COMP9418: Advanced Topics in Statistical Machine Learning

Bayesian Networks

Instructor: Gustavo Batista

University of New South Wales

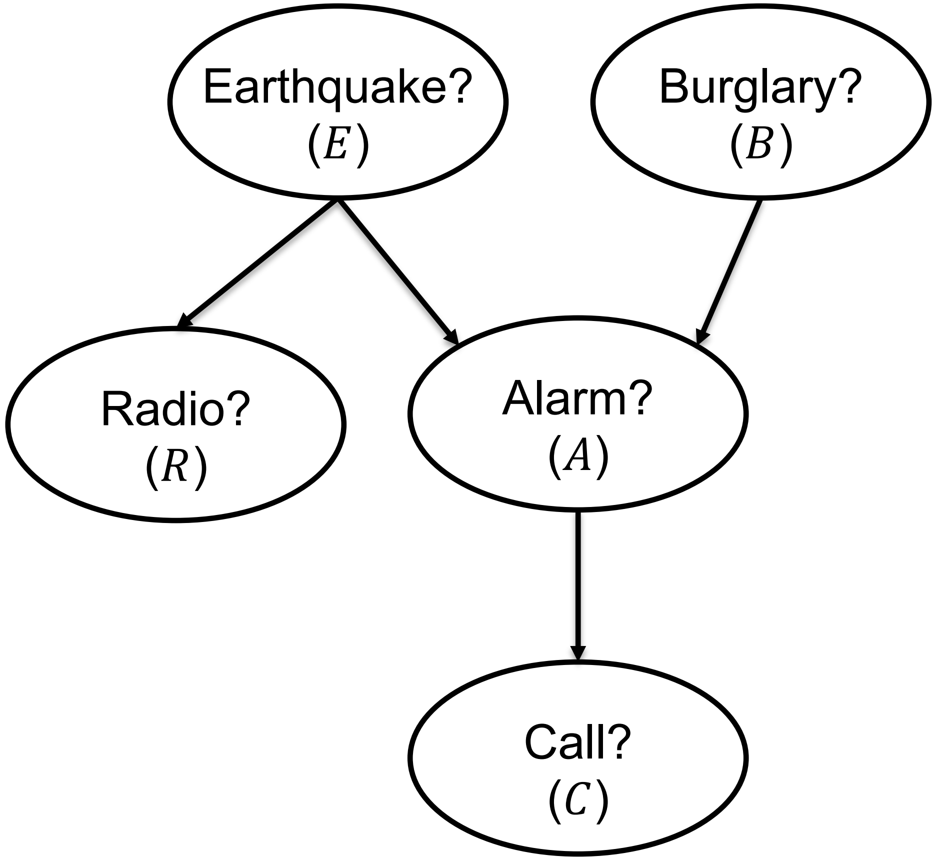
Introduction

- This lecture introduces Bayesian networks as a modelling tool to specify joint probability distributions
  - The size of a joint distribution is exponential in the number of variables
  - This causes modelling and computational difficulties
  - The specification of a joint distribution may hide some relevant properties such as independencies

- Bayesian networks is a graphical modelling tool for specifying probability distributions
  - It relies on the insight that independence is a significant aspect of beliefs, and
  - Independencies can be elicited using the language of graphs

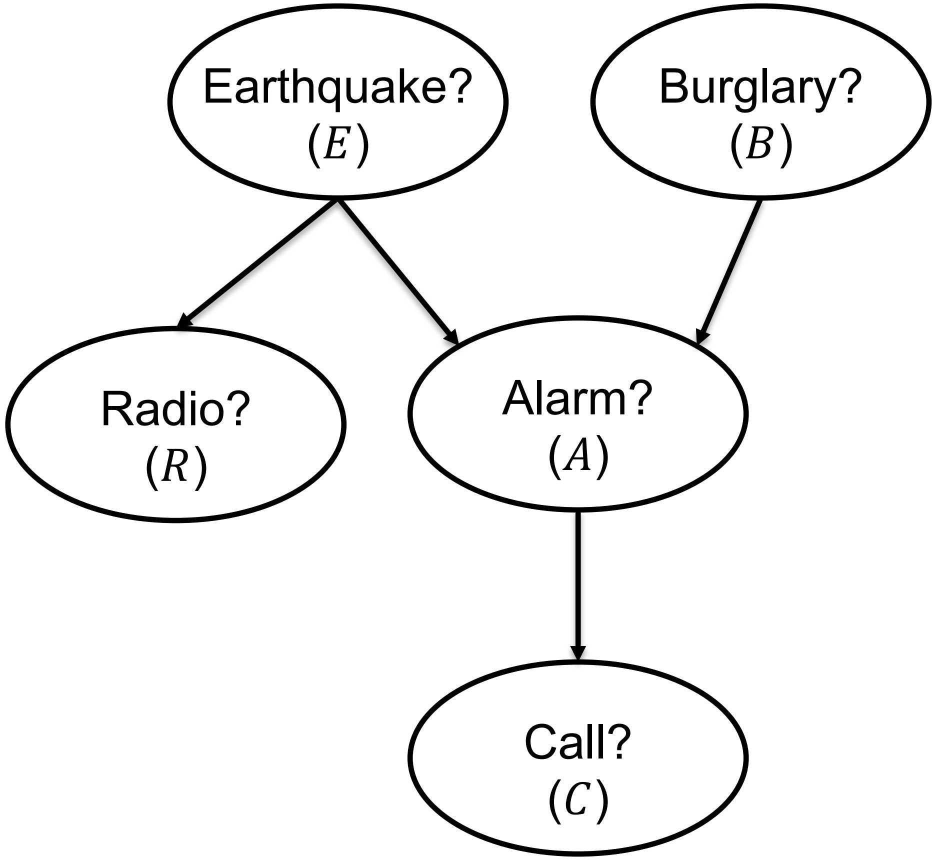
Graphs and Independence

- This figure is a *directed acyclic graph* (DAG)
  - Nodes represent variables
  - Let us assume (for now) that edges represent "direct causal influence"
  - For example, alarm triggering (A) causes a call for a neighbour (C)
- Given this representation, we expect the belief dynamic to satisfy some properties
  - For instance, C is influenced by evidence on R
  - A radio report would increase belief in Alarm. In turn, increase belief in a call from a neighbour
  - However, the belief in C would not increase if we knew the alarm did not trigger
  - $C \perp R | \neg A$



