COMP307/AIML420 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Reasoning under uncertainty:

Bayesian Networks 2

Outline

- Introduction/Review
- Number of Free Parameters
- Building a BN
- Introduction to Inference in a BN
- Summary

Introduction/Review

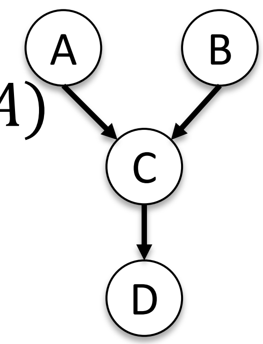
- **Naive Bayes (NB)** is a **simple type of Bayesian network**
  - **Use Bayes' theorem** to calculate the **probability of a particular class given a set of feature**

- **Bayesian networks (BN)** **“**extends” **NB** by allowing to model **dependencies between features** through a DAG
  - **Graphical models** that represent **probabilistic relationships between random variables**
  - **nodes** = random variables
  - **edges** = probabilistic dependency between variables

Introduction/Review

- **Factorisation:** a joint probability distribution is expressed as a product of simpler conditional probability distributions
- The product rule tells us we can always write

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



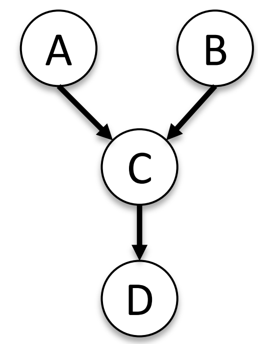
  - No structural constraints imposed
- For the example BN this simplifies to

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

  - Structural constraints imposed; fewer parameters and more robust
- The joint probability distribution over all variables in the network can be represented as a product of the conditional probabilities of each variable given its parents.

Using a Bayesian network: inference

- We are generally interested in a particular probability. For example, $P(A|D)$ in the example network:
  - Product rule says: $P(A|D) = \frac{P(A,D)}{P(D)}$
  - Here: $P(A,B,C,D) = P(D|C)P(C|A,B)P(B)PA)$
  - Total probability: $P(A,D) = \sum_{C,B} P(D|C)P(C|A,B)P(B)PA)$
  - Total probability: $P(D) = \sum_{A} P(A,D)$

