CS 1675 Introduction to Machine Learning
Lecture 14

Bayesian belief networks

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

Midterm exam

Midterm Thursday, March 2, 2017

- **in-class (75 minutes)**
- **closed book**

- No programing questions
- Know/understand well what is in lecture notes and assignments
- Know key ideas/principle behind different models, learning algorithms
- Know principles for ML and Error function optimization
- Know how to differentiate simple functions
- True/false questions will require justification

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$
$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: try to estimate the underlying true probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D

true distribution $p(\mathbf{X})$ → n samples $D = \{D_1, D_2, .., D_n\}$ → estimate $\hat{p}(\mathbf{X})$

Standard (iid) assumptions: Samples

- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**

Modeling complex distributions

Question: How to model and learn complex multivariate distributions $\hat{p}(\mathbf{X})$ with a large number of variables?

Example: modeling of disease – symptoms relations

- **Disease:** pneumonia
- **Patient symptoms (findings, lab tests)**:
  - Fever, Cough, Paleness, WBC (white blood cells) count, Chest pain, etc.
- **Model of the full joint distribution**:
  P(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain)

One probability per assignment of values to variables:
P(Pneumonia =T, Fever =T, Cought=T, WBC=High, Chest pain=T)

- **How many probabilities are there?**

Full joint distribution

- **Any joint probability over a subset of variables can be obtained via marginalization from the full joint**

$$P(Pneumonia, WBCcount, Fever) = \sum_{c,p=\{T,F\}} P(Pneumonia, WBCcount, Fever, Cough = c, Paleness = p)$$

- **Question:** Is it possible to recover the full joint from the joint probabilities over a subset of variables?

Joint probabilities

- **Is it possible to recover the full joint from the joint probabilities over a subset of variables?**

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

		WBCcount high	normal	low	$\mathbf{P}(Pneumonia)$
Pneumonia	*True*	?	?	?	0.001
	False	?	?	?	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	0.002	

Joint probabilities and independence

- **Is it possible to recover the full joint from the joint probabilities over a subset of variables?**
- Only if the variables are independent !!!

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

		WBCcount			$\mathbf{P}(Pneumonia)$
		high	*normal*	*low*	
Pneumonia	*True*	?	?	?	0.001
	False	?	?	?	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	0.002	

Variable independence

- **The two events A, B are said to be independent if:**

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

- **The variables X, Y are said to be independent if their joint can be expressed as a product of marginals:**

  P(X, Y) = **P**(X)**P**(Y)

Conditional probability

Conditional probability :

- Probability of A given B $$P(A \mid B) = \frac{P(A,B)}{P(B)}$$
- Conditional probability is defined in terms of joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities

$$P(A,B) = P(A \mid B)P(B) \quad \textbf{(product rule)}$$

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

- Conditional probability – is useful for **various probabilistic inferences**

$$P(Pneumonia = True \mid Fever = True, WBCcount = high, Cough = True)$$

Conditional probabilities

Conditional probability

- Is defined in terms of the joint probability:

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

- **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

- Defines probabilities for all possible assignments of values to target variables, given a fixed assignment of other variable values

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

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

		Pneumonia		
		True	*False*	
$\boxed{WBCcount}$	*high*	0.08	0.92	1.0
	normal	0.0001	0.9999	1.0
	low	0.0001	0.9999	$\boxed{1.0}$

Variable we condition on

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

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

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

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

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

Why this may be important?

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

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

Probabilistic inference

Various probabilistic 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).

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

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

Modeling complex distributions

- Defining the **full joint distribution** makes it possible to represent and reason with the probabilities
- We are able to handle an arbitrary inference problem

Problems:

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

  n – number of random variables, d – number of values
- **Inference (time) complexity.** To compute some queries requires $O(\mathrm{d}^{\mathrm{n}})$ steps.
- **Acquisition problem.** How to acquire/learn all these probabilities?

