

CS 3750 Machine Learning

Lecture 4

Bayesian belief networks

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

CS 3750 Advanced Machine Learning

Parametric distribution models

Distribution models discussed so far:

- Bernoulli distribution
- Binomial distribution
- Multinomial distribution
- Normal distribution

and other members of the exponential family

- let us model distributions over random variables that represent variety of real-world events and patterns
- However these models are often not feasible when considering a large number of random variables and their multivariate distribution

CS 3750 Advanced Machine Learning

Density estimation

- So far we have covered density estimation for “simple” distribution models:
 - Bernoulli
 - Binomial
 - Multinomial
 - Gaussian
 - Poisson

But what if:

- The dimension of $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$ is large
 - Example: patient data
- Compact parametric distributions do not seem to fit the data
 - E.g.: multivariate Gaussian may not fit
- We have only a “small” number of examples to do accurate parameter estimates

CS 3750 Advanced Machine Learning

Modeling of complex multivariate distributions

How to model complex multivariate distributions with a large number of variables?

Example:

- Assume n binary random variables
- The full joint model requires 2^n parameters
- Not feasible for larger n

Solution:

- **Decompose the distribution and its model**

Two models to achieve the decomposition:

- Bayesian belief networks
- Markov random fields

CS 3750 Advanced Machine Learning

Bayesian belief networks (BBNs)

Bayesian belief network:

- Represents the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Takes 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**

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

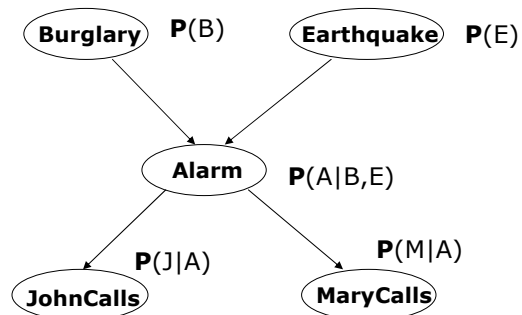
$$P(A | C, B) = P(A | C)$$

Bayesian belief network

1. Directed acyclic graph

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

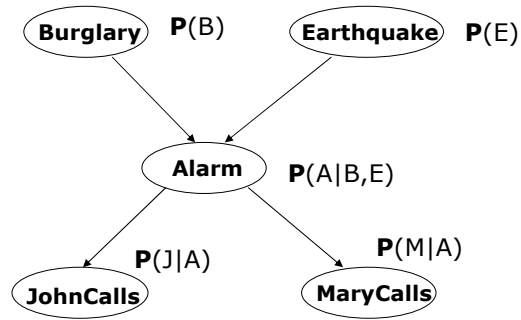
The chance of Alarm being is influenced by Earthquake,
The chance of John calling is affected by the Alarm



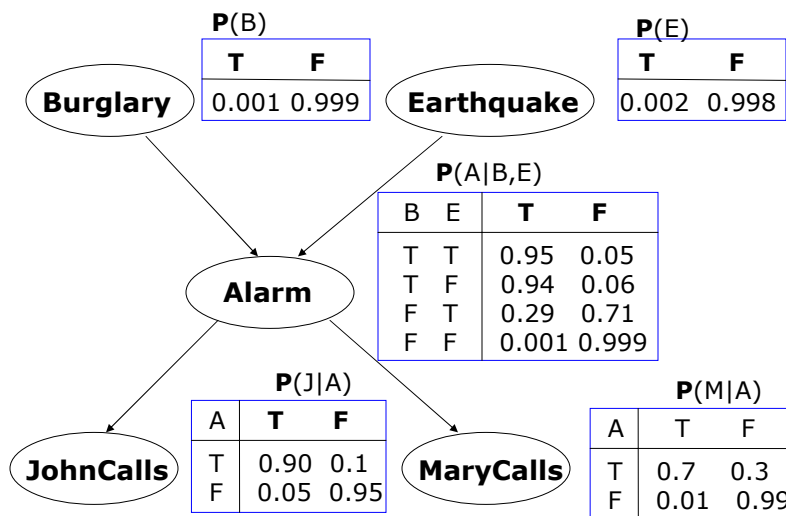
Bayesian belief network

2. Local conditional distributions

- relate 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):

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

Example:

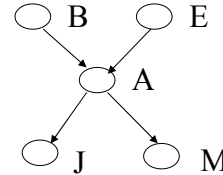
Assume the following assignment of values to random variables

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

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**

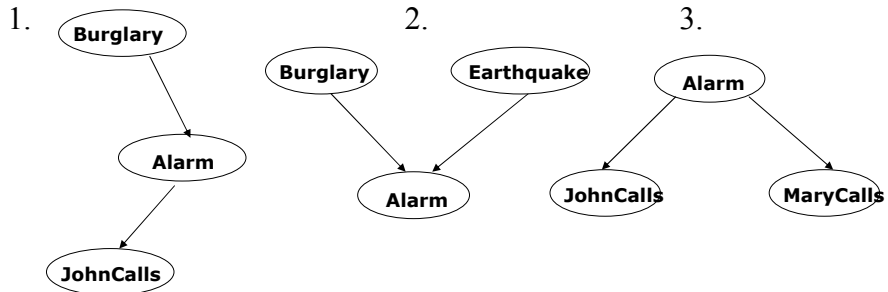
$$P(A \mid C, B) = P(A \mid C)$$

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

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

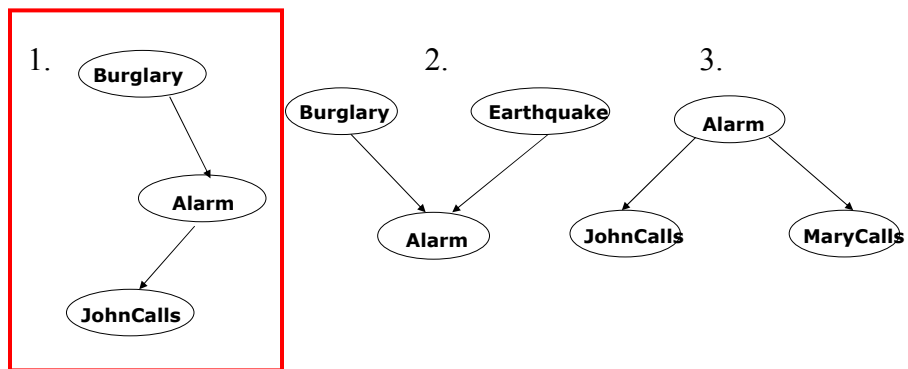
Independences in BBNs

3 basic independence structures:



CS 3750 Advanced Machine Learning

Independences in BBNs



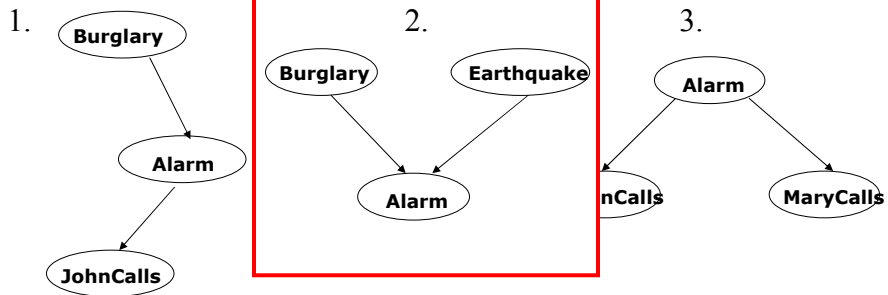
1. JohnCalls is **independent** of Burglary given Alarm

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

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

CS 3750 Advanced Machine Learning

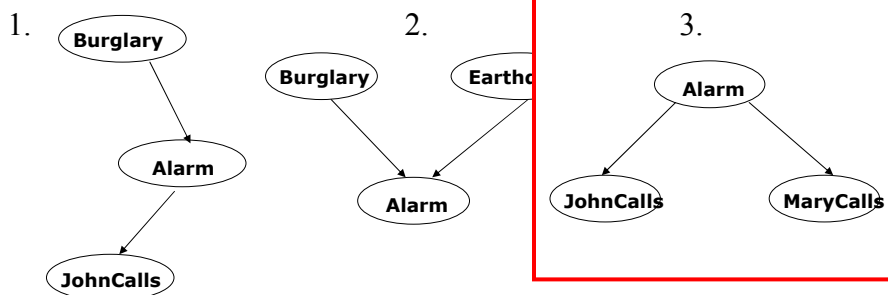
Independences in BBNs



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

$$P(J | A, M) = P(J | A)$$

$$P(J, M | A) = P(J | A)P(M | A)$$

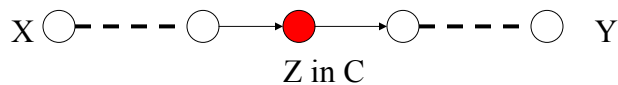
Independences in BBN

- 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**
- **Path blocking**
 - 3 cases that expand on three basic independence structures