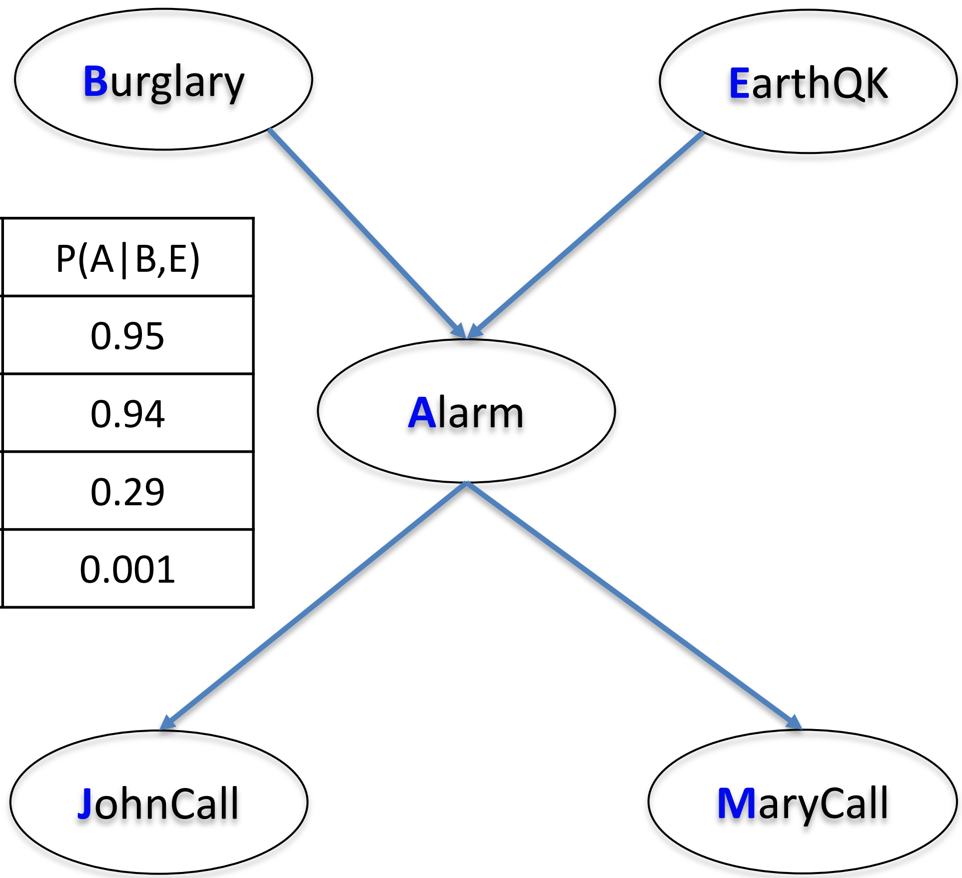


- Often faster computational methods exist

Number of free parameters

B	P(B)
T	0.001

E	P(E)
T	0.002



B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001

A	J	P(J|A)
T	T	0.9
F	T	0.05

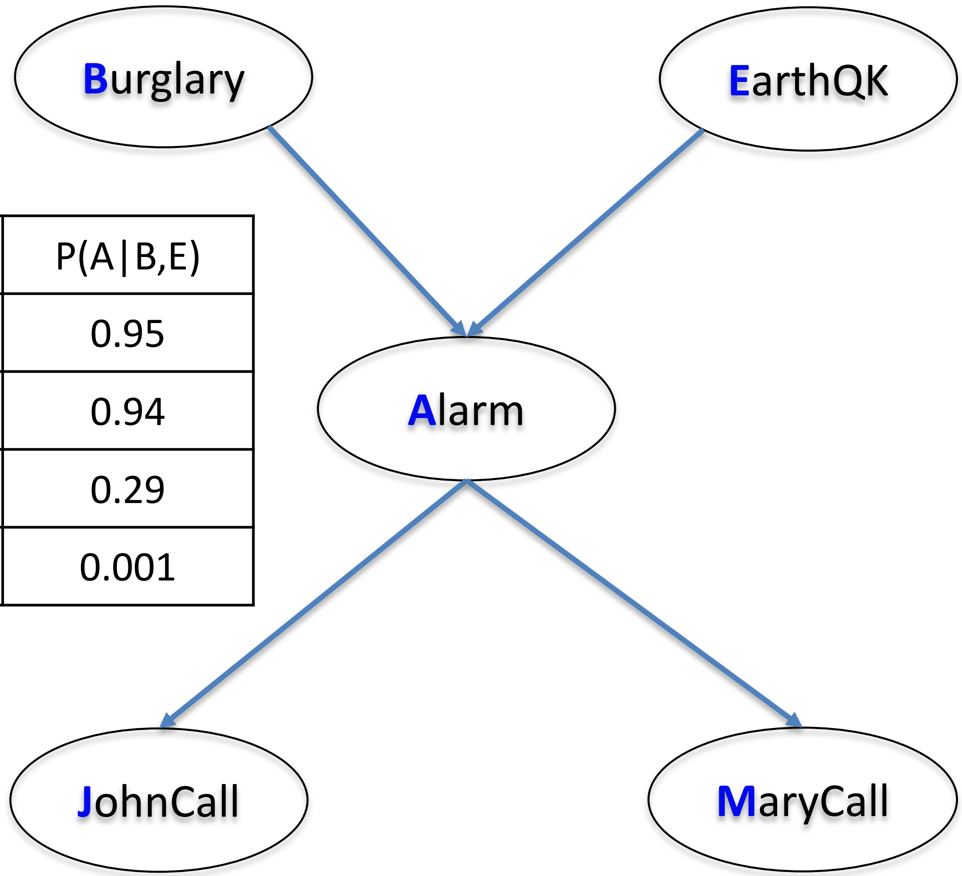
A	M	P(M|A)
T	T	0.7
F	T	0.01

Number of free parameters

- Do we need to store P(B = F), P(A = F I B = T, E = T)?
- How many probabilities need to be stored?