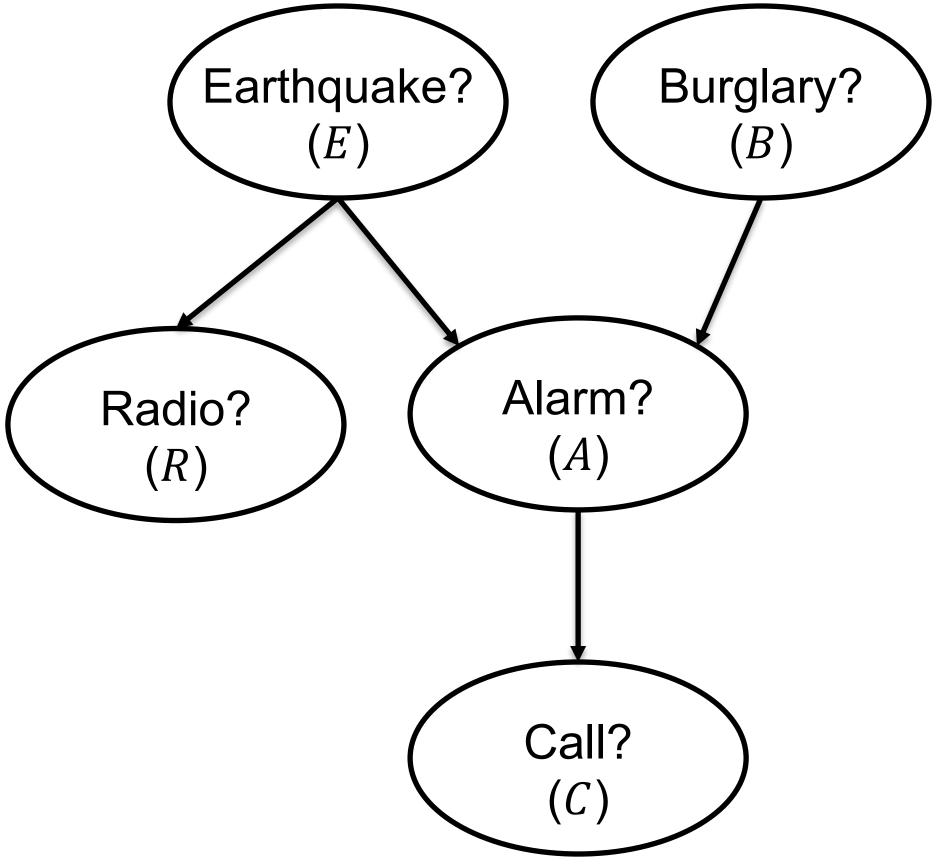
Notation

- Given a variable V in a DAG G
  - *Parents(V)* are the parents of V in DAG G, that is, the set of variables N with an edge from N to V
  - *Descendants(V)* are the descendants of V in G, that is, the set of variables N with a direct path from V to N
  - *Non_Descendants(V)* are all variables in G other than V, Parents(V), and Descendants(V)
- A DAG G is a compact representation of the following independence statements
  - $V \perp Non\_Descendants(V) \mid Parents(V)$
  - Every variable is conditionally independent of its nondescendants given its parents
  - *Markovian assumptions* of G denoted by *Markov(G)*



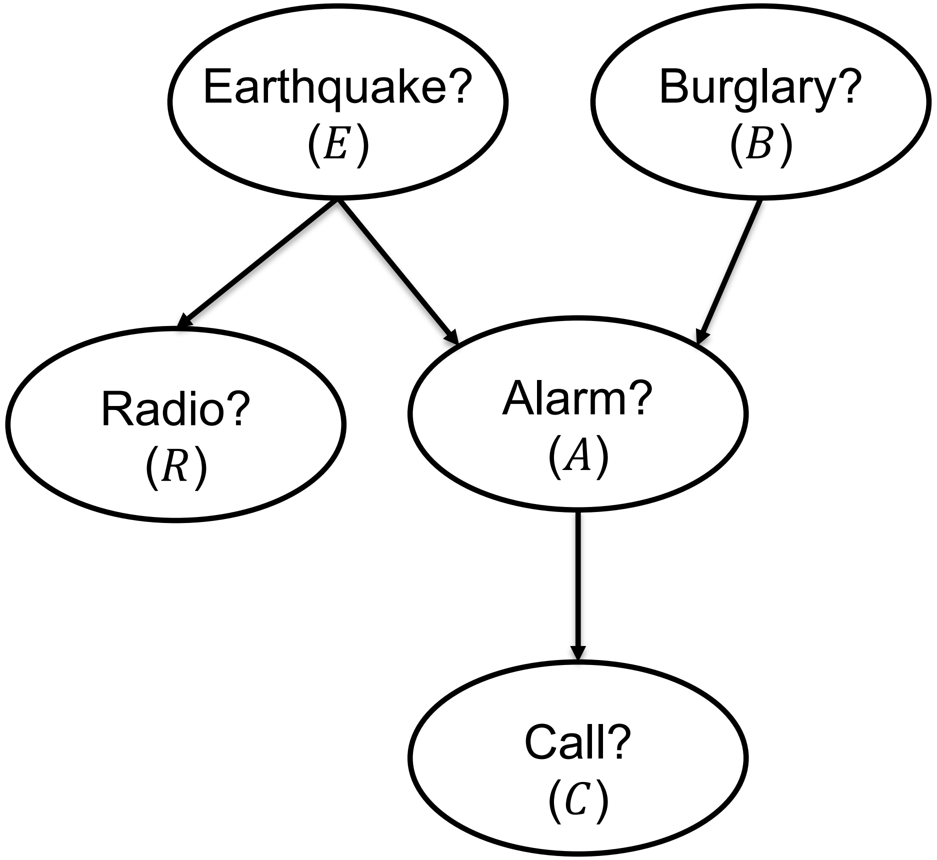
Markovian Assumptions

- If we view DAG G as a causal structure
  - Parents(V) are direct causes of V
  - Descendants(V) denotes the effects of V
- Given the direct causes of a variable, our beliefs in that variable will no longer be influenced by any other variable except possibly by its effects
- These are all the statements in this DAG



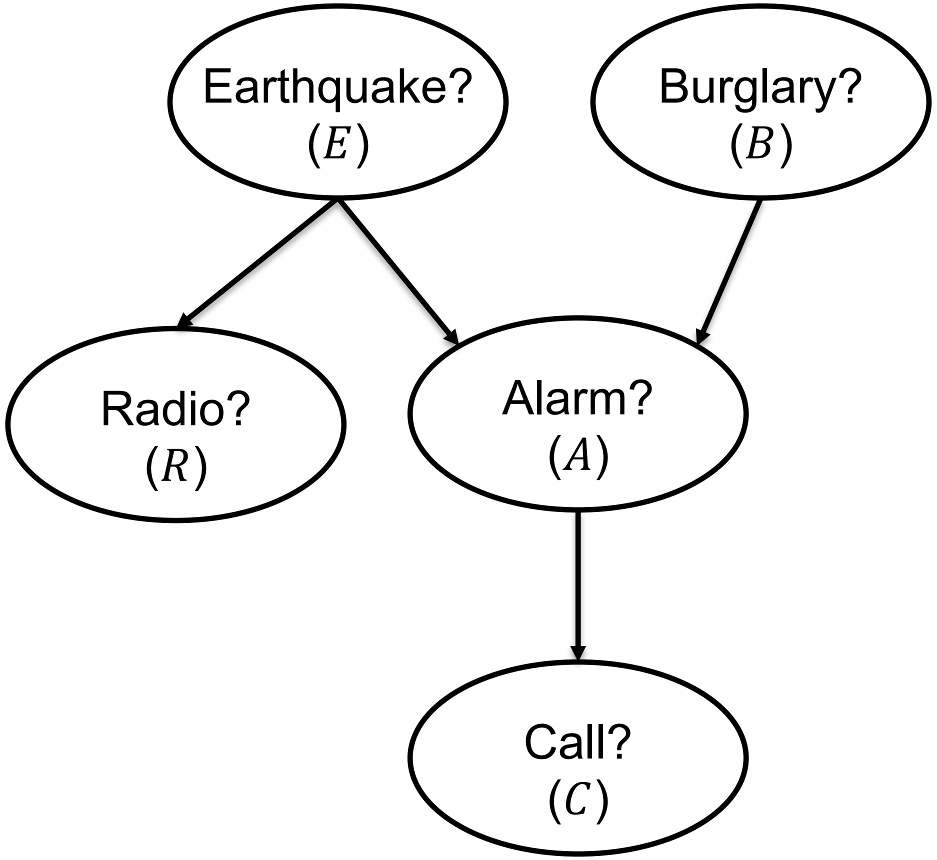
Markov Assumptions

- If we view DAG G as a causal structure
  - Parents(V) are direct causes of V
  - Descendants(V) denotes the effects of V
- Given the direct causes of a variable, our beliefs in that variable will no longer be influenced by any other variable except possibly by its effects
- These are all the statements in this DAG
  - $C \perp B, E, R \mid A$
  - $R \perp A, B, C \mid E$
  - $A \perp R \mid B, E$
  - $B \perp E, R$
  - $E \perp B$



