CS 2750 Machine Learning
Lecture 15

Bayesian belief networks

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

Midterm exam

Midterm exam

- **Thursday, March 5, 2020**
- **In-class**
- **Closed book**

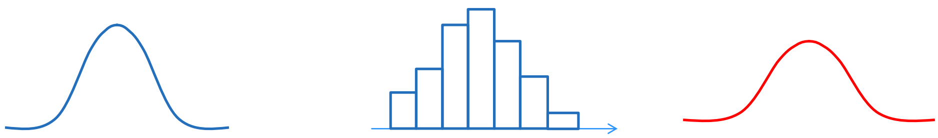
Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying 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 $\hat{p}(\mathbf{X})$

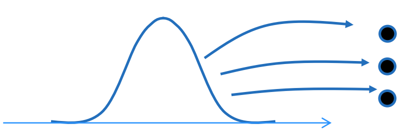


Density estimation

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 $p(\mathbf{X})$)**



Independently drawn instances from the same fixed distribution

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in ***X*** with parameters Θ :

$$\hat{p}(\mathbf{X} \mid \Theta)$$

- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: Find the parameters Θ that explain the observed data the best

Parameter estimation

- **Maximum likelihood (ML)**

  maximize $p(D \mid \Theta, \xi)$

  – yields: one set of parameters Θ_{ML}

  – the target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta}_{ML})$$

- **Bayesian parameter estimation**

  – uses the posterior distribution over possible parameters

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

  – Yields: all possible settings of Θ (and their "weights")

  – The target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\boldsymbol{\Theta}} p(X \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid D, \xi)\, d\boldsymbol{\Theta}$$

Parameter estimation

Other possible criteria:

- **Maximum a posteriori probability (MAP)**

  maximize $p(\mathbf{\Theta} \mid D, \xi)$ (mode of the posterior)

  – Yields: one set of parameters $\mathbf{\Theta}_{MAP}$

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

- **Expected value of the parameter**

  $\hat{\mathbf{\Theta}} = E(\mathbf{\Theta})$ (mean of the posterior)

  – Expectation taken with regard to posterior $p(\mathbf{\Theta} \mid D, \xi)$

  – Yields: one set of parameters

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \hat{\mathbf{\Theta}})$$

Distribution models

- **So far we have covered density estimation for "simple" distribution models:**
  – Bernoulli
  – Binomial
  – Multinomial
  – Gaussian
  – Poisson

But what if:

- The dimension of $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ 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 relatively "small" number of examples to learn many parameter estimates

Modeling complex distributions

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

Solution:
- **Decompose the distribution using conditional and marginal independence relations**
- **Decompose the parameter estimation problem to a set of smaller parameter estimation tasks**

Decomposition of distributions using conditional and marginal independence assumption is the main idea behind **Bayesian belief networks**

Example

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:
- Symptoms and disease are represented as random variables

Our objectives:
- Describe a multivariate distribution representing the relations between symptoms and disease
- Design inference and learning procedures for the multivariate model

Representation complexity

Example: modeling of disease – symptoms relations

- **Disease:** pneumonia (T?F)
- **Patient symptoms (findings, lab tests)**:
  - Fever (T/F) Cough (T/F), Paleness (T/F), WBC (white blood cells) count (High/Normal/Low), Chest pain (T/G), etc.
- **Model of the full joint distribution**: $\hat{p}(\mathbf{X})$
  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)
P(Pneumonia =T, Fever =T, Cought=T, WBC=High, Chest pain=F)
P(Pneumonia =T, Fever =T, Cought=T, WBC=Norm, Chest pain=T)

- **How many probabilities are there?**

Representation complexity

Example: modeling of disease – symptoms relations

- **Disease:** pneumonia (T?F)
- **Patient symptoms (findings, lab tests)**:
  - Fever (T/F) Cough (T/F), Paleness (T/F), WBC (white blood cells) count (High/Normal/Low), Chest pain (T/G), etc.
- **Model of the full joint distribution**: $\hat{p}(\mathbf{X})$
  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)
P(Pneumonia =T, Fever =T, Cought=T, WBC=High, Chest pain=F)
P(Pneumonia =T, Fever =T, Cought=T, WBC=Norm, Chest pain=T)