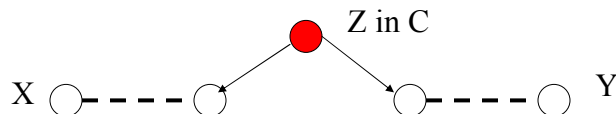
CS 3750 Advanced Machine Learning

Undirected path blocking

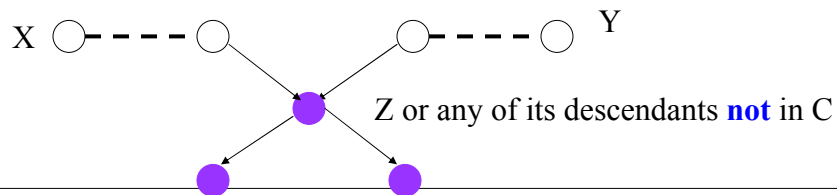
- 1. With linear substructure



- 2. With wedge substructure

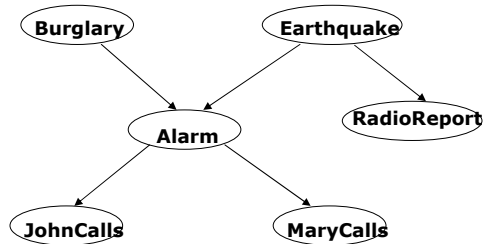


- 3. With vee substructure



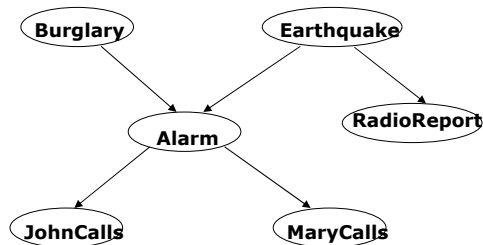
CS 3750 Advanced Machine Learning

Independences in BBNs



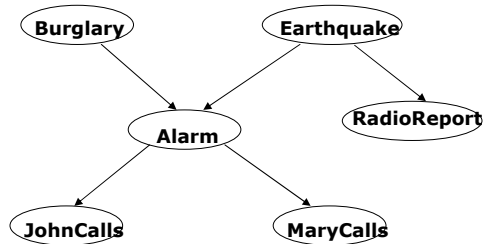
- Earthquake and Burglary are independent given MaryCalls ?

Independences in BBNs



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

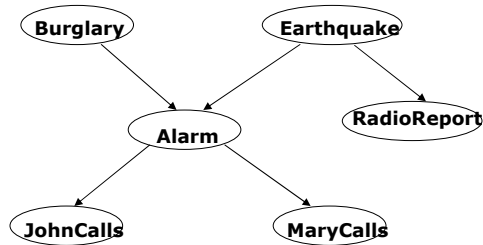
Independences in BBNs



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

CS 3750 Advanced Machine Learning

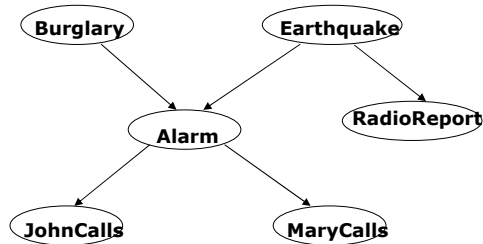
Independences in BBNs



- 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 **?**

CS 3750 Advanced Machine Learning

Independences in BBNs



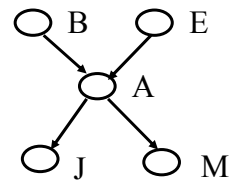
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **T**
- Burglary and RadioReport are independent given MaryCalls **F**