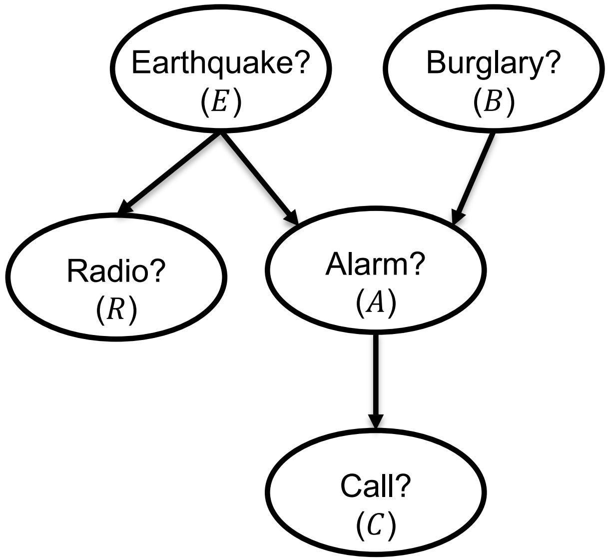
Markov Assumptions

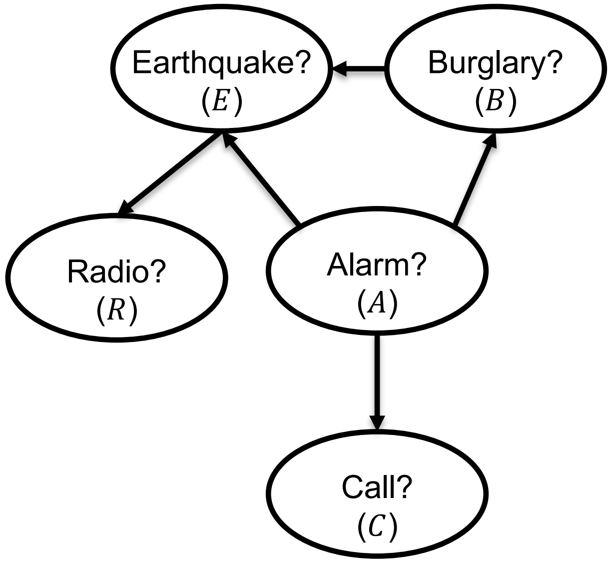
- Suppose we want to make a probability distribution that captures the state of belief
  - The first step is to construct the graph, ensuring the independences on G matches our beliefs
  - The DAG G is a partial specification. It says that P must satisfy Markov(G)
- The specification of G restricts the choices for the distribution P
  - However, it does not uniquely define it
  - We need to augment G with a set of conditional probabilities
  - The conditional probabilities and G are guaranteed to uniquely define the distribution P



Causality

- The formal interpretation of a DAG is a set of conditional independences
  - It makes no reference to causality
  - However, we used causality to motivate this interpretation
- It is perfectly possible to have a DAG that does not match our causal perception
  - We will see that every independence in the first graph is also present in the second
  - We discuss next the graph parametrization (quantifying dependencies between notes and parents)
  - This process is much easier to accomplish by an expert if the DAG corresponds to causal perceptions

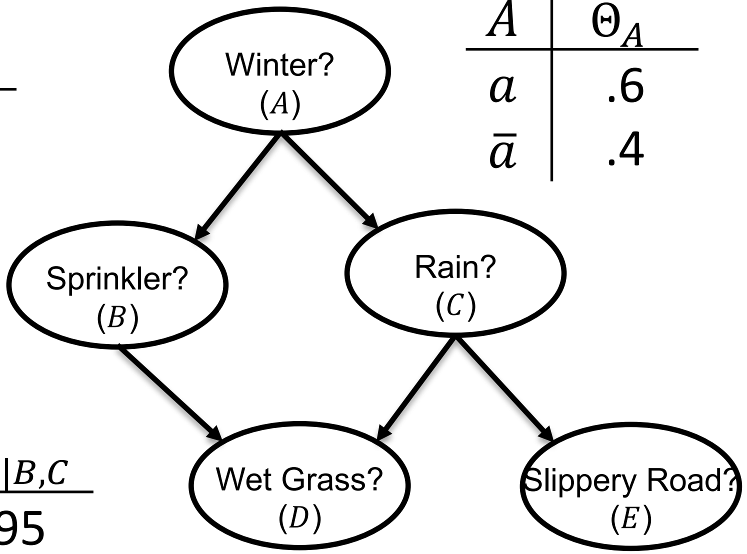




Parametrisation

- The conditional probabilities we need to specify are
  - For every variable X in DAG G and its parents $\boldsymbol{U}$
  - Provide the probabilities $P(x|\boldsymbol{u})$ for every value x of X and every instantiation $\boldsymbol{u}$ of parents $\boldsymbol{U}$
- For example, for this graph, we need to specify
  - $P(B|A), P(E|C), P(C|A), P(A), P(D|B,C)$
  - Each table is known as a *conditional probability table* (CPT)
  - Notice that $\sum_x P(x|\boldsymbol{u}) = 1$ for each $\boldsymbol{u} \in \boldsymbol{U}$
- Therefore, we only need 11 probabilities to specify the CPTs of this graph

A	B	$\Theta_{B\|A}$
a	b	.2
a	$\bar{b}$	.8
$\bar{a}$	b	.75
$\bar{a}$	$\bar{b}$	.25



A	Θ_A
a	.6
$\bar{a}$	.4

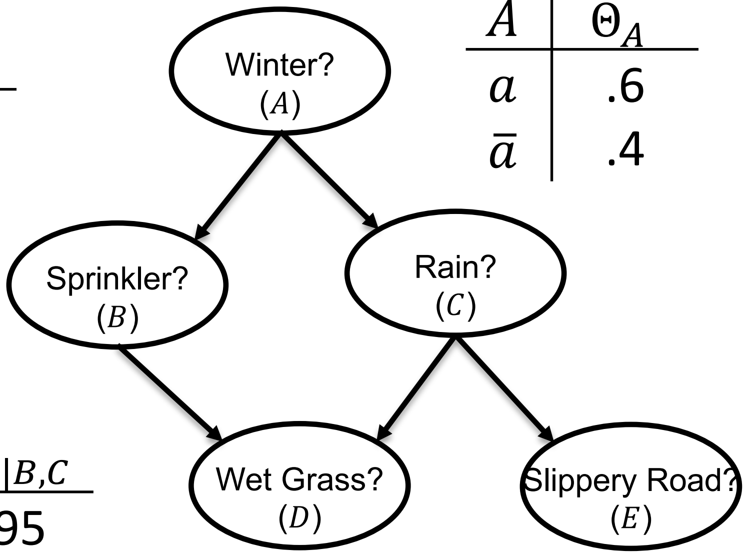
B	C	D	$\Theta_{D\|B,C}$
b	c	d	.95
b	c	$\bar{d}$	.05
b	$\bar{c}$	d	.9
b	$\bar{c}$	$\bar{d}$	.1
$\bar{b}$	c	d	.8
$\bar{b}$	c	$\bar{d}$	.2
$\bar{b}$	$\bar{c}$	d	0
$\bar{b}$	$\bar{c}$	$\bar{d}$	1

C	E	$\Theta_{E\|C}$
c	e	.7
c	$\bar{e}$	.3
$\bar{c}$	e	0
$\bar{c}$	$\bar{e}$	1

A	C	$\Theta_{C\|A}$
a	c	.8
a	$\bar{c}$	.2
$\bar{a}$	c	.1
$\bar{a}$	$\bar{c}$	.9

Bayesian Networks: Definition

- A Bayesian network for variables $\boldsymbol{Z}$ is a pair (G, Θ), where
  - G is a directed acyclic graph over variables $\boldsymbol{Z}$, called the *network structure*
  - Θ is a set of CPTs, one for each variable in $\boldsymbol{Z}$, called the *network parametrization*
- We use
  - $\Theta_{X|\boldsymbol{U}}$ to denote the CPT for variable X and its parents $\boldsymbol{U}$
  - $X\boldsymbol{U}$ to denote a set of variables known as *network family*
  - $\theta_{x|\boldsymbol{u}}$ is the value of $P(x|\boldsymbol{u})$ known as *network parameter*



A	Θ_A
a	.6
$\bar{a}$	.4

A	B	$\Theta_{B\|A}$
a	b	.2
a	$\bar{b}$	.8
$\bar{a}$	b	.75
$\bar{a}$	$\bar{b}$	.25