- **How many probabilities are there?** $2^5*3 = 32*3 = 96$
  $O(a^k)$ where k is the number of variables

Marginalization

Joint probability distribution (for a set variables)

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

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

$\mathbf{P}(Pneumonia)$

WBCcount

		high	*normal*	*low*	
Pneumonia	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
		0.005	0.993	0.002	

$\mathbf{P}(WBCcount)$

Marginalization (summing of rows, or columns)
- summing out variables

Joint distribution over a subset variables

- **Full joint distribution is defined over all variables we use in the model**

  E.g. **P**(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain)

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

  E.g.

  $$P(Pneumonia, WBCcount, Fever) =$$

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

- **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

Pneumonia		*high*	*normal*	*low*	$\mathbf{P}(Pneumonia)$
	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 !!!

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

WBCcount

Pneumonia		*high*	*normal*	*low*	$\mathbf{P}(Pneumonia)$
	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 probabilities can be expressed as a product of marginal probabilities:**

  $\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$

Bayesian belief networks (BBNs)

Proposed in 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**

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

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

Conditional probability: definitions

Conditional probability :

- Probability of A given B $$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

- Conditional probability is defined in terms of the joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities

Product rule

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

Chain rule

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

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

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

Variable we condition on

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

Bayesian belief networks (BBNs)

Proposed in 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**

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

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

Alarm system example

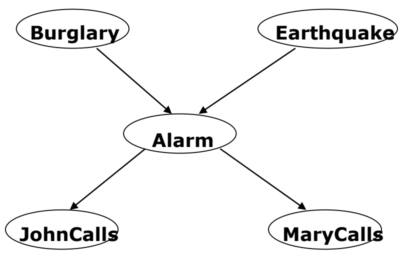
Story: Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.

We want to represent the relations among the events:

- Burglary, Earthquake, Alarm, Mary calls and John calls

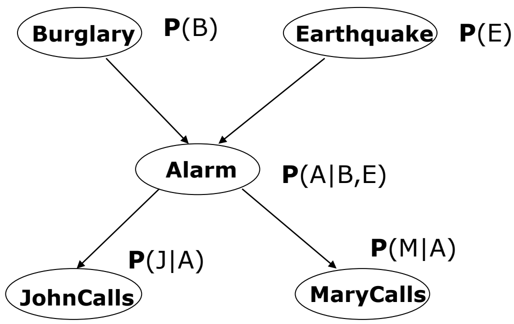
From the story we can extract (typically causal) relations among the events