B	P(B)
T	0.001

E	P(E)
T	0.002



B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001

A	J	P(J|A)
T	T	0.9
F	T	0.05

A	M	P(M|A)
T	T	0.7
F	T	0.01

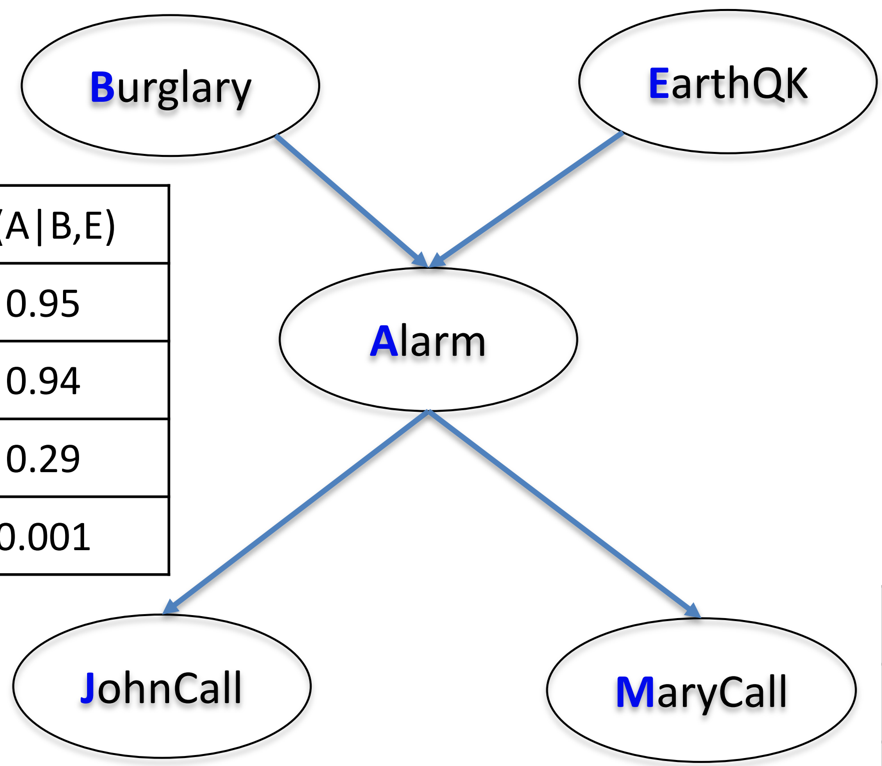
Number of free parameters

- Conditional Prob Table (CPT) size: no of classes minus 1
- Number of free parameters in a model is the number of variables/probabilities that cannot be derived, but has to be estimated
  - Number of free parameters in the alarm network?

B	P(B)
T	0.001

E	P(E)
T	0.002

B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001



A	J	P(J|A)
T	T	0.9
F	T	0.05

A	M	P(M|A)
T	T	0.7
F	T	0.01

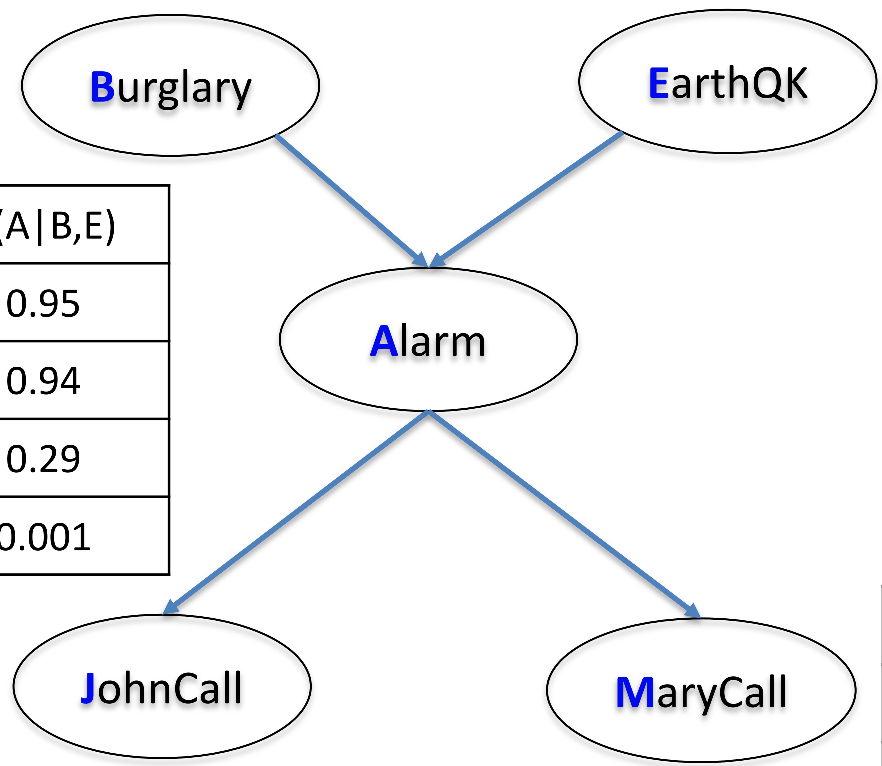
Number of free parameters

- Number of free parameters in the alarm network?