A	C	$\Theta_{C\|A}$
a	c	.8
a	$\bar{c}$	.2
$\bar{a}$	c	.1
$\bar{a}$	$\bar{c}$	.9

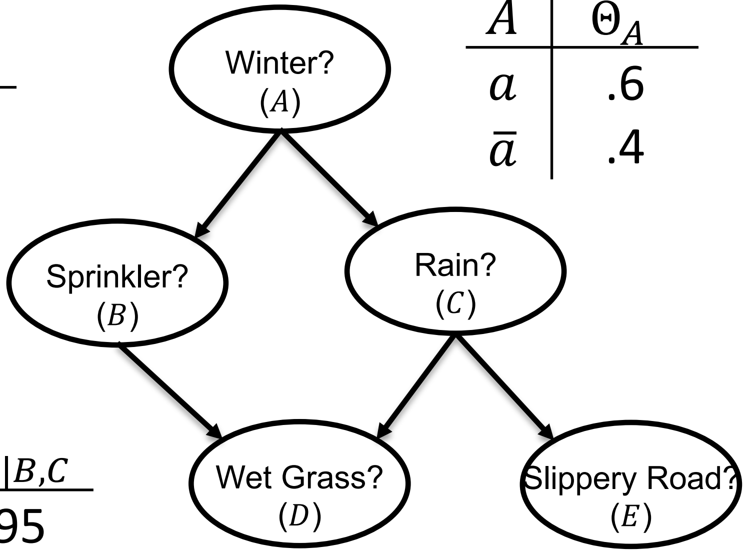
B	C	D	$\Theta_{D\|B,C}$
b	c	d	.95
b	c	$\bar{d}$	.05
b	$\bar{c}$	d	.9
b	$\bar{c}$	$\bar{d}$	.1
$\bar{b}$	c	d	.8
$\bar{b}$	c	$\bar{d}$	.2
$\bar{b}$	$\bar{c}$	d	0
$\bar{b}$	$\bar{c}$	$\bar{d}$	1

C	E	$\Theta_{E\|C}$
c	e	.7
c	$\bar{e}$	.3
$\bar{c}$	e	0
$\bar{c}$	$\bar{e}$	1

Bayesian Networks : More Definition

- *Network instantiation* is an assignment of all network variables
  - A network parameter $\theta_{x|\mathbf{u}}$ is compatible with a network instantiation $\mathbf{z}$ when $x\mathbf{u}$ and $\mathbf{z}$ agree on common variables
  - We write $\theta_{x|\mathbf{u}} \sim \mathbf{z}$
  - For instance, θ_a, $\theta_{b|a}$, $\theta_{\bar{c}|a}$, $\theta_{d|b,\bar{c}}$, and $\theta_{\bar{e}|\bar{c}}$ are parameters compatible with the instantiation $a, b, \bar{c}, d, \bar{e}$.

A	B	$\Theta_{B\|A}$
a	b	.2
a	$\bar{b}$	.8
$\bar{a}$	b	.75
$\bar{a}$	$\bar{b}$	.25



A	Θ_A
a	.6
$\bar{a}$	.4

B	C	D	$\Theta_{D\|B,C}$
b	c	d	.95
b	c	$\bar{d}$	.05
b	$\bar{c}$	d	.9
b	$\bar{c}$	$\bar{d}$	.1
$\bar{b}$	c	d	.8
$\bar{b}$	c	$\bar{d}$	.2
$\bar{b}$	$\bar{c}$	d	0
$\bar{b}$	$\bar{c}$	$\bar{d}$	1

C	E	$\Theta_{E\|C}$
c	e	.7
c	$\bar{e}$	.3
$\bar{c}$	e	0
$\bar{c}$	$\bar{e}$	1

A	C	$\Theta_{C\|A}$
a	c	.8
a	$\bar{c}$	.2
$\bar{a}$	c	.1
$\bar{a}$	$\bar{c}$	.9

Bayesian Networks: More Definition

- Only one probability distribution satisfies the constrains imposed by a Bayesian network
  - The distribution is given by

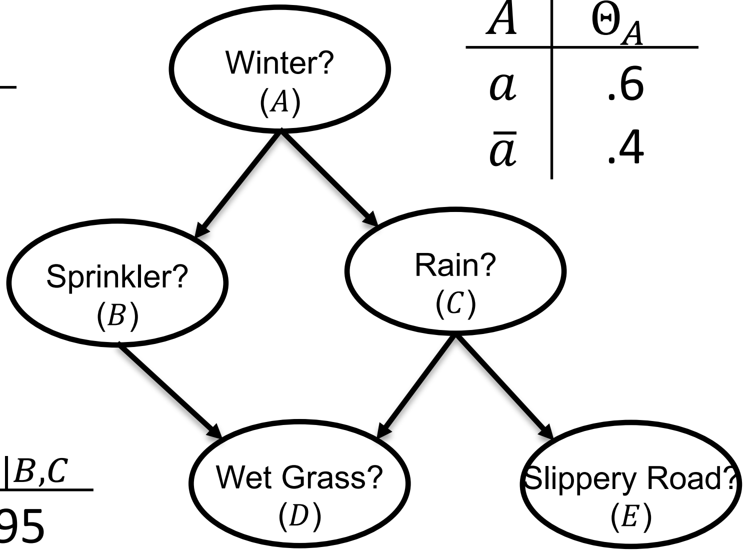
$$P(\mathbf{z}) \stackrel{\text{def}}{=} \prod_{\Theta_{x|\mathbf{u}\sim\mathbf{z}}} \theta_{x|\mathbf{u}}$$

  - This equation is known as the *chain rule* for Bayesian networks
- For instance,
  - $P(a,b,\bar{c},d,\bar{e}) = \theta_a \theta_{b|a} \theta_{\bar{c}|a} \theta_{d|b,\bar{c}} \theta_{\bar{e}|\bar{c}}$

$$= (.6)(.2)(.2)(.9)(1)$$

$$= .0216$$

A	B	$\Theta_{B\|A}$
a	b	.2
a	$\bar{b}$	.8
$\bar{a}$	b	.75
$\bar{a}$	$\bar{b}$	.25



A	Θ_A
a	.6
$\bar{a}$	.4

B	C	D	$\Theta_{D\|B,C}$
b	c	d	.95
b	c	$\bar{d}$	.05
b	$\bar{c}$	d	.9
b	$\bar{c}$	$\bar{d}$	.1
$\bar{b}$	c	d	.8
$\bar{b}$	c	$\bar{d}$	.2
$\bar{b}$	$\bar{c}$	d	0
$\bar{b}$	$\bar{c}$	$\bar{d}$	1

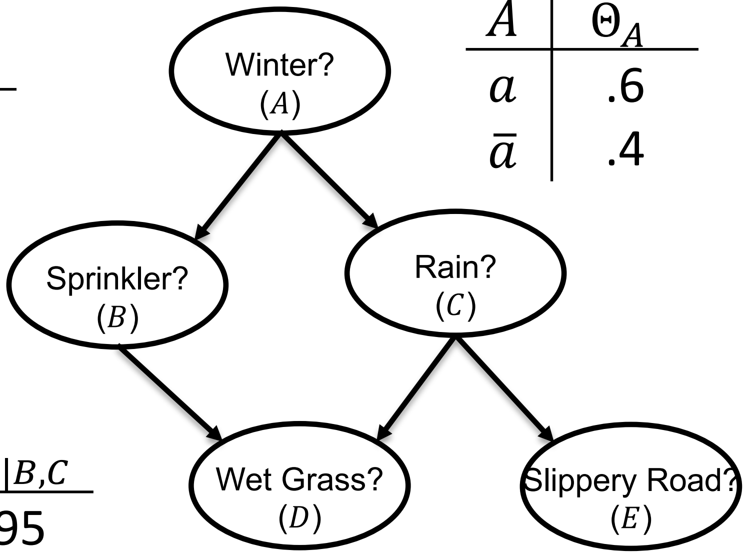
C	E	$\Theta_{E\|C}$
c	e	.7
c	$\bar{e}$	.3
$\bar{c}$	e	0
$\bar{c}$	$\bar{e}$	1