Causal relations



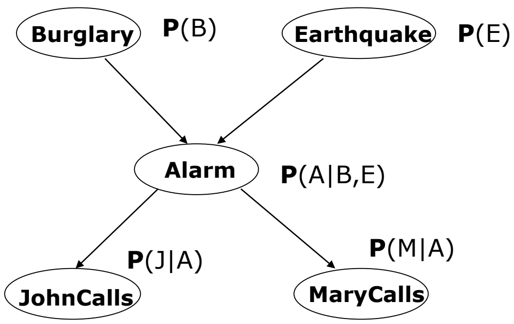
Bayesian belief network

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

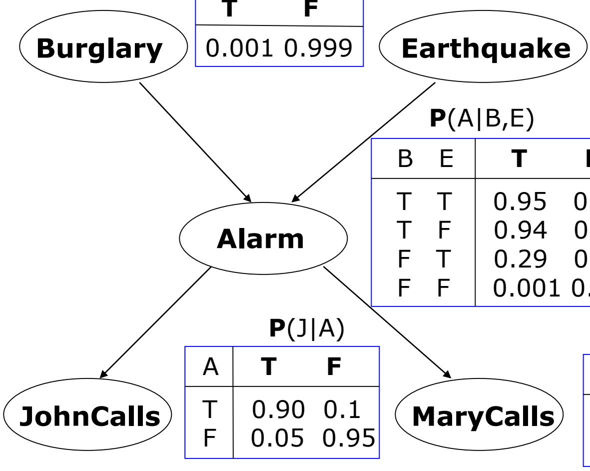


Bayesian belief network

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

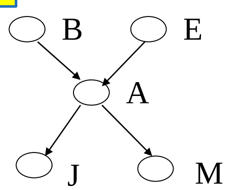
P(M|A)

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**

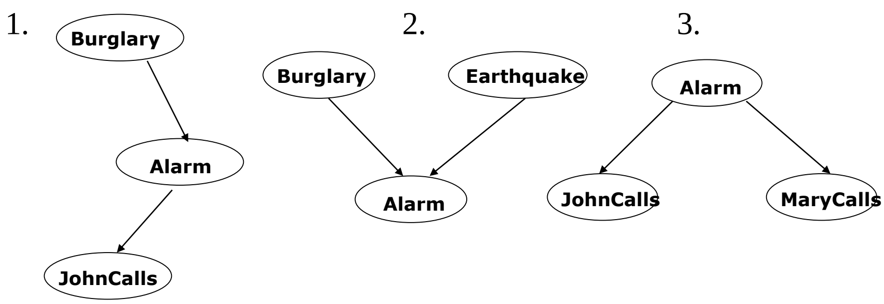
$$P(A \mid C,B) = P(A \mid C)$$

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

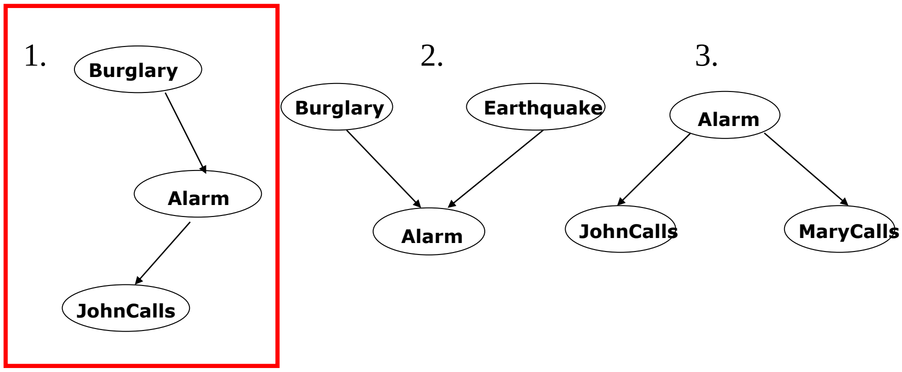
- **The graph structure implies the decomposition !!!**

Independences in BBNs

3 basic independence structures:



Independences in BBNs

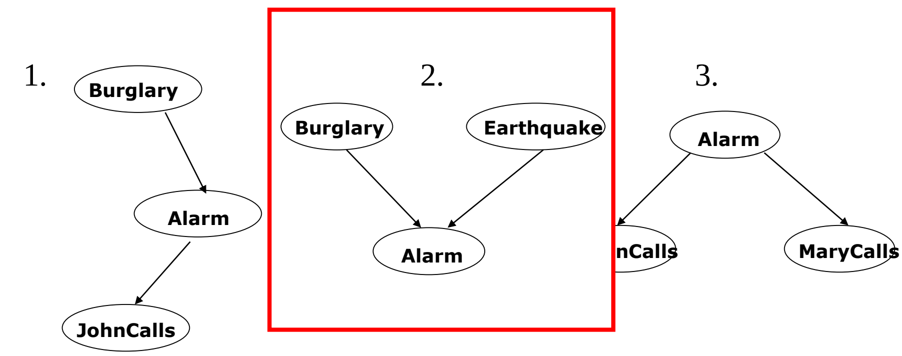


1. JohnCalls **is independent** of Burglary given Alarm

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

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

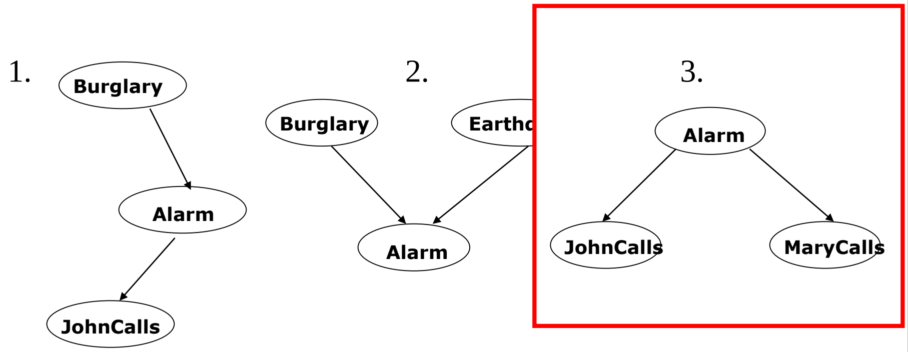
Independences in BBNs



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

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

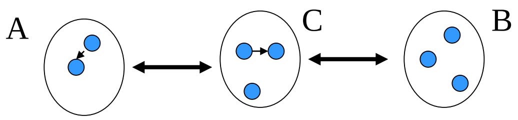
$$P(J, M \mid A) = P(J \mid A) P(M \mid A)$$

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

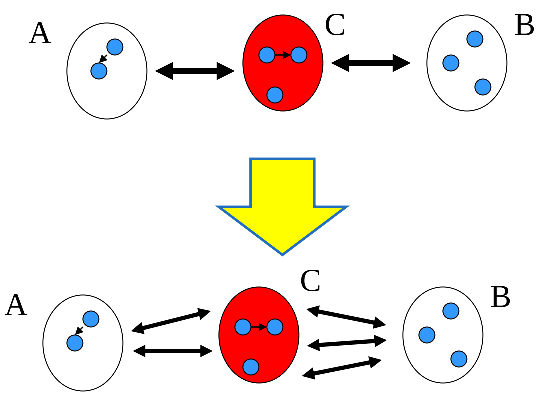
Undirected path blocking

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



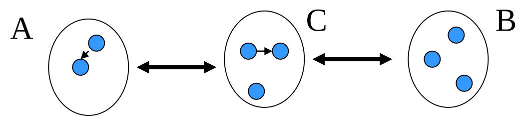
Undirected path blocking

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

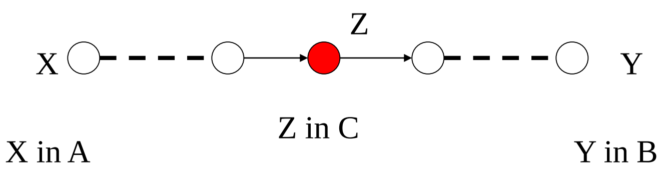


Undirected path blocking

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



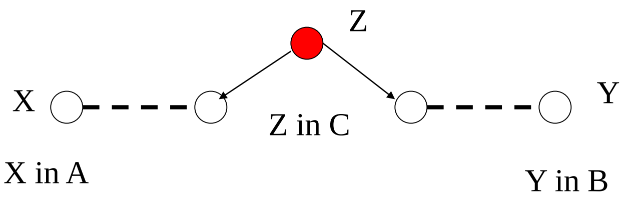
- **1. Path blocking with a linear substructure**



Undirected path blocking

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

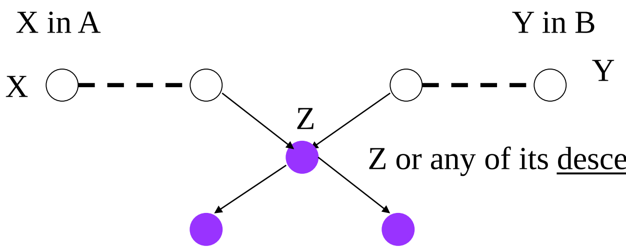
- **2. Path blocking with the wedge substructure**



