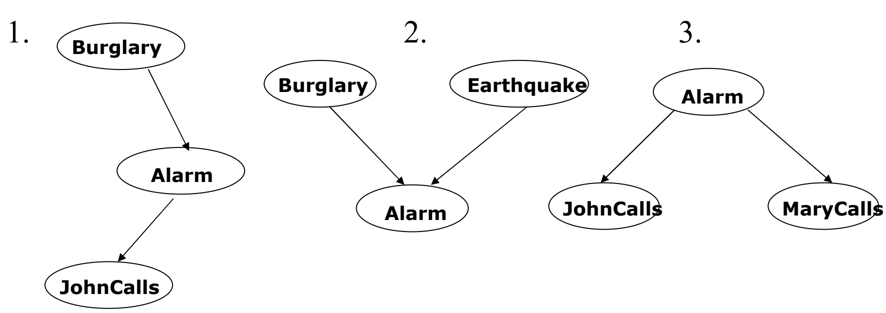
$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

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

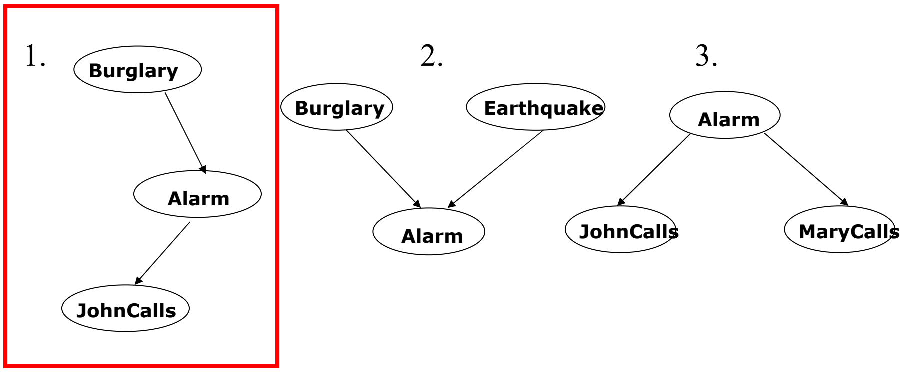


# Independences in BBNs

**3 basic independence structures:**



# Independences in BBNs



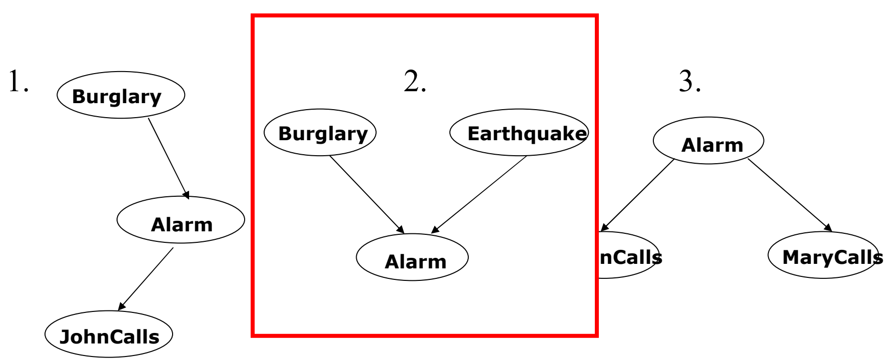

1. JohnCalls **is independent** of Burglary given Alarm

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

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



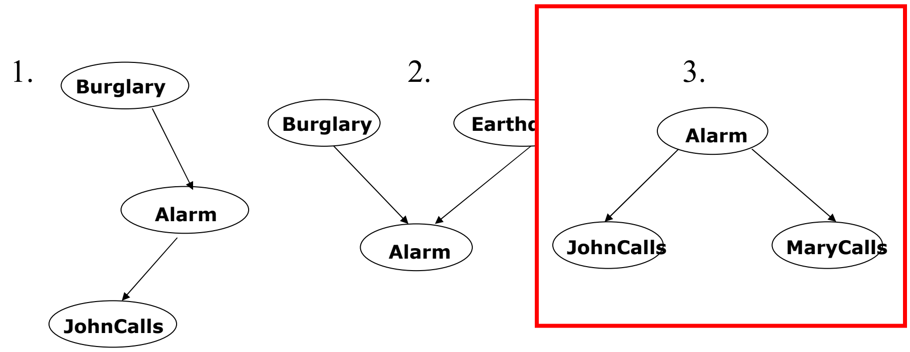
# Independences in BBNs




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

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

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



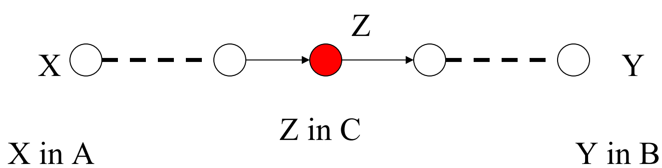
# Independences in BBN

- 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**

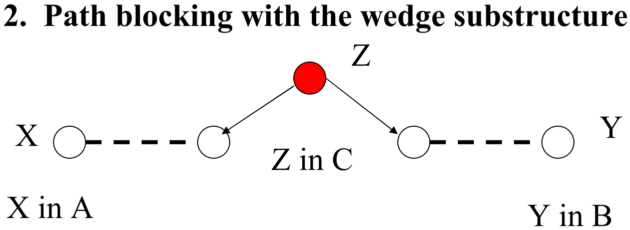





# Undirected path blocking

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

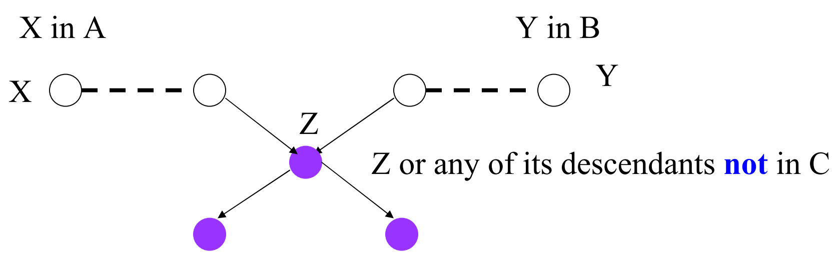
- **2. Path blocking with the wedge substructure**



# Undirected path blocking

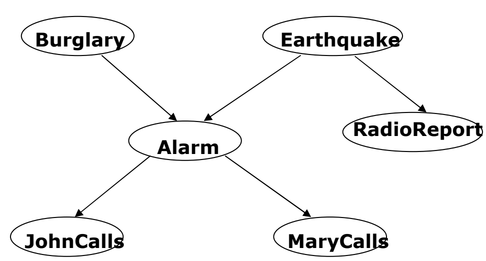
A is d-separated from B given C if every undirected path between them is **blocked**

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






# Independences in BBNs




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

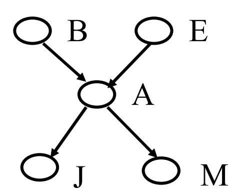
- Represents the full joint distribution over the variables more compactly using the product of local conditionals.
- **So how did we get to local parameterizations?**

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

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



# Full joint distribution in BBNs

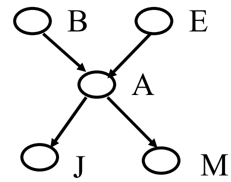



**Rewrite the full joint probability using the product rule:**

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

# Full joint distribution in BBNs

**Rewrite the full joint probability using the product rule:**



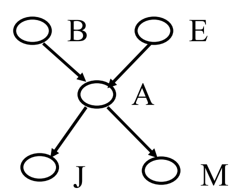

