CS 2750 Machine Learning
Lecture 16

Bayesian belief networks II

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

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)

Bayesian belief networks (BBNs)

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

Key features:

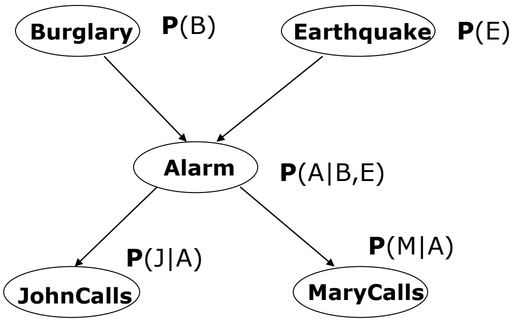
- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables
- **X and Y are independent** $P(X,Y) = P(X)P(Y)$
- **X and Y are conditionally independent given Z**

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

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

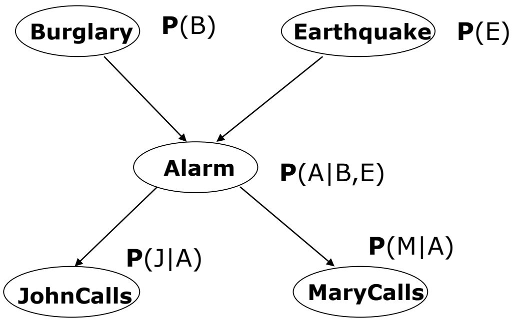
Bayesian belief network

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

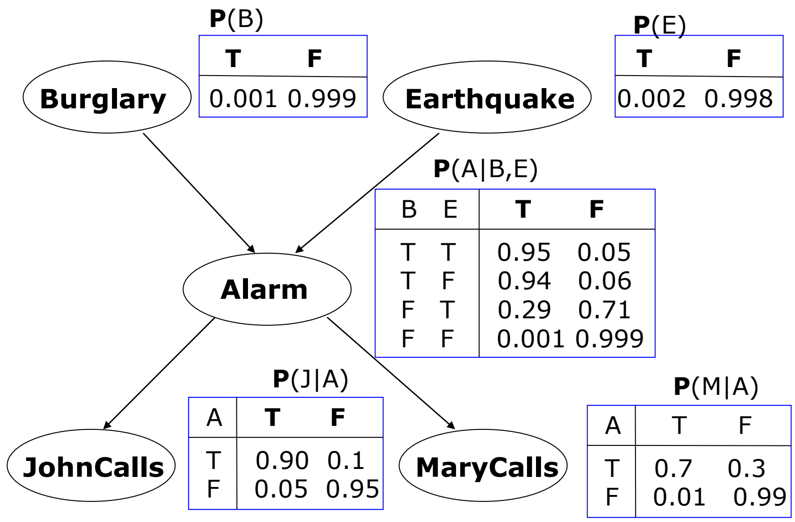


Bayesian belief network

2. **Local conditional distributions**
   - relating variables and their parents



Bayesian belief network

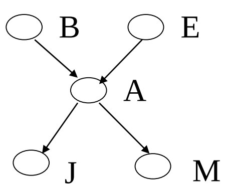


Full joint distribution in BBNs

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?**

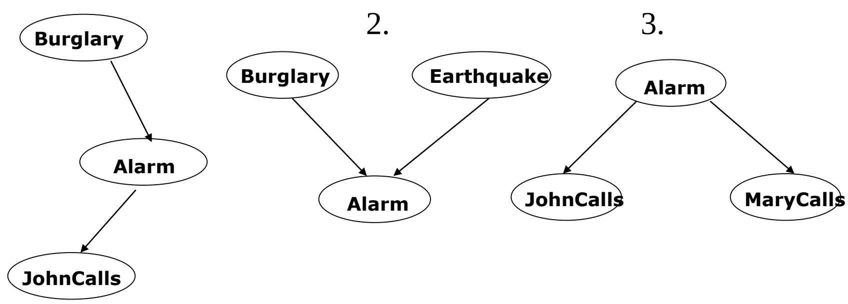
Answer:

- **Chain rule +**
- **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 !!!**

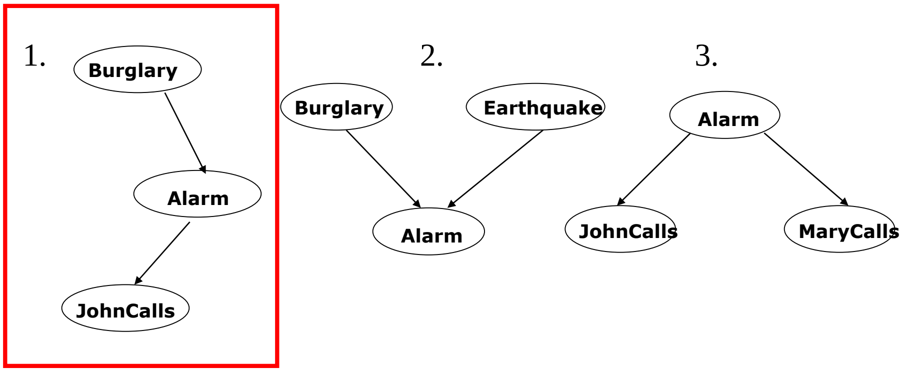
Independences in BBNs

3 basic independence structures:

1. 2. 3.



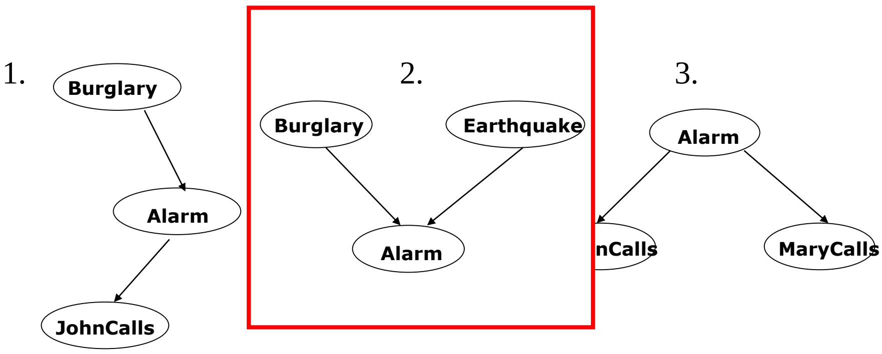
Independences in BBNs



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

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

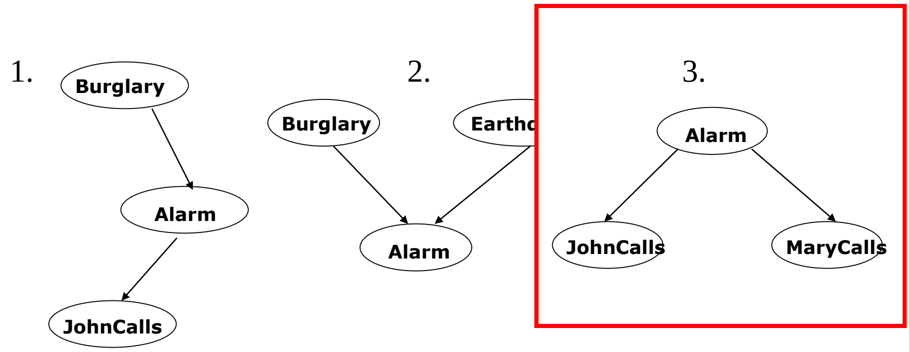
Independences in BBNs



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

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

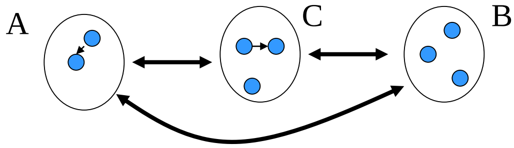
$$P(J, M \mid A) = P(J \mid A)P(M \mid A)$$