Undirected path blocking

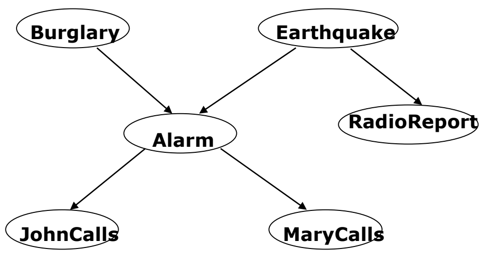
A is d-separated from B given C if every undirected path between them is **blocked**

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



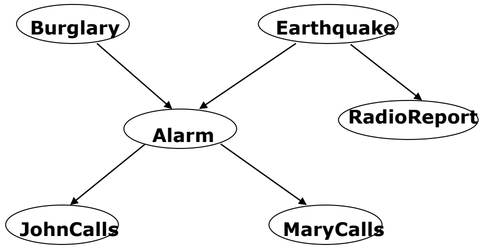
Z or any of its descendants **not** in C

Independences in BBNs



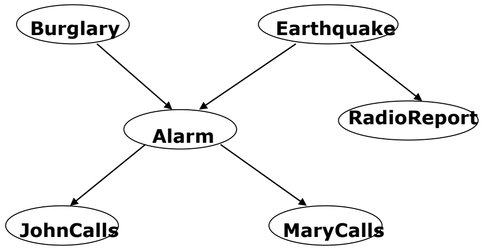
- Earthquake and Burglary are independent given MaryCalls **?**

Independences in BBNs



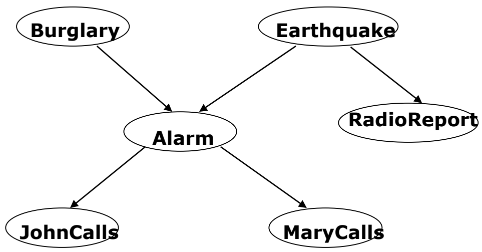
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **?**

Independences in BBNs



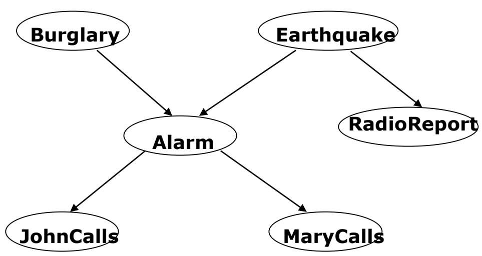
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **?**

Independences in BBNs



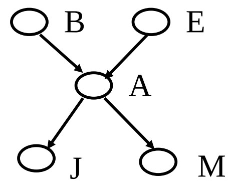
- Earthquake and Burglary are independent given MaryCalls **F**
- Burglary and MaryCalls are independent (not knowing Alarm) **F**
- Burglary and RadioReport are independent given Earthquake **T**
- Burglary and RadioReport are independent given MaryCalls **?**

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**

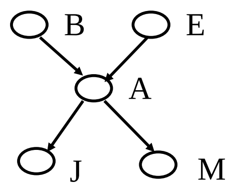
Full joint distribution in BBNs



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:

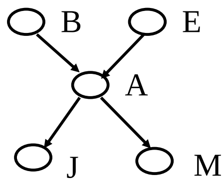
$$P(B=T,E=T,A=T,J=T,M=F)=$$

Product rule

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

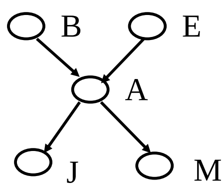


$P(B=T,E=T,A=T,J=T,M=F)=$ **Product rule**

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

Full joint distribution in BBNs

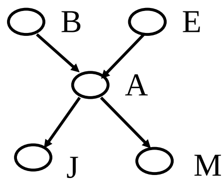
Rewrite the full joint probability using the product rule:



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

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

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

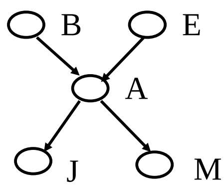
$$=P(J=T\mid B=T,E=T,A=T,M=F)P(B=T,E=T,A=T,M=F)$$

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

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

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

Full joint distribution in BBNs



Rewrite the full joint probability using the product rule:

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

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

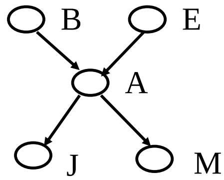
$$=\underline{P(J=T\mid A=T)}P(B=T,E=T,A=T,M=F)$$

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

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

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

Full joint distribution in BBNs

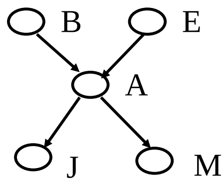


Rewrite the full joint probability using the product rule:

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

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

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:

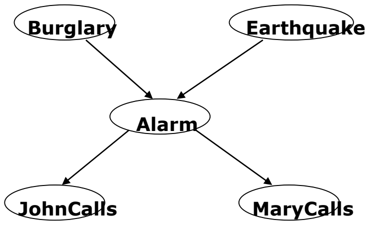
$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

- **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:

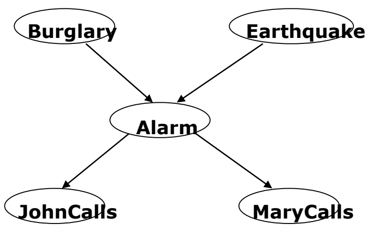
$$2^5 = 32$$

One parameter is for free:

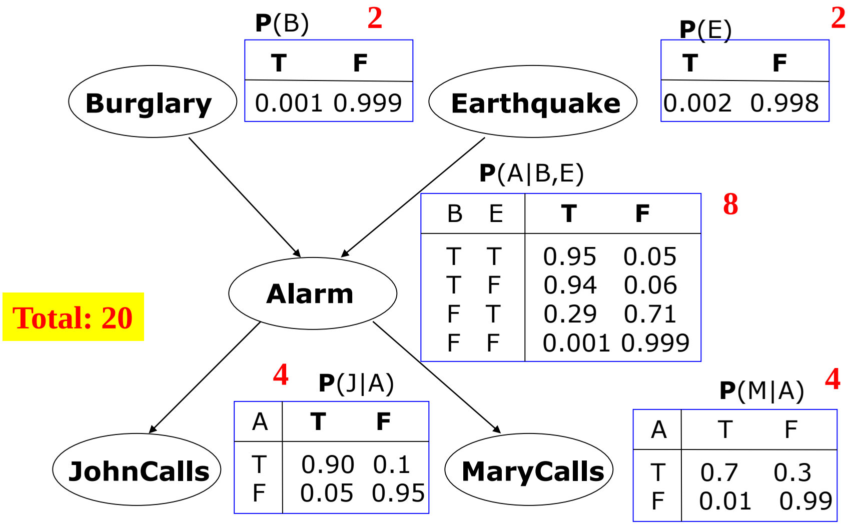
$$2^5 - 1 = 31$$

# 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

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:

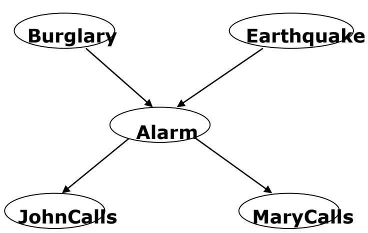
$$2^5 = 32$$

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

Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

P(B) 1

T	F
0.001	0.999

P(E) 1

T	F
0.002	0.998

= 1- 0.002

P(A|B,E) 4

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

= 1- 0.95

Total free params: 10

2 **P**(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A) 2

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:

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

- **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:

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