CS 3750 Advanced Machine Learning

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

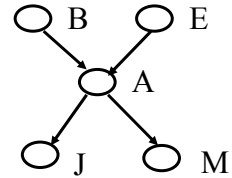
$$P(B=T, E=T, A=T, J=T, M=F) =$$



CS 3750 Advanced Machine Learning

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



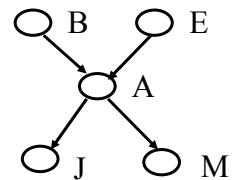
$$P(B=T, E=T, A=T, J=T, M=F) =$$

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



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

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

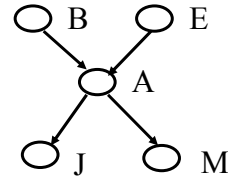
$$= \underline{P(J=T \mid A=T)} P(B=T, E=T, A=T, M=F)$$

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



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

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

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

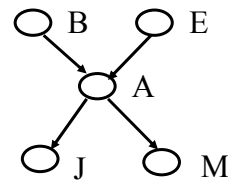
$$P(M=F \mid B=T, E=T, A=T) P(B=T, E=T, A=T)$$

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



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

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

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

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

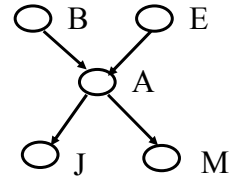
$$\underline{P(M=F \mid A=T)} P(B=T, E=T, A=T)$$

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:



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

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

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

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

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

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

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

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

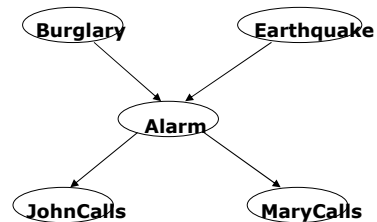
CS 3750 Advanced Machine Learning

Parameter complexity problem

- In the BBN the full joint distribution is expressed as a product of conditionals (of smaller) complexity

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

Parameters:
full joint: ?
BBN: ?



CS 3750 Advanced Machine Learning

Parameter complexity problem

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

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

- What did we save?**

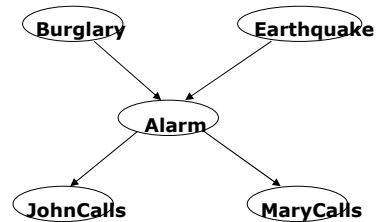
Alarm example: 5 binary (True, False) variables

of parameters of the full joint:

$$2^5 = 32$$

One parameter is for free:

$$2^5 - 1 = 31$$



Parameter complexity problem

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

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

- What did we save?**

Alarm example: 5 binary (True, False) variables

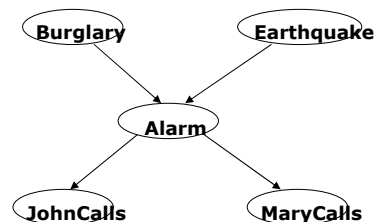
of parameters of the full joint:

$$2^5 = 32$$

One parameter is for free:

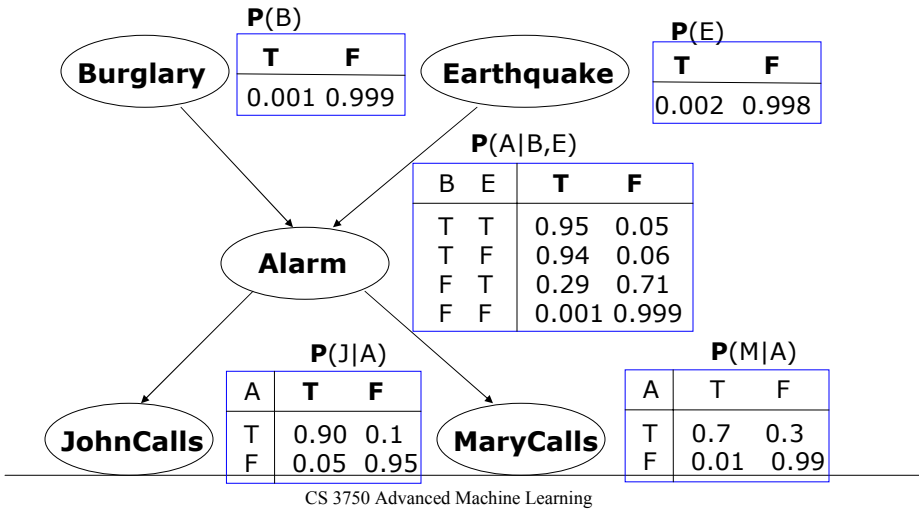
$$2^5 - 1 = 31$$

of parameters of the BBN: ?