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

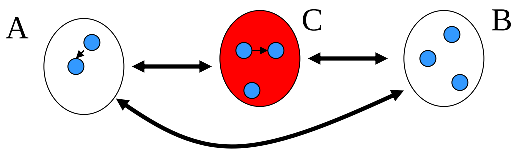
Undirected path blocking

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



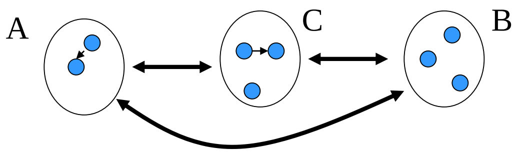
Undirected path blocking

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

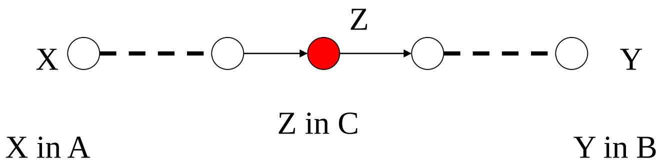


Undirected path blocking

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



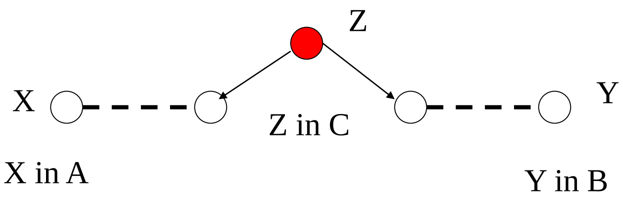
- **1. Path blocking with a linear substructure**



Undirected path blocking

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

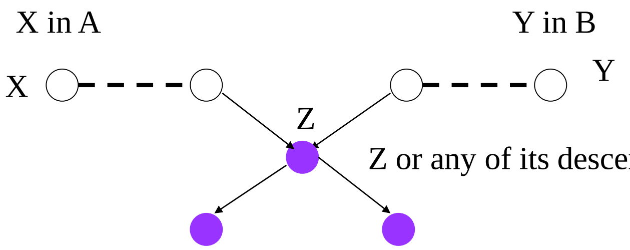
- **2. Path blocking with the wedge substructure**



Undirected path blocking

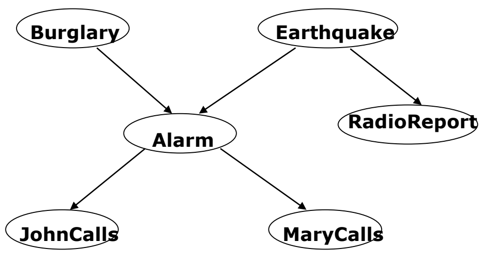
A is d-separated from B given C if every undirected path between them is **blocked**

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



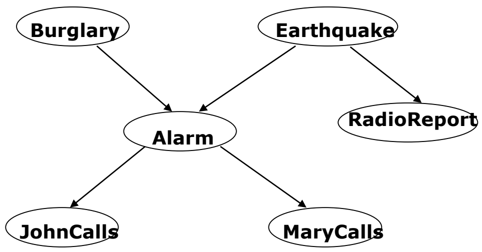
Z or any of its descriptions **not** in C

Independences in BBNs



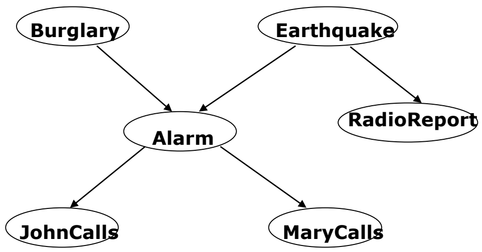
- Earthquake and Burglary are independent given MaryCalls **?**