$$1+1+4+2+2=10$$

B	P(B)
T	0.001

E	P(E)
T	0.002



B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001

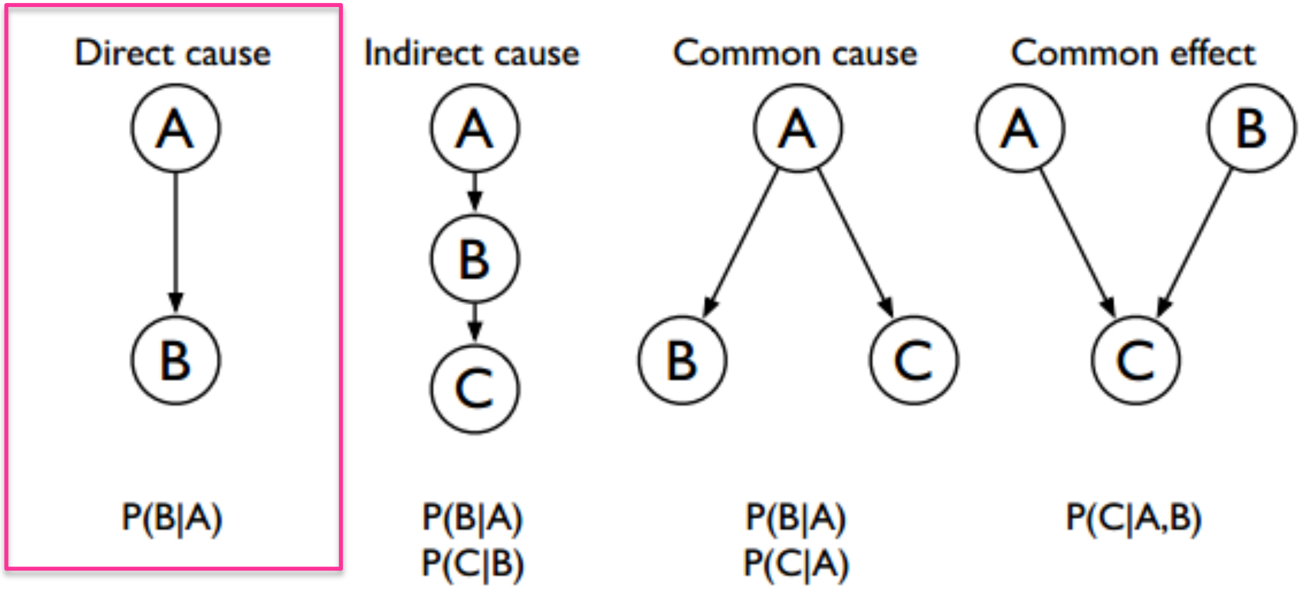
A	J	P(J|A)
T	T	0.9
F	T	0.05

A	M	P(M|A)
T	T	0.7
F	T	0.01

Number of free parameters

- Calculate the CPT size (number of free parameters) for the following
  - Assume: $|A| = 2, |B| = 2, |C| = 2$, they are all Boolean (binary) variables

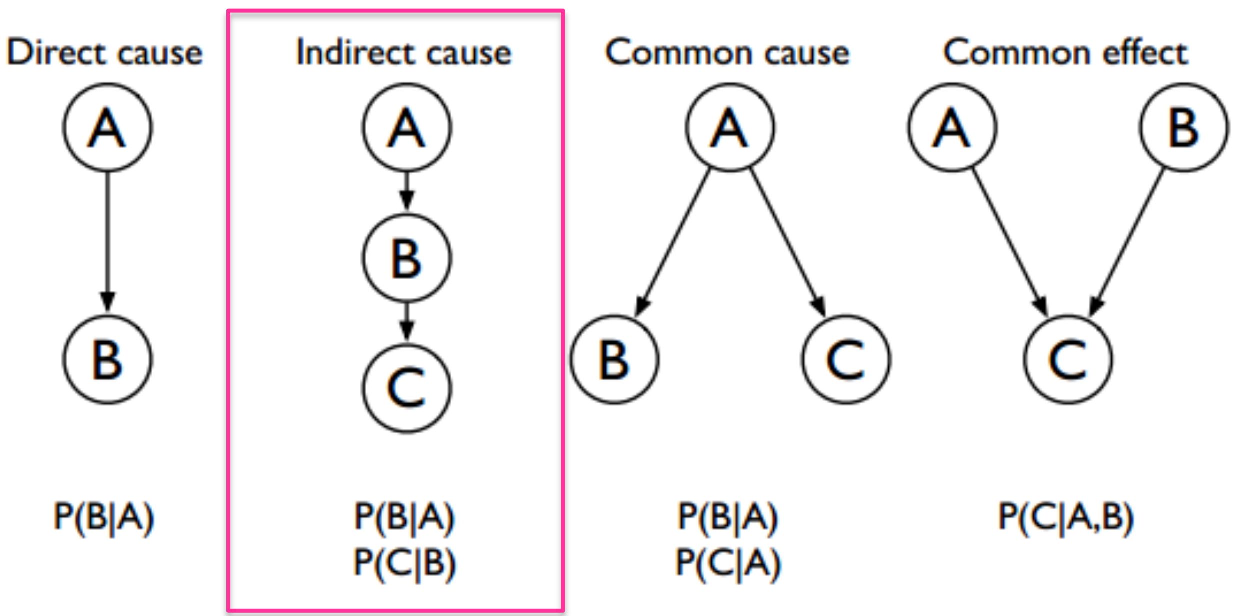
- Example: direct cause
  - $|A| - 1 + |A| \times (|B| - 1) = 2 - 1 + 2 \times 1 = 3$
- Other cases?