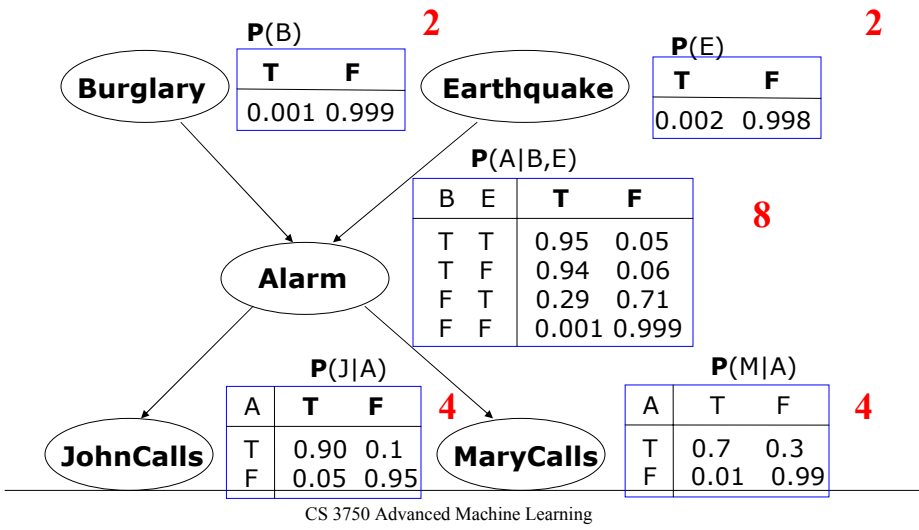
Bayesian belief network.

- In the BBN the **full joint distribution** is expressed using a set of local conditional distributions



Bayesian belief network.

- In the BBN the **full joint distribution** is expressed using a set of local conditional distributions



Parameter complexity problem

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

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

- 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:

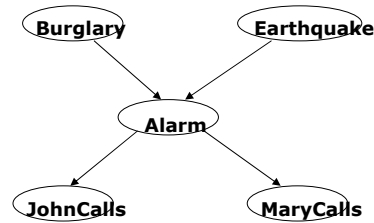
$$2^5 - 1 = 31$$

of parameters of the BBN:

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

One parameter in every conditional is for free:

?



Parameter complexity problem

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

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

- 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:

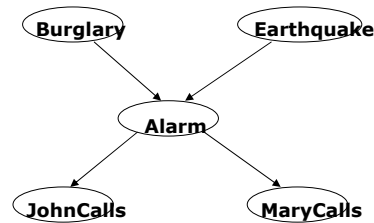
$$2^5 - 1 = 31$$

of parameters of the BBN:

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

One parameter in every conditional is for free:

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



Model acquisition problem

The structure of the BBN typically reflects causal relations

- BBNs are also sometime referred to as **causal networks**
- Causal structure is very intuitive in many applications domain and it is relatively easy to obtain from the domain expert

Probability parameters of BBN correspond to conditional distributions relating a random variable and its parents only

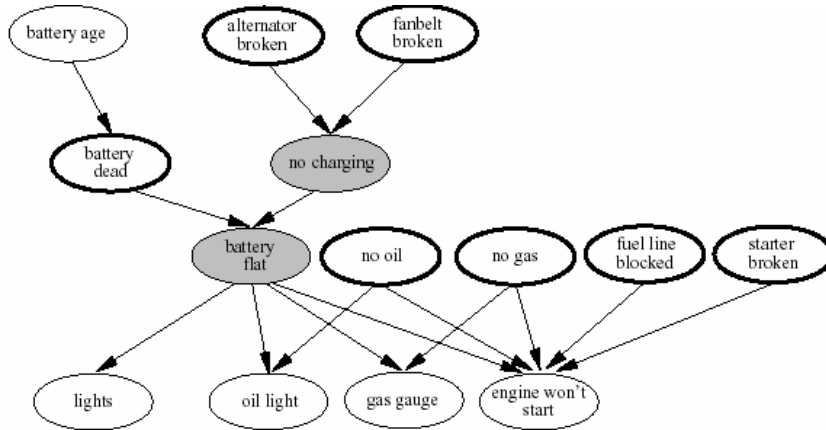
- Their complexity much smaller than the full joint
- Easier to come up (estimate) the probabilities from expert or automatically by learning from data