Independences in BBNs



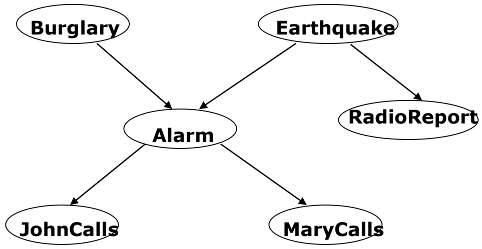
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **?**

Independences in BBNs



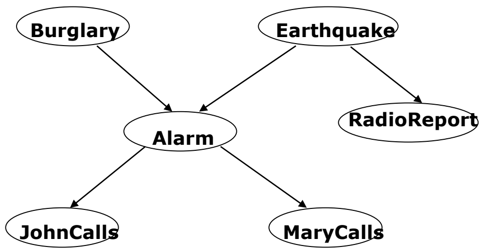
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **?**

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

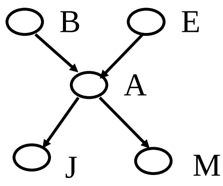


- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **T**

- Burglary and RadioReport are independent given MaryCalls **F**

Full joint distribution in BBNs

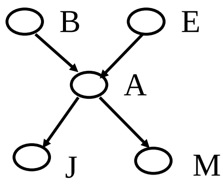
Rewrite the full joint probability using the product rule:



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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



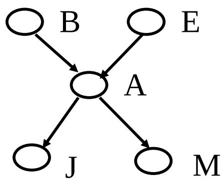
$$P(B=T, E=T, A=T, J=T, M=F) =$$

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:



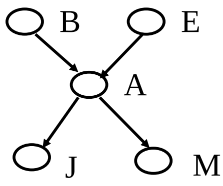
$P(B=T,E=T,A=T,J=T,M=F)=$ **Product rule**

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



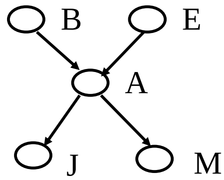
$P(B=T,E=T,A=T,J=T,M=F)=$

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

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

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

Full joint distribution in BBNs

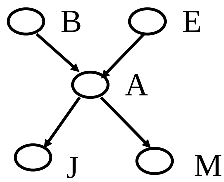


Rewrite the full joint probability using the product rule:

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

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

Full joint distribution in BBNs



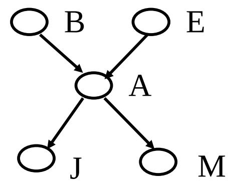
Rewrite the full joint probability using the product rule:

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

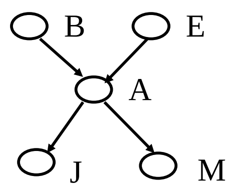


$$P(B=T,E=T,A=T,J=T,M=F)=$$

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



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

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

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

Parameter complexity problem

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

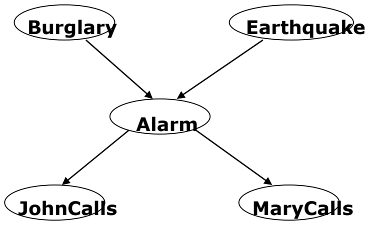
$$\mathbf{P}(X_1, X_2,.., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

- **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:

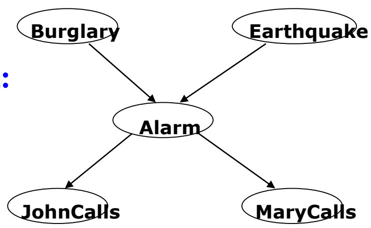
$$2^5 = 32$$

One parameter depends on the rest:

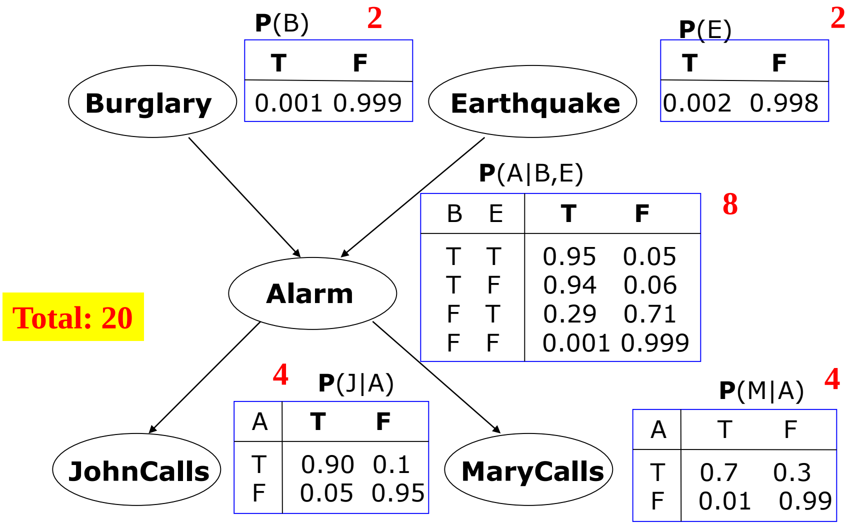
$$2^5 - 1 = 31$$

# 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:

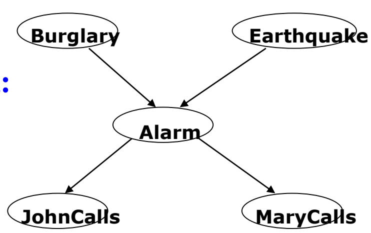
$$2^5 = 32$$

One parameter depends on the rest:

$$2^5 - 1 = 31$$

# of parameters of the BBN:

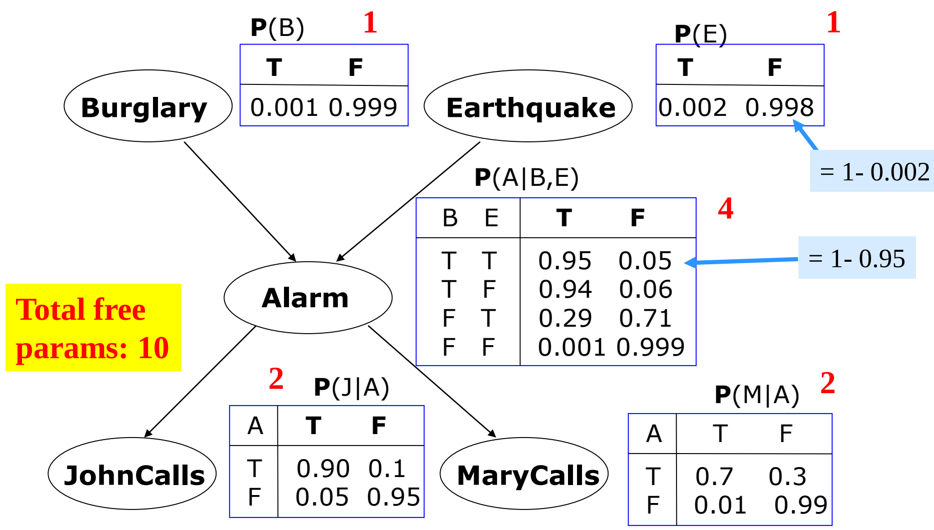
$$2^3 + 2(2^2) + 2(2) = 20$$



One parameter in every conditional depends on the rest:

?