Pneumonia 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

Bayesian belief networks (BBNs)

Bayesian belief networks (late 80s, beginning of 90s)

- Give solutions to the space, acquisition bottlenecks
- Partial solutions for time complexities

Key features:

- 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
- **X and Y are independent** $P(X,Y) = P(X)P(Y)$
- **X and Y are conditionally independent given Z**

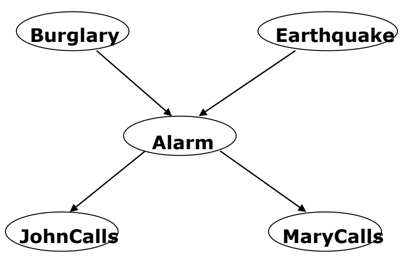
$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

$$P(X \mid Y,Z) = P(X \mid Z)$$

Alarm system example

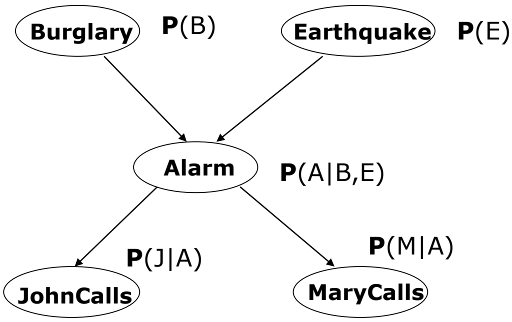
- Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is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

Causal relations



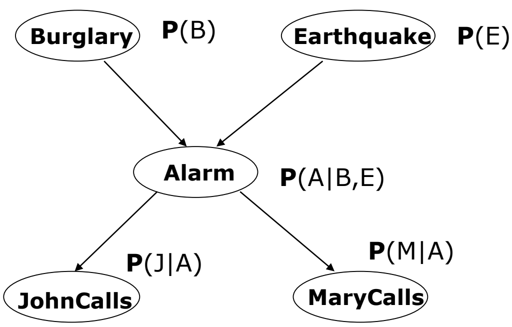
Bayesian belief network

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

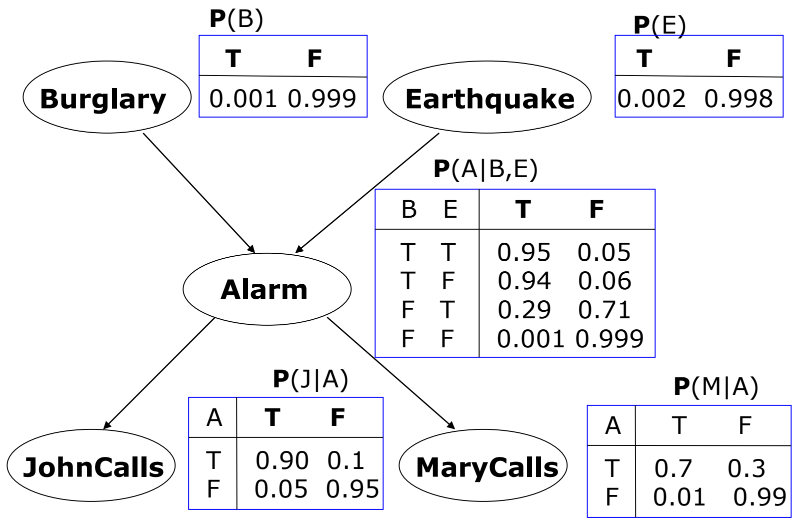


Bayesian belief network

2. **Local conditional distributions**
   - relating 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)

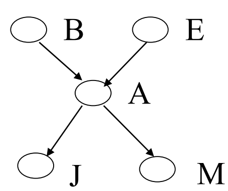
A	T	F
T	0.7	0.3
F	0.01	0.99

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**

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

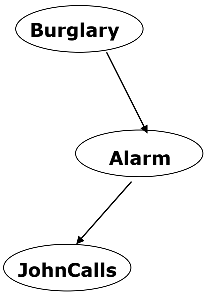
$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

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

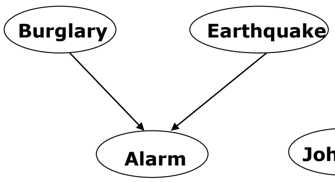
Independences in BBNs

3 basic independence structures:

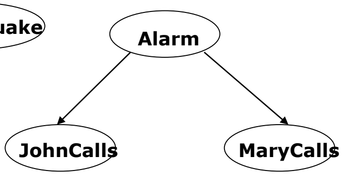
1.