A	C	$\Theta_{C\|A}$
a	c	.8
a	$\bar{c}$	.2
$\bar{a}$	c	.1
$\bar{a}$	$\bar{c}$	.9

Bayesian Networks: Complexity

- The size of the CPT $\Theta_{X|\boldsymbol{U}}$ is exponential in the number of parents $\boldsymbol{U}$
  - If the maximal number of parents for every variable is k then the size of any CPT is $O(d^{k+1})$, where d is the number of values
  - With n network variables, the total number of variables is bounded by $O(nd^{k+1})$
- This number is reasonable if the number of parents is small
  - We will discuss techniques to represent CPTs when the number of parents is large

A	B	$\Theta_{B\|A}$
a	b	.2
a	$\bar{b}$	.8
$\bar{a}$	b	.75
$\bar{a}$	$\bar{b}$	.25



A	Θ_A
a	.6
$\bar{a}$	.4

B	C	D	$\Theta_{D\|B,C}$
b	c	d	.95
b	c	$\bar{d}$	.05
b	$\bar{c}$	d	.9
b	$\bar{c}$	$\bar{d}$	.1
$\bar{b}$	c	d	.8
$\bar{b}$	c	$\bar{d}$	.2
$\bar{b}$	$\bar{c}$	d	0
$\bar{b}$	$\bar{c}$	$\bar{d}$	1

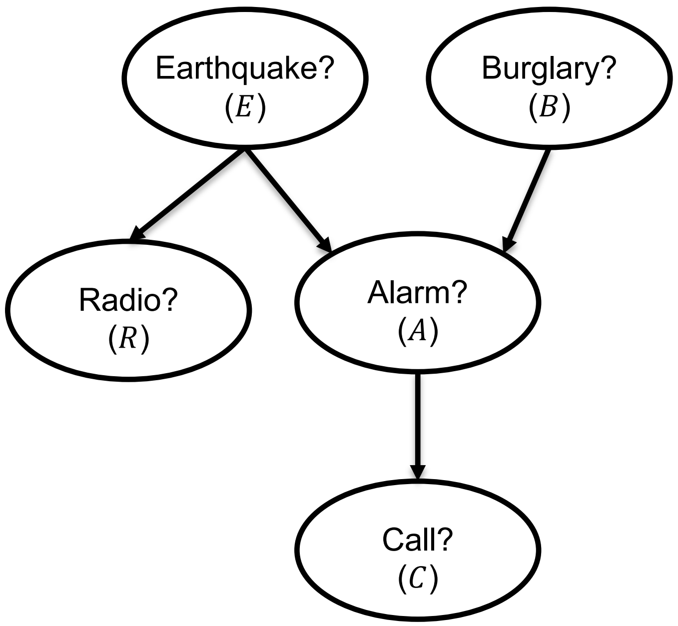
C	E	$\Theta_{E\|C}$
c	e	.7
c	$\bar{e}$	.3
$\bar{c}$	e	0
$\bar{c}$	$\bar{e}$	1

A	C	$\Theta_{C\|A}$
a	c	.8
a	$\bar{c}$	.2
$\bar{a}$	c	.1
$\bar{a}$	$\bar{c}$	.9

Properties of Independence

- The distribution P specified by a Bayesian network (G, Θ) satisfies the independence assumptions in Markov(G)
  - However, these are not the only independences satisfied by P
  - For example, $R \perp A \mid E$
- This independence and other ones follow the ones in Markov(G)
  - If we use a set of properties known as *graphoid axioms*
  - These properties include symmetry, decomposition, weak union and contraction

$$X \perp Non\_Descendants(X) \mid Parents(X)$$

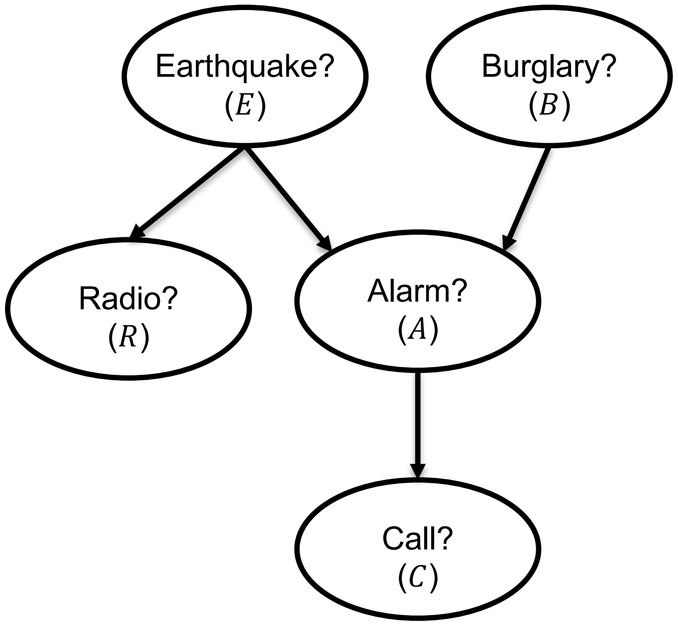


Properties of Independence: Symmetry

- *Symmetry* is the simplest property of probabilistic independence
  - If learning $\boldsymbol{y}$ does not influence our belief in $\boldsymbol{x}$, then learning $\boldsymbol{x}$ does not influence our belief in $\boldsymbol{y}$.

$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ if and only if $\boldsymbol{Y} \perp \boldsymbol{X} \mid \boldsymbol{Z}$

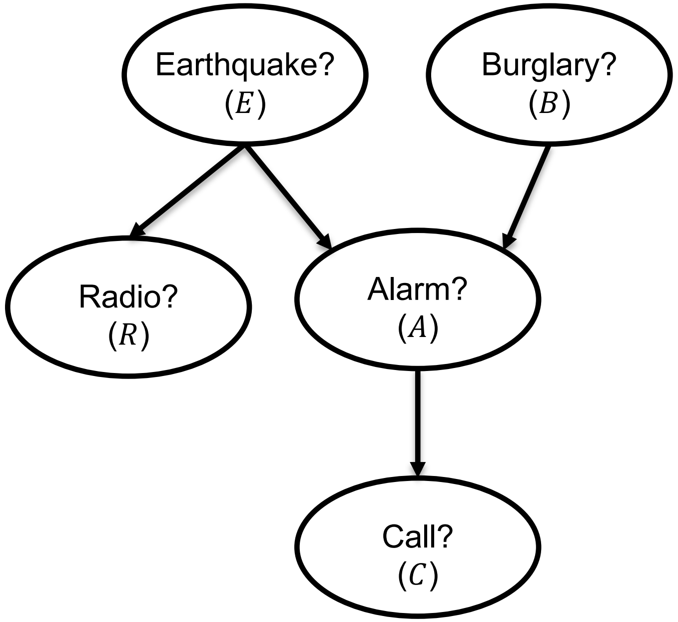
- In the example graph
  - $A \perp R \mid B, E$ (Markovian property for A)
  - $R \perp A \mid B, E$ (using symmetry)



Properties of Independence: Decomposition

- The second property is *decomposition*
  - If learning $\boldsymbol{yw}$ does not influence our belief in $\boldsymbol{x}$, then learning $\boldsymbol{y}$ alone, or learning $\boldsymbol{w}$ alone, does not influence our belief in $\boldsymbol{y}$.
- In the example graph
  - $R \perp A, C, B \mid E$ (Markovian property for A)
  - $R \perp A \mid E$ (using decomposition)
  - $R \perp C \mid E$ (using decomposition)
  - $R \perp B \mid E$ (using decomposition)
- Decomposition allow us to state the following
  - $X \perp \boldsymbol{W}$ for every $\boldsymbol{W} \subseteq$ Non_Descendants(X)
  - Notice $\boldsymbol{W}$ can be any subset of Non_descendants(X)

$\boldsymbol{X} \perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$ only if
$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ and $\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}$