Bayesian belief network: free parameters



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

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:

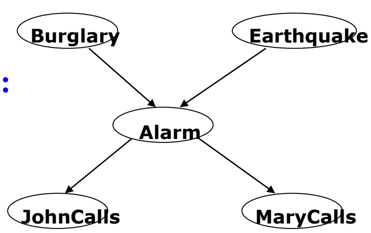
$$2^5 = 32$$

One parameter depends on the rest:

$$2^5 - 1 = 31$$

# of parameters of the BBN:

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



One parameter in every conditional depends on the rest:

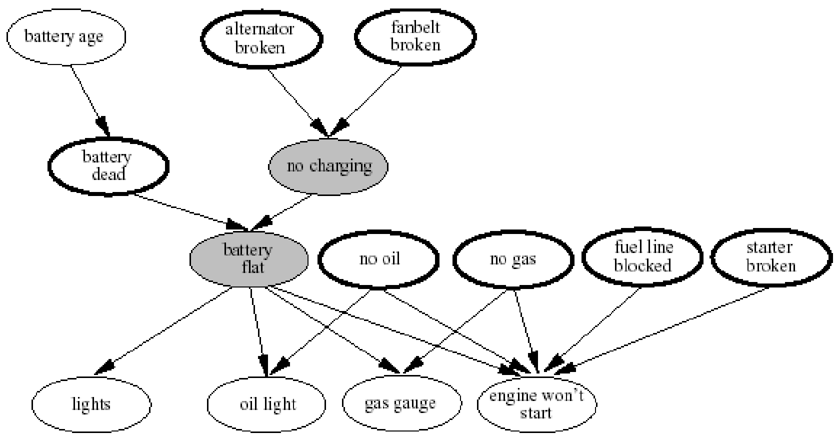
$$2^2 + 2(2) + 2(1) = 10$$

BBNs examples

- **In various areas:**
  - Intelligent user interfaces (Microsoft)
  - Troubleshooting, diagnosis of a technical device
  - Medical diagnosis:
    - Pathfinder CPSC
    - Munin
    - QMR-DT
  - Collaborative filtering
  - Military applications
  - Insurance, credit applications