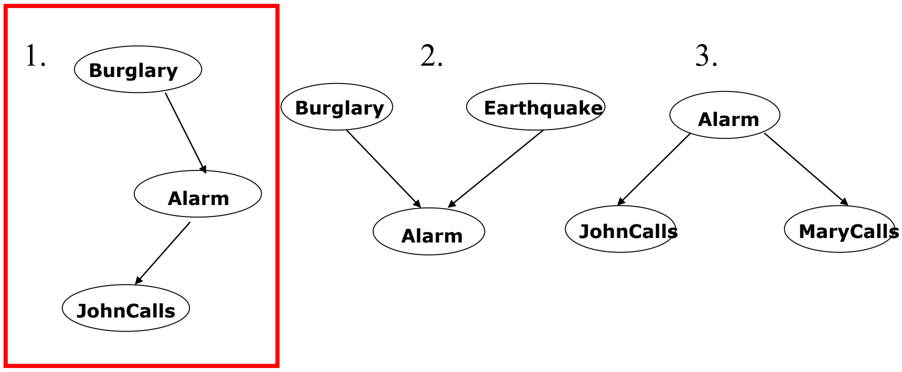
2.



3.



Independences in BBNs

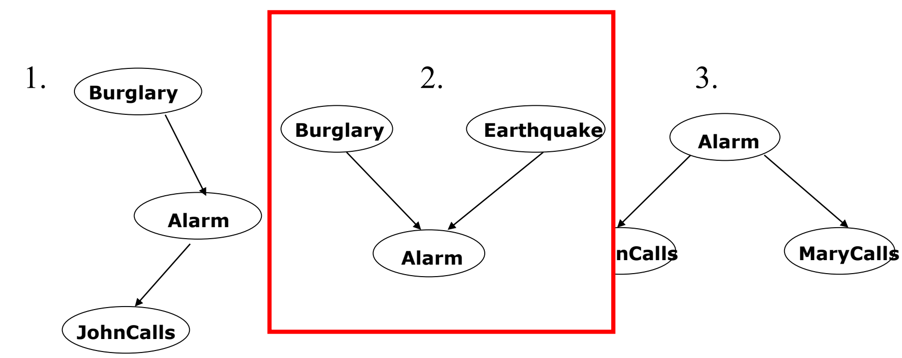


1. JohnCalls **is independent** of Burglary given Alarm

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

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

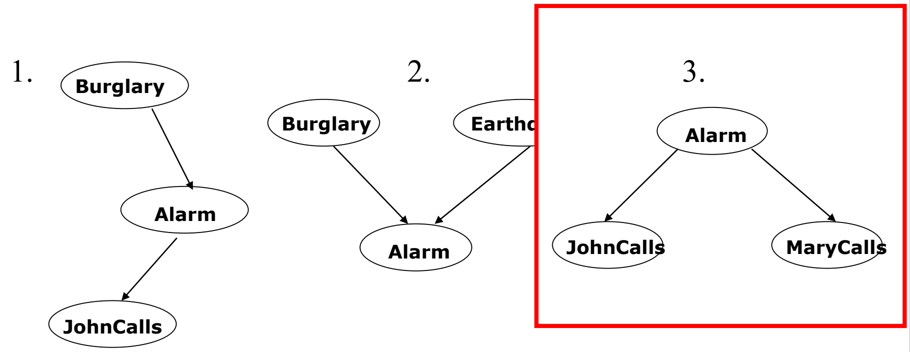
Independences in BBNs



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

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

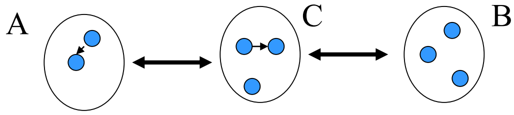
$$P(J, M \mid A) = P(J \mid A)P(M \mid A)$$

Independence in BBN

- BBN distribution models many conditional independence relations relating distant variables and sets
- These are defined in terms of the graphical criterion called d-separation
- **D-separation in the graph**
  - 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