BBNs built in practice

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

Diagnosis of car engine

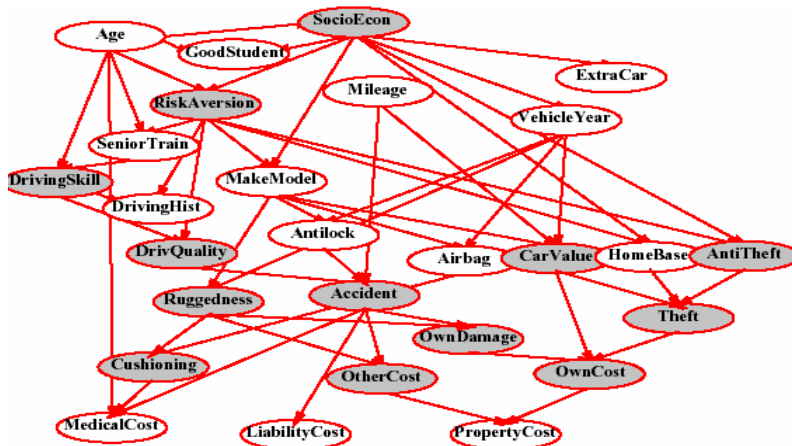
- Diagnose the engine start problem



CS 3750 Advanced Machine Learning

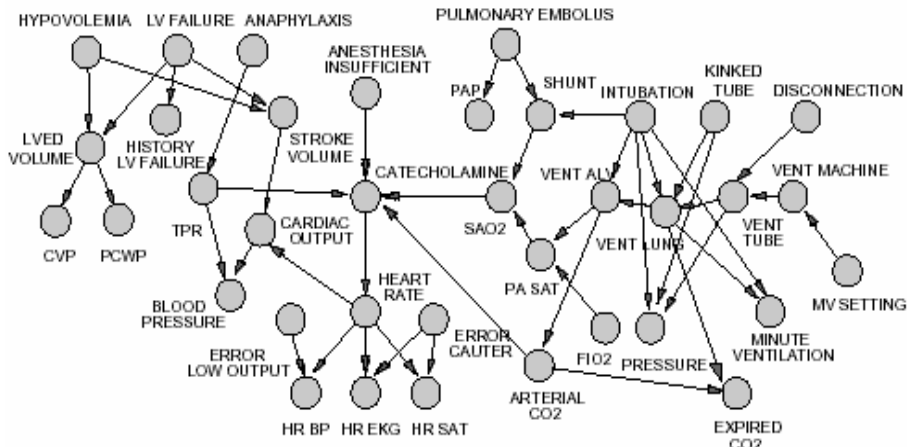
Car insurance example

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



CS 3750 Advanced Machine Learning

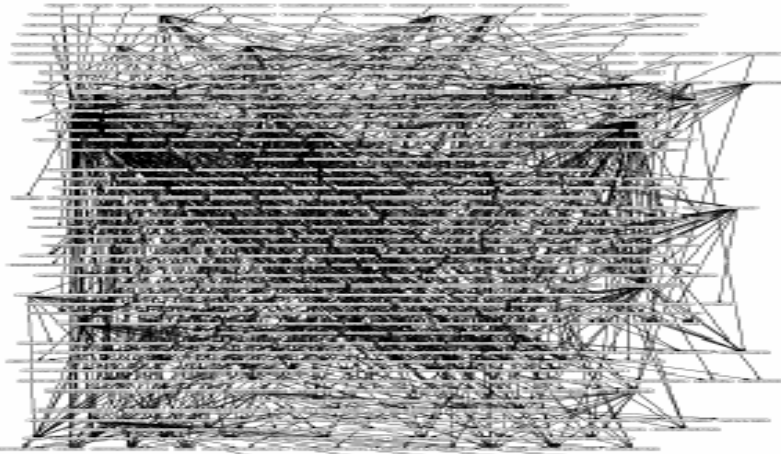
(ICU) Alarm network



CS 3750 Advanced Machine Learning

CPCS

- Computer-based Patient Case Simulation system (CPCS-PM) developed by Parker and Miller (at University of Pittsburgh)
- 422 nodes and 867 arcs



CS 3750 Advanced Machine Learning

QMR-DT

- **Medical diagnosis in internal medicine**

Bipartite network of disease/findings relations