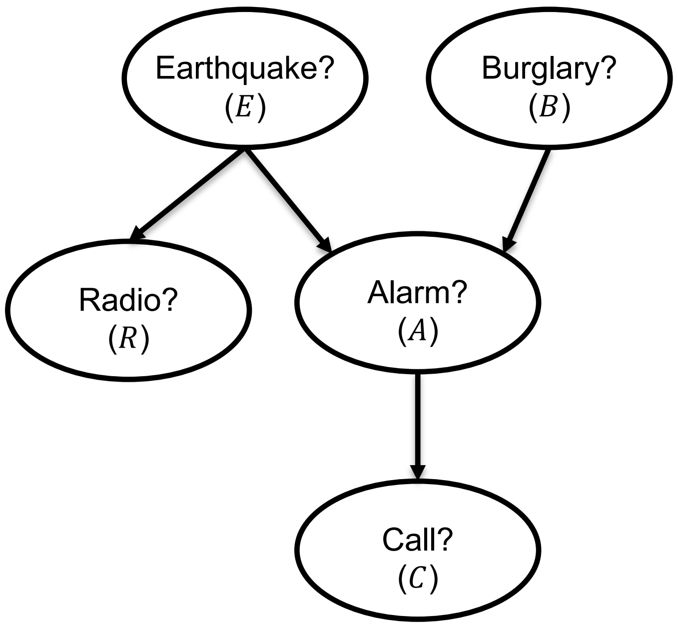


Properties of Independence: Decomposition

- Decomposition allow us to prove the chain rule for Bayesian networks

$$P(\mathbf{z}) \stackrel{\text{def}}{=} \prod_{\Theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}}$$

- For this example network we have
  - $P(e,b,r,a,c) = \theta_e \theta_b \theta_{r|e} \theta_{a|e,b} \theta_{c|a}$
  - $P(e,b,r,a,c) = P(e)P(b)P(r|e)P(a|e,b)P(c|a)$

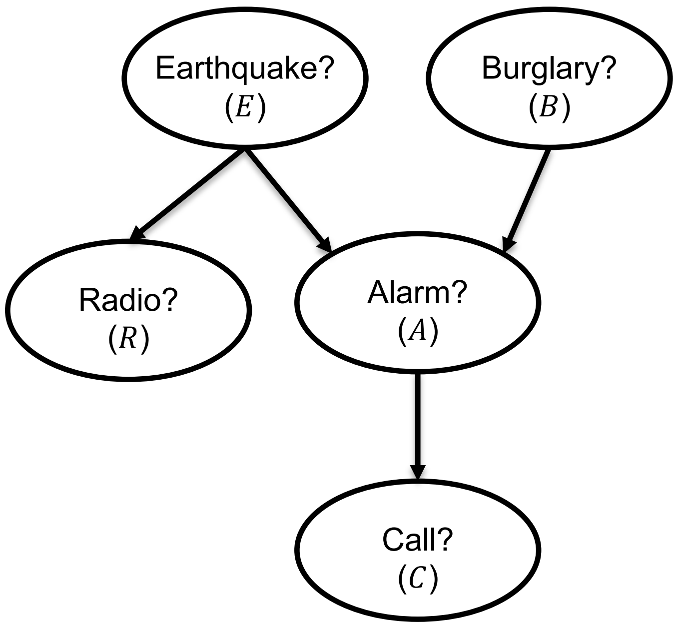


- By the chain rule
  - $P(e,b,r,a,c) = P(e)P(b|e)P(r|b,e)P(a|e,b,r)P(c|a,e,b,r)$

Properties of Independence: Weak Union

- The next property is *weak union*
  - If the information $\boldsymbol{yw}$ is not relevant to our belief in $\boldsymbol{x}$, then the partial information $\boldsymbol{y}$ will not make the rest of the information, $\boldsymbol{w}$, relevant
- In the example graph
  - $C \perp B, E, R \mid A$ (Markovian property for A)
  - $C \perp R \mid A, B, E$ (using decomposition)
- Decomposition allow us to state the following
  - $X \perp \mathrm{Non\_Descendants}(X) \setminus \boldsymbol{W} \mid Parents(X) \cup \boldsymbol{W}$ for every $\boldsymbol{W} \subseteq$ Non_Descendants(X)
  - This can be viewed as strengthening of the independences declared by Markov(G)

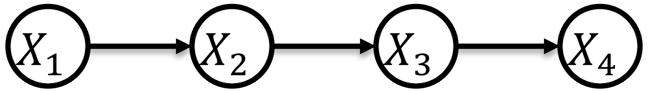
$\boldsymbol{X} \perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$ only if $\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z} \cup \boldsymbol{Y}$



Properties of Independence: Contraction

- The fourth property is *contraction*
  - If after learning the irrelevant information $\boldsymbol{y}$ the information $\boldsymbol{w}$ is found to be irrelevant to our belief in $\boldsymbol{x}$, then the combined information $\boldsymbol{yw}$ must have been irrelevant from the beginning

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \text{ and } \boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z} \cup \boldsymbol{Y} \text{ only if } \boldsymbol{X} \perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$$

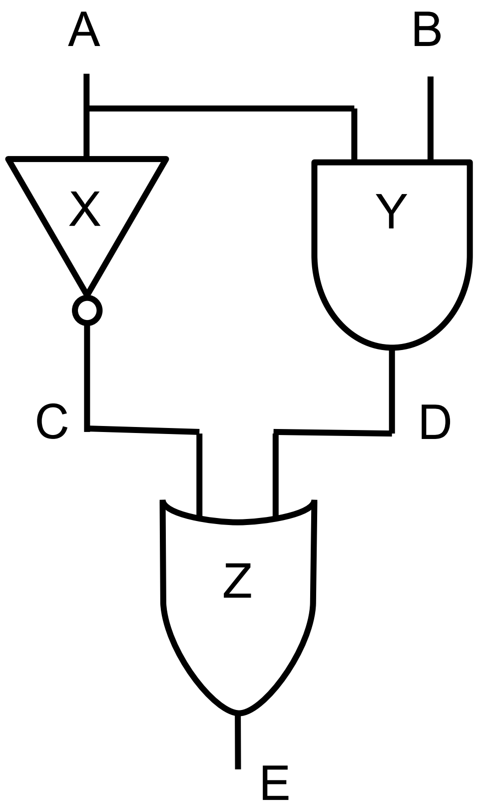


Properties of Independence: Intersection

- The final axiom is *intersection*
  - It holds only for the class strictly positive distributions
  - If information $\boldsymbol{w}$ is irrelevant given $\boldsymbol{y}$ and information $\boldsymbol{y}$ is irrelevant given $\boldsymbol{w}$, then the combined information $\boldsymbol{yw}$ is irrelevant to start with

- Symmetry, decomposition, weak union and contraction
  - Plus the property of triviality $(X \perp \emptyset \mid Z)$
  - Form the *graphoid axioms*
  - Plus intersection, the set is known as *positive graphoid axioms*

$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \cup \boldsymbol{W}$ and $\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z} \cup \boldsymbol{Y}$ only if $\boldsymbol{X} \perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$

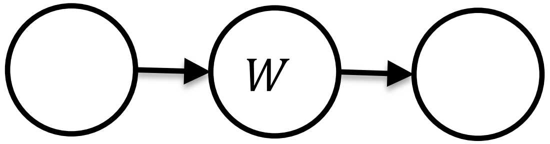


Graphical Test of Independence

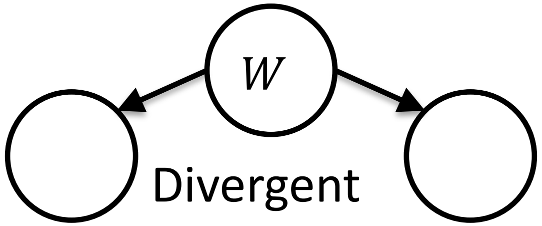
- P is a distribution induced by the Bayesian network (G, Θ)
  - P satisfies independences that go beyond what is in Markov(G)
  - Graphoid axioms derive new independences
  - However, this derivation is not trivial
- A graphical test known as *d-separation* can capture the inferential power of graphoid axioms
  - Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three disjoint sets of variables
  - $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in DAG G, if every path between a node in $\boldsymbol{X}$ and a node in $\boldsymbol{Y}$ is blocked by $\boldsymbol{Z}$
  - If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ then $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ for every probability distribution induced by G