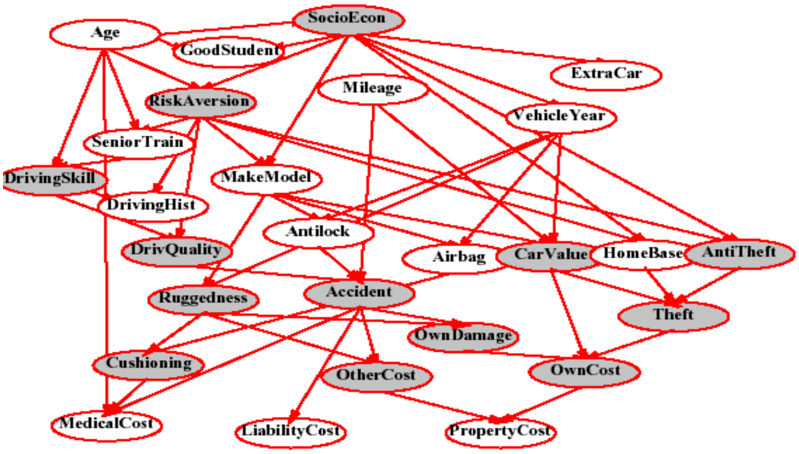
Diagnosis of car engine

- Diagnose the engine start problem

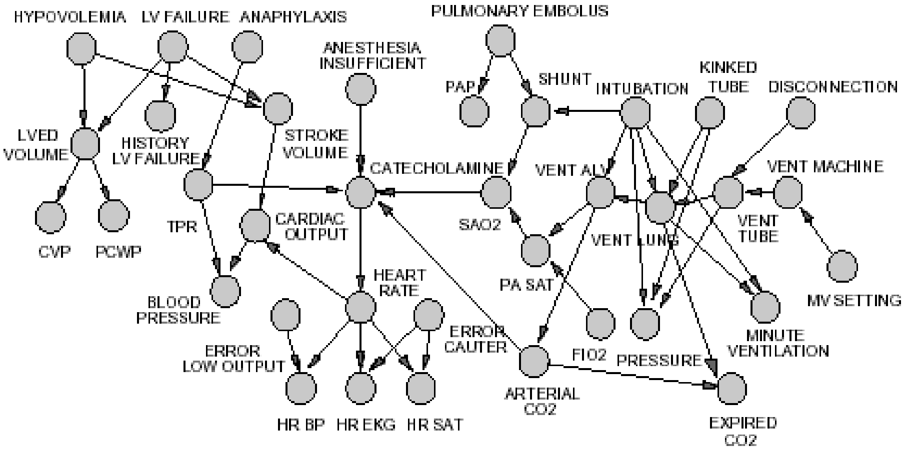


Car insurance example

- Predict claim costs (medical, liability) based on application data

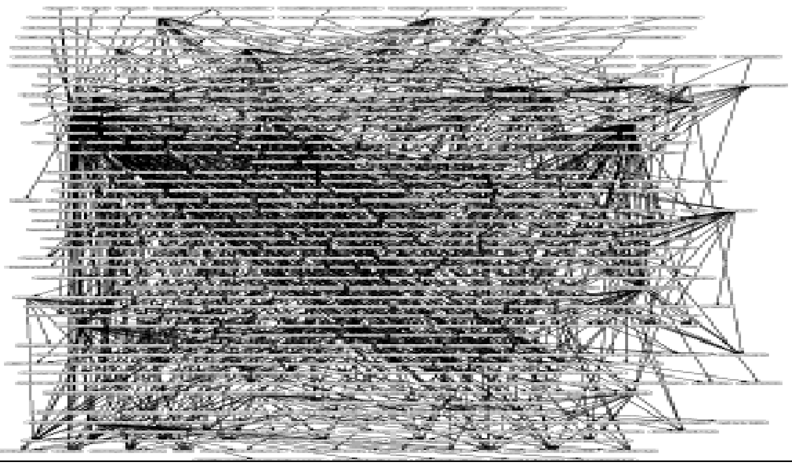


(ICU) Alarm network



CPCS

- **C**omputer-based **P**atient **C**ase **S**imulation system (CPCS-PM) developed by Parker and Miller (at University of Pittsburgh)
- 422 nodes and 867 arcs



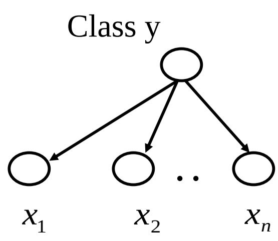
Naïve Bayes model

A **special (simple) Bayesian belief network**

- **Defines a generative classifier model**
- Model of P(**x** ,y) = P(**x** | y) P(y)
  - **Class variable y**

    p(y)
  - **Attributes are independent given y**

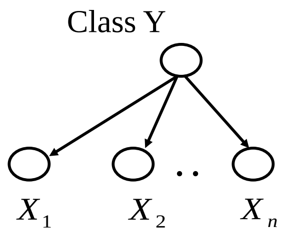
$$p(\mathbf{x} \mid y = i) = \prod_{j=1}^{d} p(x_j \mid y = i)$$



Learning:

- Parameterize models of p(y) and all p(x_j | y=i)
- ML estimates of the parameters

Naïve Bayes model



A **special (simple) Bayesian belief network**

- **Defines a generative classifier model**
- Model of P(**x** ,y) = P(**x** | y) P(y)

Classification: given ***x*** select the class

- Select the class with the maximum posterior
- Calculation of a posterior is an example of BBN inference

$$p(y=i \mid \mathbf{x}) = \frac{p(y=i)\, p(\mathbf{x} \mid y=i)}{\sum_{u=1}^{k} p(y=u)\, p(\mathbf{x} \mid y=u)} = \frac{p(y=i) \prod_{j=1}^{d} p(x_j \mid y=i)}{\sum_{u=1}^{k} p(y=u) \prod_{j=1}^{d} p(x_j \mid y=u)}$$

Remember: we can calculate the probabilities from the full joint