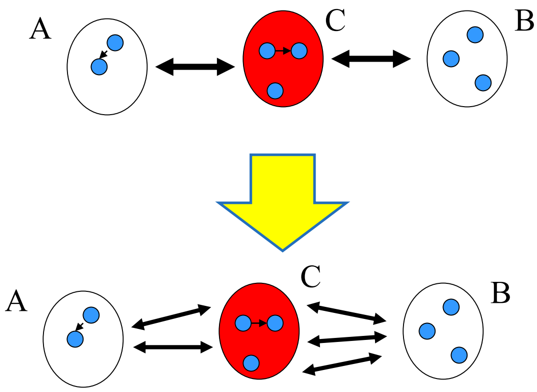
Undirected path blocking

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



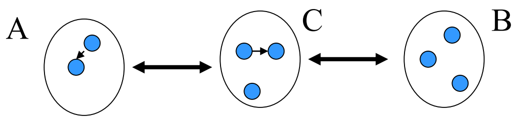
Undirected path blocking

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

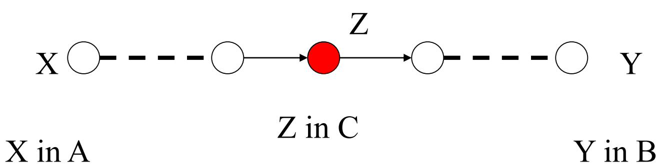


Undirected path blocking

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



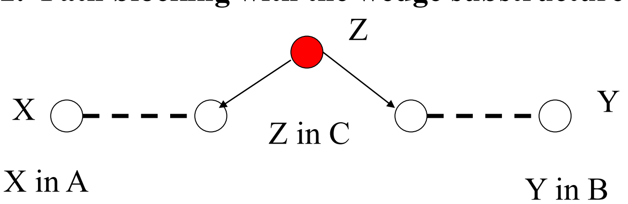
- **1. Path blocking with a linear substructure**



Undirected path blocking

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

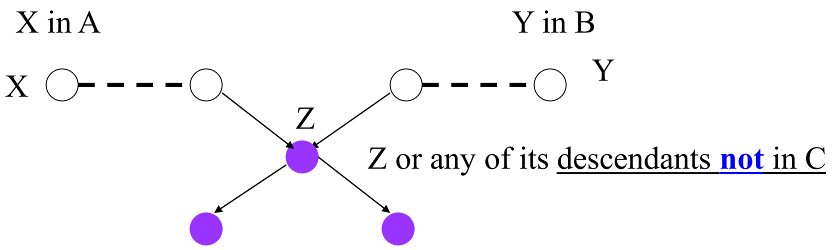
- **2. Path blocking with the wedge substructure**



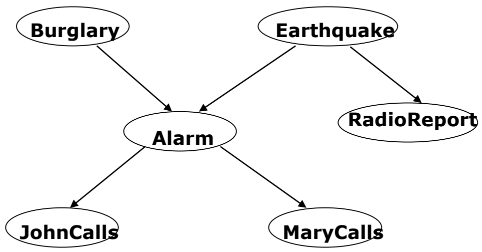
Undirected path blocking

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

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

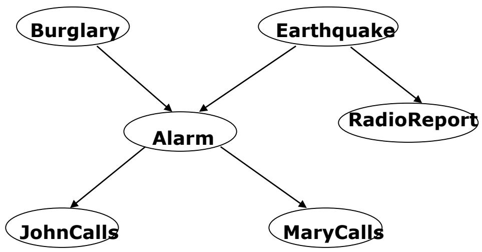


Independences in BBNs



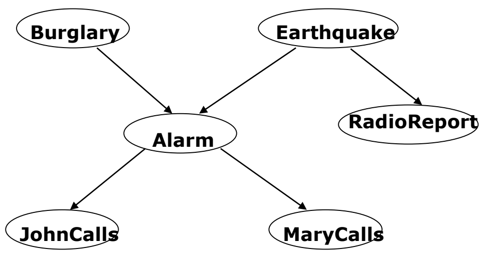
- Earthquake and Burglary are independent given MaryCalls ?