Number of free parameters

- Calculate the CPT size (number of free parameters) for the following
  - Assume: $|A| = 2, |B| = 2, |C| = 2$, they are all Boolean (binary) variables

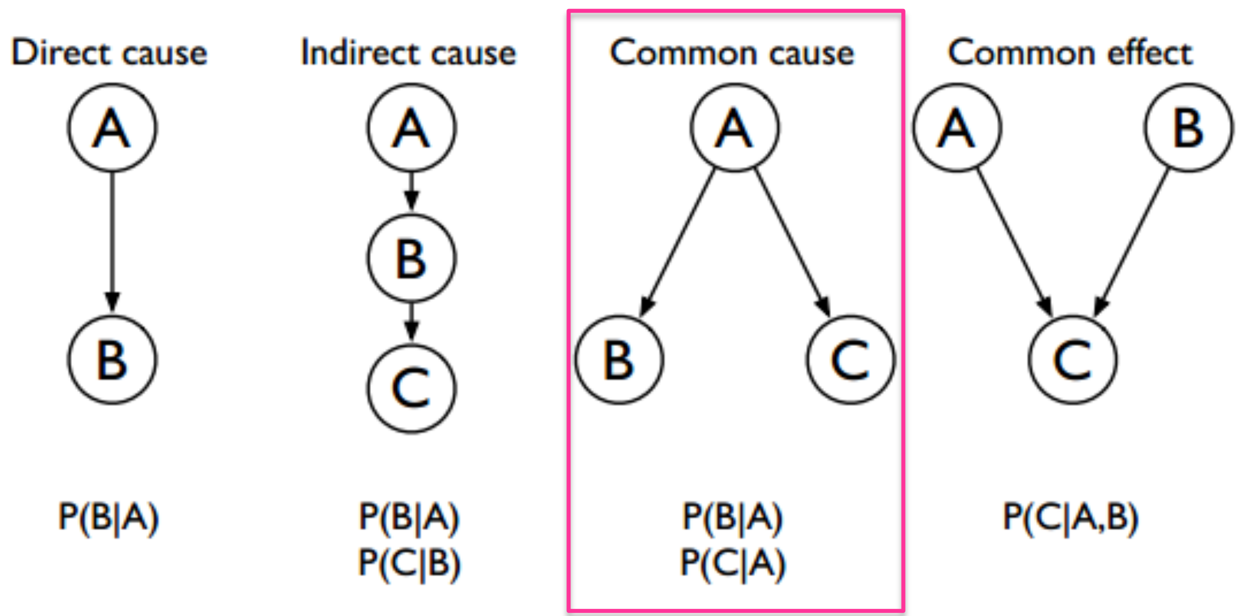
- Example: direct cause
  - $|A| - 1 + |A| \times (|B| - 1) = 2 - 1 + 2 \times 1 = 3$
- Other cases?
  - Indirect cause: $|A| - 1 + |A|(|B| - 1) + |B|(|C| - 1) = 2 - 1 + 2 \times 1 + 2 \times 1 = 5$



Number of free parameters

- Calculate the CPT size (number of free parameters) for the following
  - Assume: $|A| = 2, |B| = 2, |C| = 2$, they are all Boolean (binary) variables

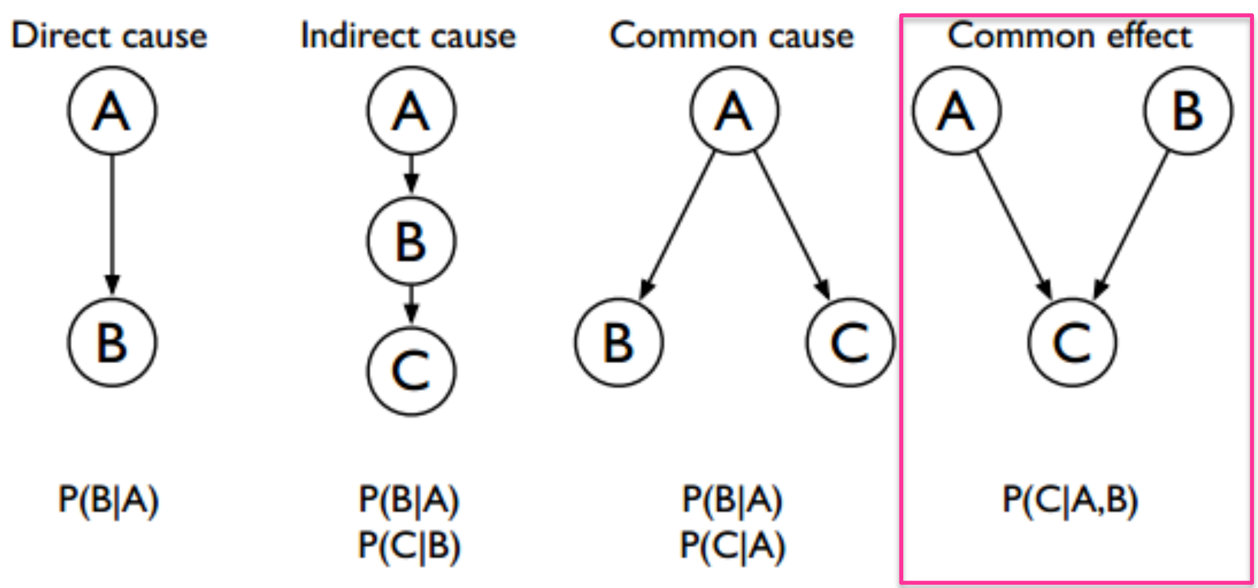
- Example: direct cause
  - $|A| - 1 + |A|\times(|B| - 1) = 2 - 1 + 2\times1 = 3$
- Other cases?
  - Indirect cause: $|A| - 1 + |A|(|B| - 1) + |B|(|C| - 1) = 2 - 1 + 2\times1 + 2\times1 = 5$
  - Common cause: $|A| - 1 + |A|(|B| - 1) + |A|(|C| - 1) = 2 - 1 + 2\times1 + 2\times1 = 5$