$$P(B=T,E=T,A=T,J=T,M=F)=$$

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



# Full joint distribution in BBNs

**Rewrite the full joint probability using the product rule:**

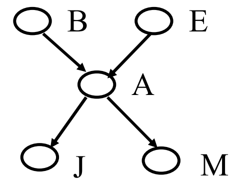



$$P(B=T,E=T,A=T,J=T,M=F)=$$

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

# Full joint distribution in BBNs

**Rewrite the full joint probability using the product rule:**




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

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

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

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

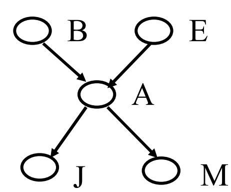
$$\underline{P(M=F \mid A=T)}\underline{P(B=T,E=T,A=T)}$$

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



# Full joint distribution in BBNs

**Rewrite the full joint probability using the product rule:**




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

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

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

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

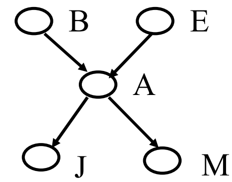
$$\underline{P(M=F \mid A=T)}\underline{P(B=T,E=T,A=T)}$$

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

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

CS 2740 Knowledge Representation M. Hauskrecht

# Full joint distribution in BBNs




**Rewrite the full joint probability using the product rule:**

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

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

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

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

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

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

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

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