Independences in BBNs



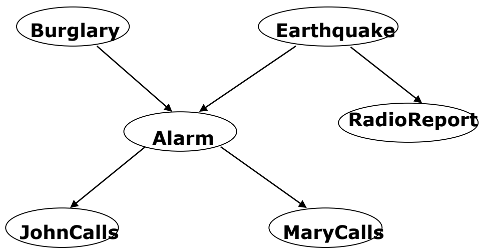
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **?**

Independences in BBNs



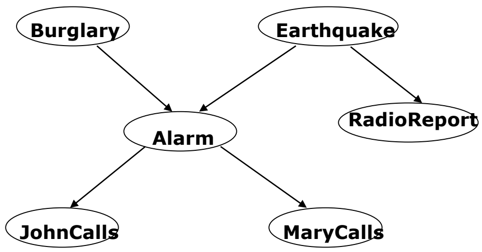
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **?**

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 **?**

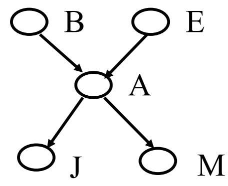
Independences in BBNs



- 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**

Full joint distribution in BBNs

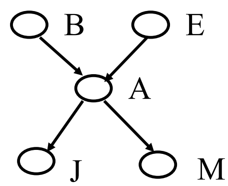
Rewrite the full joint probability using the product rule:



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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

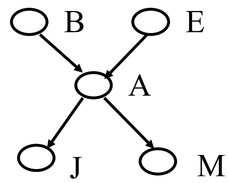


$$P(B=T,E=T,A=T,J=T,M=F)=$$

Product rule

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

Full joint distribution in BBNs



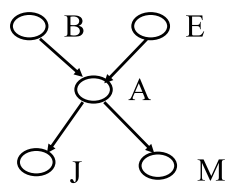
Rewrite the full joint probability using the product rule:

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

Product rule

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

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

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

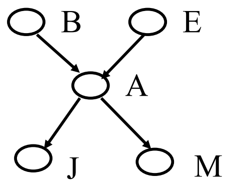
$$=P(J=T \mid B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$
$$=\underline{P(J=T \mid A=T)}P(B=T,E=T,A=T,M=F)$$

Product rule

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

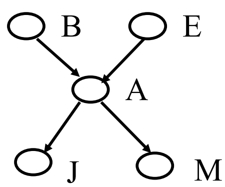


$$P(B=T,E=T,A=T,J=T,M=F)=$$

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

Full joint distribution in BBNs

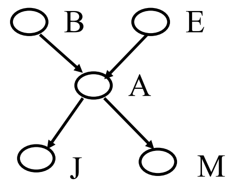
Rewrite the full joint probability using the product rule:



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

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

Full joint distribution in BBNs

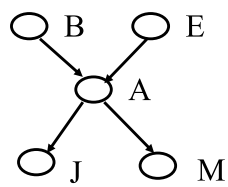


Rewrite the full joint probability using the product rule:

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

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

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

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

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

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

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:

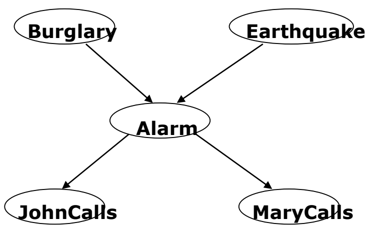
$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

- **What did we save?**

Alarm example: binary (True, False) variables

# of parameters of the full joint:

?



Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:

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

- **What did we save?**

Alarm example: binary (True, False) variables

# of parameters of the full joint:

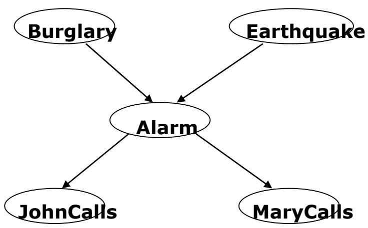
$$2^5 = 32$$

One parameter is for free:

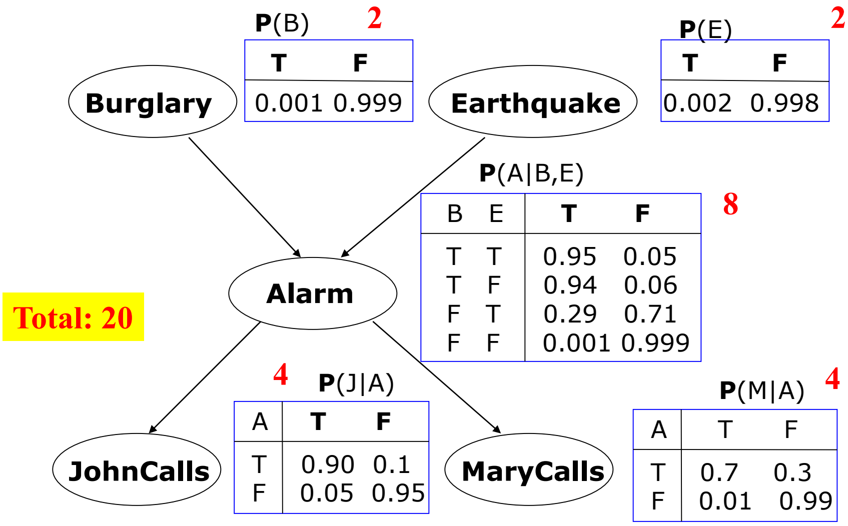
$$2^5 - 1 = 31$$

# of parameters of the BBN:

?



Bayesian belief network: parameters count



P(B) — 2

T	F
0.001	0.999

P(E) — 2

T	F
0.002	0.998

P(A|B,E) — 8

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) — 4

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A) — 4

A	T	F
T	0.7	0.3
F	0.01	0.99

Total: 20

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:

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

- **What did we save?**

Alarm example: 5 binary (True, False) variables

# of parameters of the full joint:

$$2^5 = 32$$

One parameter is for free:

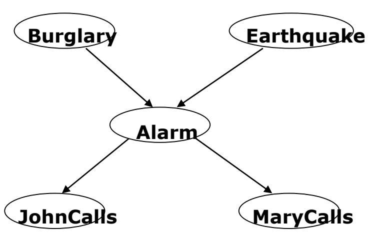
$$2^5 - 1 = 31$$

# of parameters of the BBN:

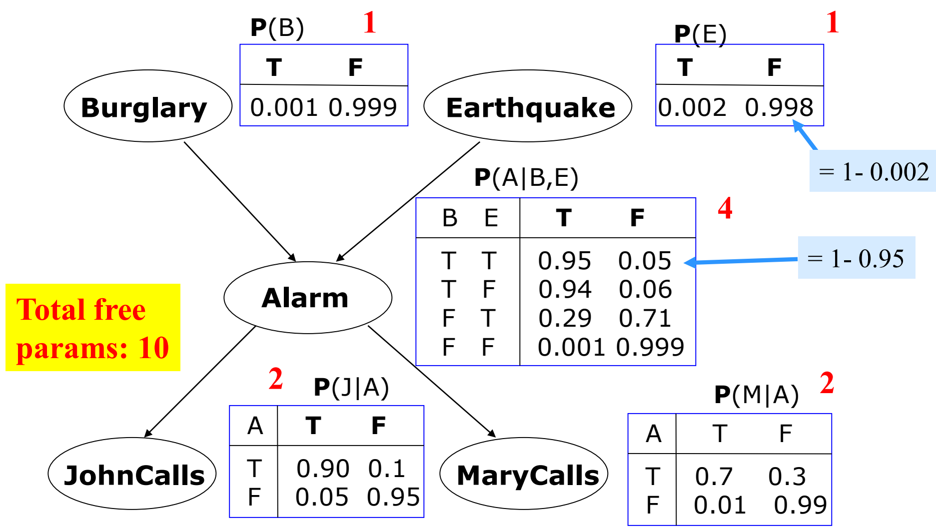
$$2^3 + 2(2^2) + 2(2) = 20$$

One parameter in every conditional is for free:

?



Bayesian belief network: free parameters



Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:

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

- **What did we save?**

Alarm example: 5 binary (True, False) variables

# of parameters of the full joint:

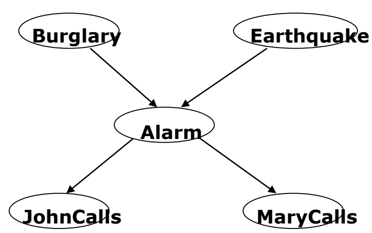
$$2^5 = 32$$

One parameter is for free:

$$2^5 - 1 = 31$$

# of parameters of the BBN:

$$2^3 + 2(2^2) + 2(2) = 20$$



One parameter in every conditional is for free:

$$2^2 + 2(2) + 2(1) = 10$$