Number of free parameters

- Calculate the CPT size (number of free parameters) for the following
  - Assume: $|A| = 2, |B| = 2, |C| = 2$, they are all Boolean (binary) variables

- Example: direct cause
  - $|A| - 1 + |A| \times (|B| - 1) = 2 - 1 + 2 \times 1 = 3$
- Other cases?
  - Indirect cause: $|A| - 1 + |A|(|B| - 1) + |B|(|C| - 1) = 2 - 1 + 2 \times 1 + 2 \times 1 = 5$
  - Common cause: $|A| - 1 + |A|(|B| - 1) + |A|(|C| - 1) = 2 - 1 + 2 \times 1 + 2 \times 1 = 5$
  - Common effect: $|A| - 1 + |B| - 1 + |A||B|(|C| - 1) = 2 - 1 + 2 - 1 + 2 \times 2 \times 1 = 6$

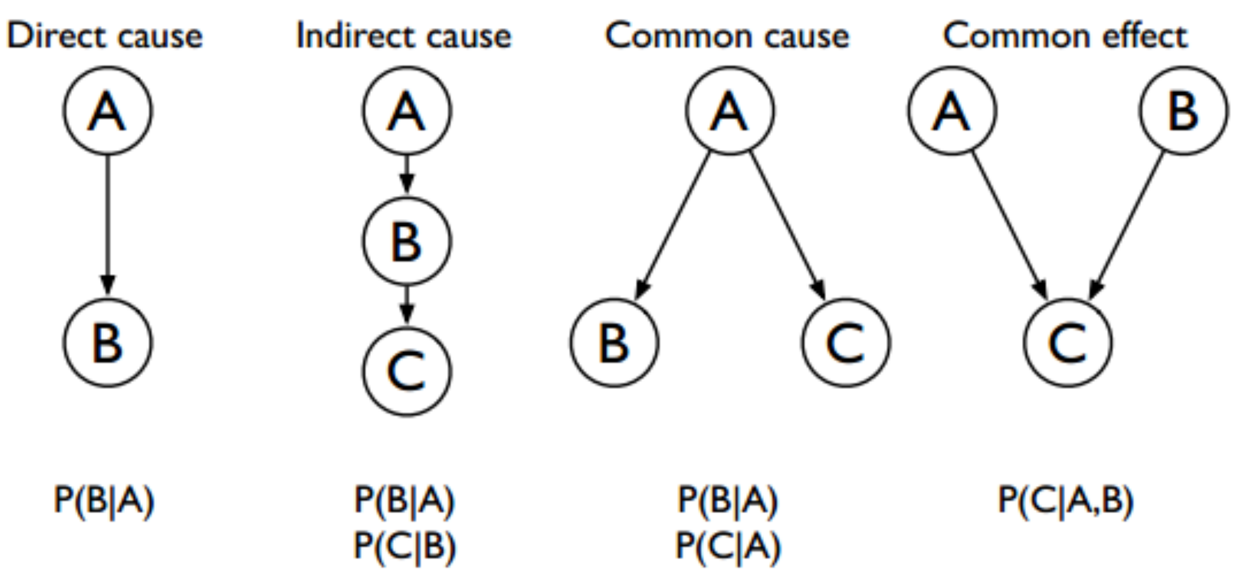


Number of Free Parameters

- Try calculate the CPT size (number of free parameters) for the following
  - Assume: $|A| = 2, |B| = 2, |C| = 2$, they are all Boolean (binary) variables

Note we are summing the free parameters for every variable

- Example: direct cause
  - $|A| - 1 + |A| \times (|B| - 1) = 2 - 1 + 2 \times 1 = 3$
- Other cases?
  - Indirect cause: $|A| - 1 + |A|(|B| - 1) + |B|(|C| - 1) = 2 - 1 + 2 \times 1 + 2 \times 1 = 5$
  - Common cause: $|A| - 1 + |A|(|B| - 1) + |A|(|C| - 1) = 2 - 1 + 2 \times 1 + 2 \times 1 = 5$
  - Common effect: $|A| - 1 + |B| - 1 + |A||B|(|C| - 1) = 2 - 1 + 2 - 1 + 2 \times 2 \times 1 = 6$