Graphical Test of Independence: Blocking

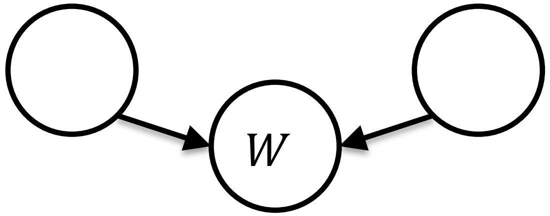
- Consider this path (note that it ignores the edges direction)
  - We will view this path as a pipe and each variable W on the path as a valve
  - A valve W is either open or closed, depending on some condition
  - If at least one of the valves on the path is closed, then the whole path is blocked
  - Otherwise the path is not blocked
- There are three types of valves
  - They are determined by its relationship to its neighbours on the path



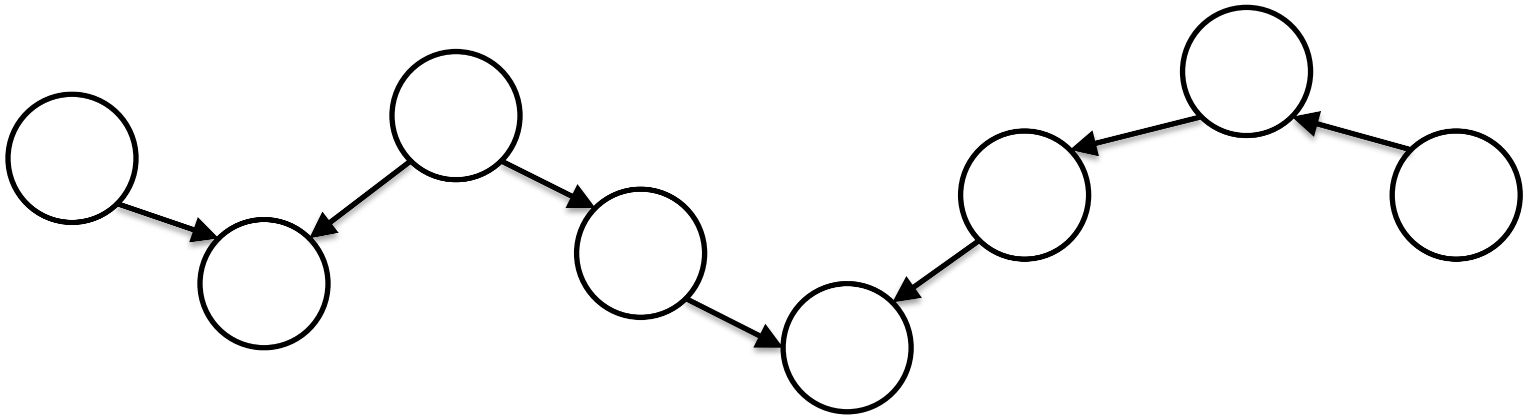
Sequential



Divergent

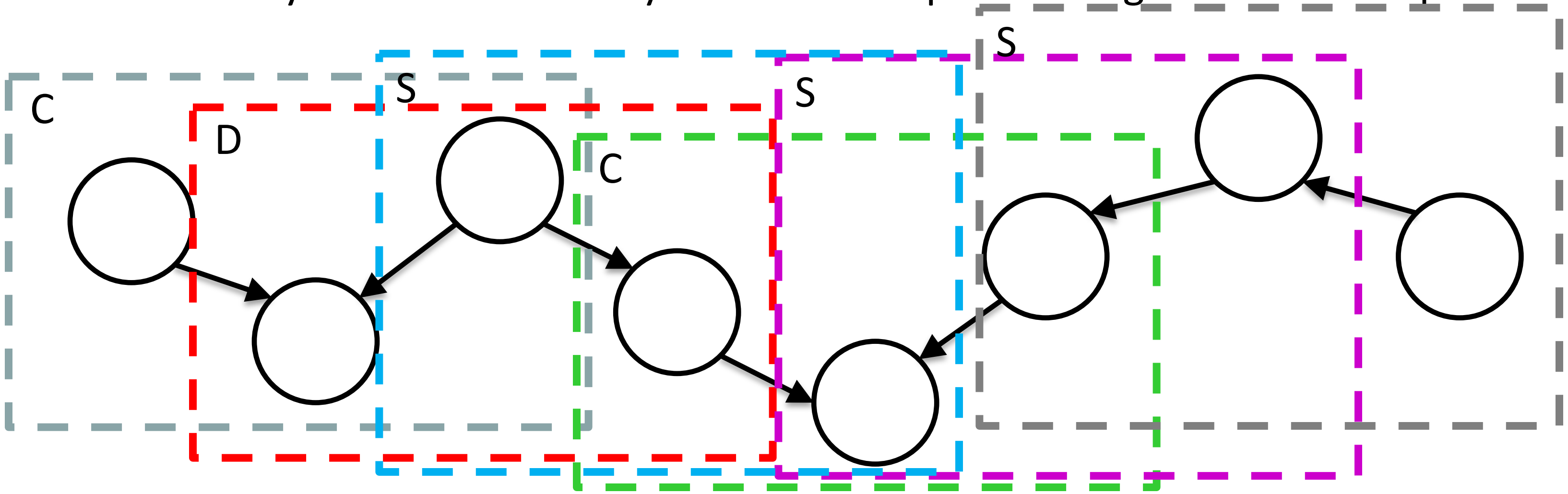


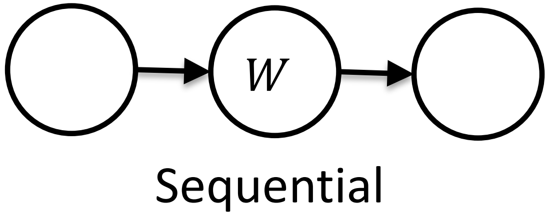
Convergent

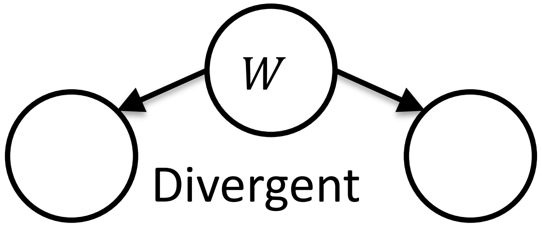


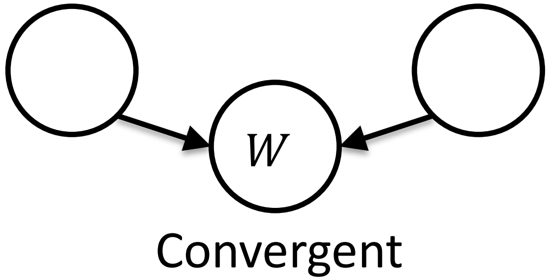
Graphical Test of Independence: Blocking

- Consider this path (note that it ignores the edges direction)
  - We will view this path as a pipe and each variable W on the path as a valve
  - A valve W is either open or closed, depending on some condition
  - If at least one of the valves on the path is closed, then the whole path is blocked
  - Otherwise the path is not blocked
- There are three types of valves
  - They are determined by its relationship to its neighbours on the path