Number of free parameters

- In general, for a Bayesian network with factorization

$$P(X_1, \ldots, X_n) = P(X_1|\text{parents}(X_1)) * \cdots * P(X_n|\text{parents}(X_n))$$

- The number of free parameters of X_i is

Number of probs estimated for each condition

$$\boxed{(|X_i| - 1)} * \boxed{\prod_{Y \in \text{parents}(X_i)} |Y|}$$

Number of conditions

- A Bayesian network with a small number of free parameters is desirable because it
  - Requires less memory
  - Is efficient to do reasoning (less variables involved for calculating posterior probabilities)

- Ideally, when building a Bayesian network, we should minimise the number of parents of each variable

Building a BN

1. Specify the **random variables**

Example: *"If it is raining, then students might not attend the lecture"*

*Variables **R**aining and **A**ttend*

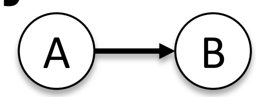
Building a BN

1. Specify the **random variables**

Example: *"If it is raining, then students might not attend the lecture"*

*Variables **R**aining and **A**ttend*

2. Specify the variables **dependencies and build DAG**

Example: 

Building a BN

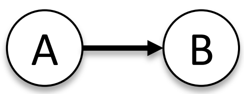
1. Specify the **random variables**

   Example: *"If it is raining, then students might not attend the lecture"*

   *Variables **R**aining and **A**ttend*

2. Specify the variables **dependencies and build DAG**

Example:



3. Assign conditional probabilities to each variable given its parents

Example: $P(R) = 0.7, P(A|R) = 0.6$

The conditional probabilities can be obtained from data, expert knowledge, or both

Building the DAG (Step 2)

- **Expert Knowledge**
  - Experts in the domain construct the network by specifying the causal relationships among variables
- **Constraint-Based Algorithms**
  - Identify conditional independence constraints with statistical tests, and link nodes that are not found to be independent
  - E.g., FCI (Fast Causal Inference)
- **Score-based Algorithms**
  - Applications of general optimisation techniques; each candidate DAG is assigned a network score maximise as the objective function
  - E.g., Tabu search, Simulated Annealing

Compactness and Node Ordering

- Compactness:
  - The more compact the BN model is, the smaller the CPT size
    - (CPT = conditional probability table)
  - Less computer memory, more computationally efficient
  - Over-dense networks fail to represent independencies explicitly
  - Over-dense networks fail to represent the causal dependencies in the domain

- The compactness depends on getting the node ordering "right." The optimal order is to add the root causes first, then the variable(s) they influence directly, and continue until leaves are reached

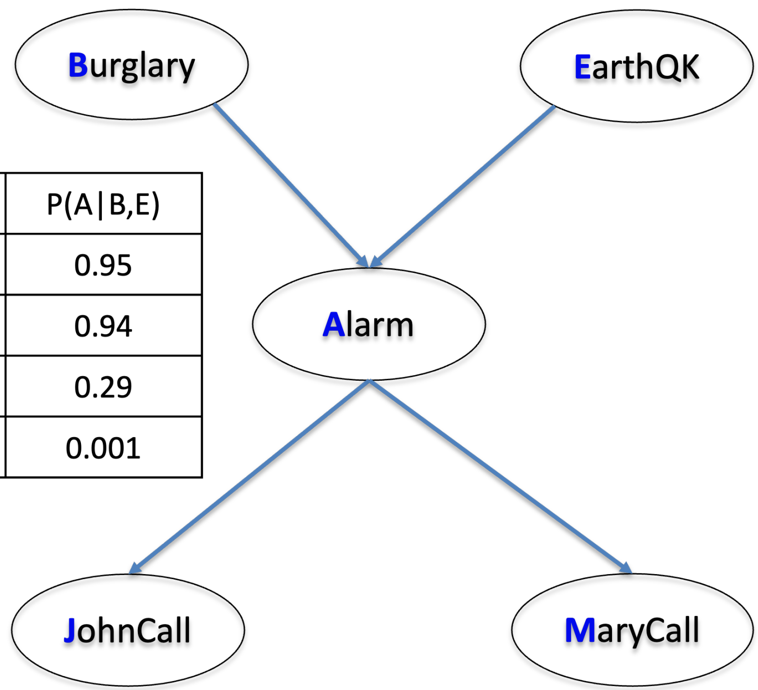
Inference in a BN

- If there was an earthquake, how likely Mary will call you?
- If both John and Mary called you, how likely there was a burglary?
- If Mary called you, how likely John will call you as well?

- Answering questions like above is **inference in a BN**

B	P(B)
T	0.001

E	P(E)
T	0.002



B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001

A	J	P(J|A)
T	T	0.9
F	T	0.05

A	M	P(M|A)
T	T	0.7
F	T	0.01

Recall basic inference

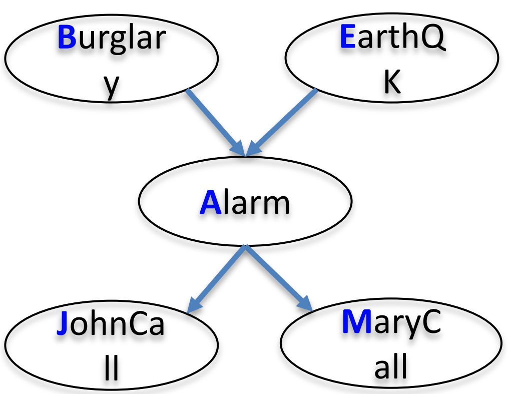
- Consider $P(M, J, A, E, B\ ) = \ P(M|A)P(J|A)(A|E, B)P(E)P(B)$
- Example: probability of burglary if Mary calls
  - Use law of total probability ("marginalise out" unknown variables)
  - Recall $P(J, A, E, B|M) = P(M, J, A, E, B)$ / $P(M)$
  - We compute: $P(B = 1|\ M = 1) = \sum_{a,j,e} P(M = 1, J = j, A = a, E = e, B = 1)/P(M1)$

  $\sum_{a,j,e} P(M = 1|A = a)P(J = j|A = a)P(A = a|E = e, B = 1)P(E = e)P(B = 1)/\ P(M = 1)$
- If we only want to know what is more probable, $P(B = 1|\ M = 1)$ vs $P(B = 0|\ M = 1)$, then we can omit the denominator:

  $P(B|\ M = 1) \propto \sum_{a,j,e} P(M = 1|A = a)P(J = j|A = a)(A = a|E = e, B)P(E = e)P(B)$
- Basic inference approach does *not* exploit network structure
- Very slow for large Bayes networks; faster but approximate, methods are useful

Inference in a BN

- (Probabilistic) inference: computing some useful quantity from the joint distribution
  - Posterior probability distribution of a variable given observation of a subset of other variables (evidence); $P(Q|E_1 = e_1, \cdots, E_k = e_k)$
  - Most probable explanation of a variable given observation of a subset of other variables (evidence); $q^* = argmax_q \, P(Q = q|E_1 = e_1, \cdots, E_k = e_k)$
- Inference in Bayesian networks is very flexible, as evidence can be entered for any node while beliefs in any other nodes can be computed
  - Causal Reasoning: P(Effect I Cause)
  - Diagnostic Reasoning: P(Cause I Effect)
  - Inter-causal Reasoning: the query nodes are common causes of the evidence nodes.



Inference in a BN

- The calculation of exact probabilities can be computationally expensive or even infeasible for very large networks

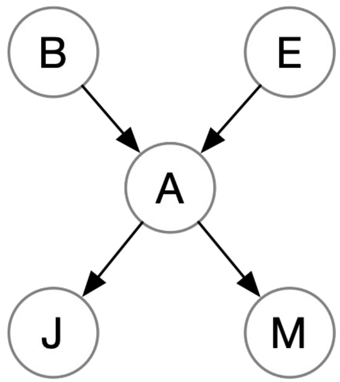
- Approximations or sampling methods are attractive but may introduce additional uncertainty

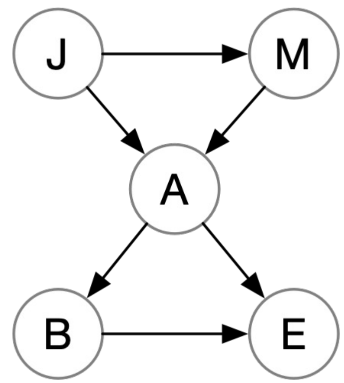
- **Exact algorithms:** "Brute Force approach" (the basic method) or "Inference by Enumeration", "Variable Elimination", …

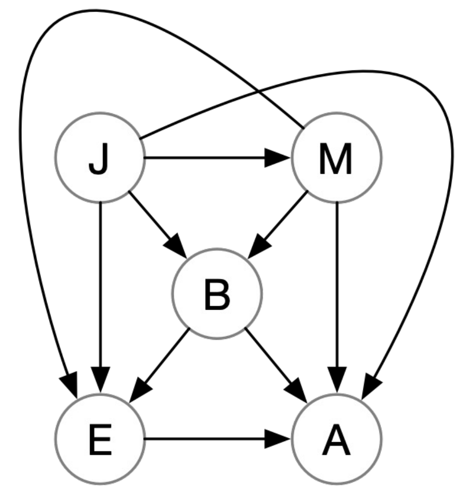
- **Non-exact algorithms:** Belief Propagation, Gibbs Sampling, …

- Discussed in AIML429

Ordering and Compactness







- Different algorithms can generate different BNs even if given the same **variables** and **CPTs**

- Do you think this influences the **inferences** using such BNs?

Summary

- Number of free parameters gives an estimation of complexity
- Building a BN is not trivial for large networks
- We can make inferences to learn about probabilities given some evidence

Coming up next…

- Tutorial
- Next week: Planning and Scheduling (Aaron)