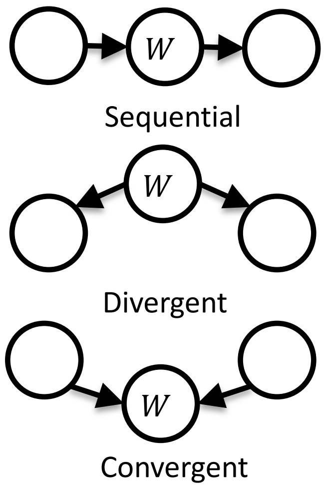


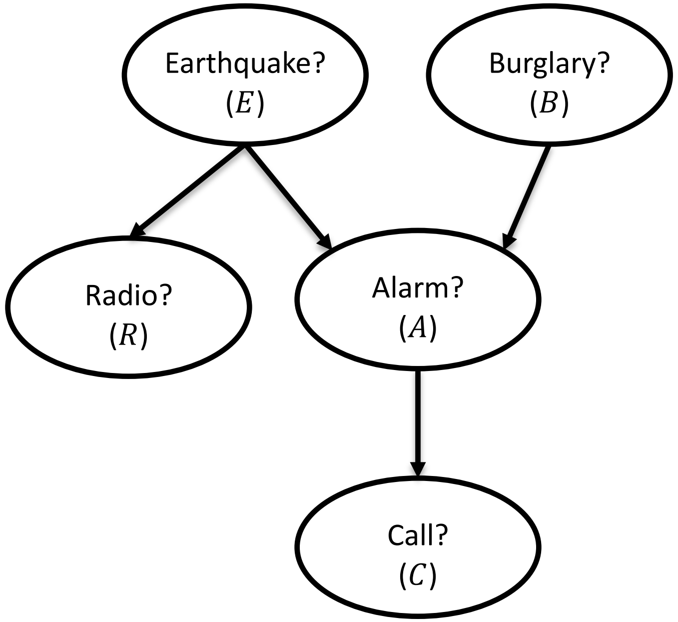




Graphical Test of Independence: Blocking

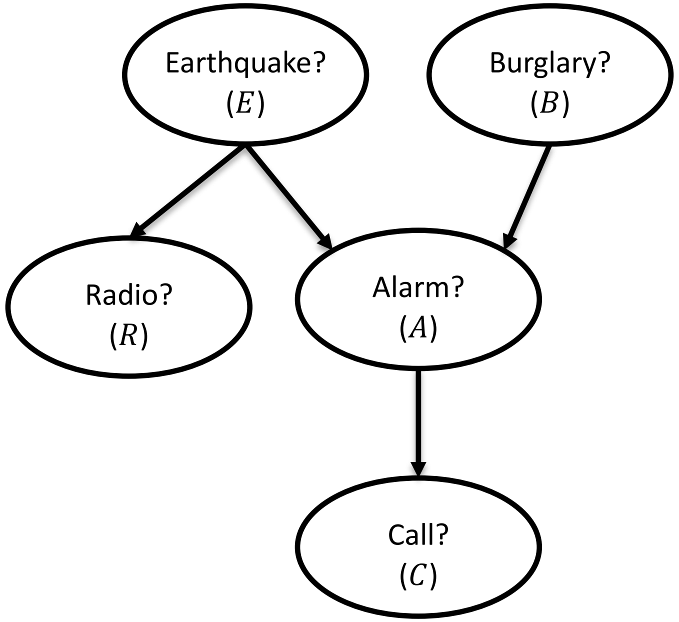
- To gain more intuition, let us use a causal interpretation
  - A sequential valve $N_1 \rightarrow W \rightarrow N_2$ declares W as an intermediary between cause N_1 and its effect N_2
  - A divergent valve $N_1 \leftarrow W \rightarrow N_2$ declares W as a common cause of two effects N_1 and N_2
  - A convergent valve $N_1 \rightarrow W \leftarrow N_2$ declares W as a common effect of two causes N_1 and N_2
- Now, we can better motivate the conditions for closed valves
  - A sequential valve is closed if W appears in $\boldsymbol{Z}$
  - A divergent valve is closed if W appears in $\boldsymbol{Z}$
  - A convergent valve is closed iff neither W nor any of its descendants appears in $\boldsymbol{Z}$





D-Separation: Definition

- Formal definition of d-separation
  - Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be disjoint sets of nodes in a DAG $\boldsymbol{G}$. We will say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, written $dsep_G(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$, iff every path between a node in $\boldsymbol{X}$ and a node in $\boldsymbol{Y}$ is blocked by $\boldsymbol{Z}$.
  - A path is blocked by $\boldsymbol{Z}$ iff at least one valve on the path is closed given $\boldsymbol{Z}$
  - Notice that a path with no valves ($X \rightarrow \mathrm{Y}$) is never blocked

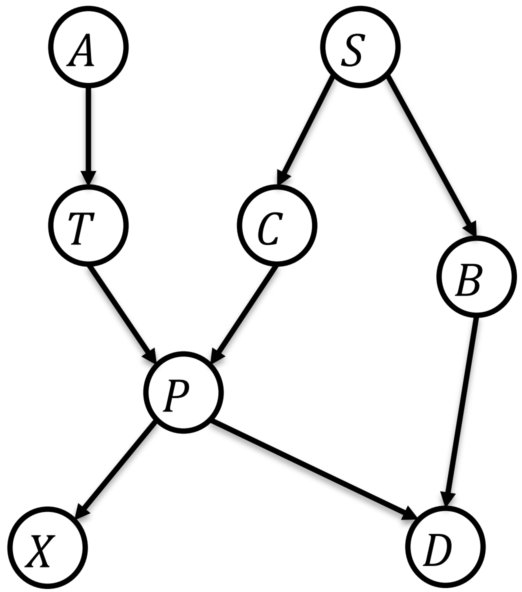


D-Separation: Complexity

- The definition of d-separation calls for considering all paths connecting a node in $\boldsymbol{X}$ with a node in $\boldsymbol{Y}$
  - The number of paths can be exponential
  - But we can implement a test without enumerating these paths

- Testing whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in DAG G is equivalent to testing whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disconnected in a new DAG G', obtained as follows
  - We delete any leaf node W from G if W does not belong to $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$. This process is repeated until no more nodes can be deleted
  - We delete all edges outgoing from nodes in $\boldsymbol{Z}$

  - The connectivity test on DAG G' ignores edge direction
  - This procedure time and space are linear in the size of the DAG G



A, S d-separated from D, X by B, P?
T, C d-separated from B by S, X?

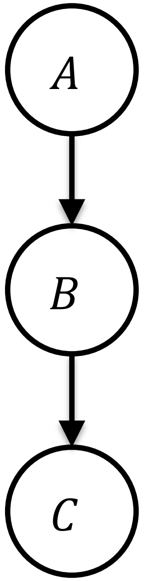
D-Separation: Soundness and Completeness

- The d-separation test is *sound*
  - If P is a probability distribution induced by a Bayesian network (G, Θ) then $dsep_G(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ only if $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$
  - We can safely use d-separation test to derive independence statements about the probability distributions induced by Bayesian networks
  - The proof is constructive and shows that every independence claimed by d-separation can be derived using the graphoid axioms
- The d-separation test is not *complete*
  - It is not capable of inferring every possible independence statement that holds in the induced distribution P
  - The explanation is that some independences may be hidden in the network parameters

A	θ_A
a	.6
$\bar{a}$	.4

A	B	$\theta_{B\|A}$
a	b	.8
a	$\bar{b}$	.2
$\bar{a}$	b	.8
$\bar{a}$	$\bar{b}$	.2

B	C	$\theta_{C\|B}$
b	c	.7
$\bar{b}$	$\bar{c}$	.3
b	c	.1
$\bar{b}$	$\bar{c}$	.9



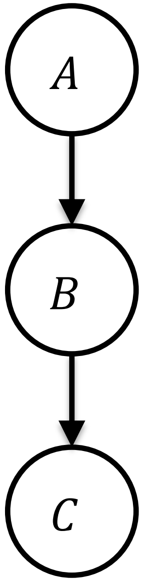
D-Separation: Soundness and Completeness

- Therefore, if we choose the parametrization carefully, we establish independences that d-separation cannot detect
  - This is not surprising since d-separation has no access to the graph parametrization
- We can conclude that, given a distribution P induced by a Bayesian network (G, Θ)
  - If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given $\boldsymbol{Z}$ for any parametrization Θ
  - If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not d-separated by $\boldsymbol{Z}$, then whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent given $\boldsymbol{Z}$ depends on the specific parametrization Θ

A	θ_A
a	.6
$\bar{a}$	.4

A	B	$\theta_{B\|A}$
a	b	.8
a	$\bar{b}$	.2
$\bar{a}$	b	.8
$\bar{a}$	$\bar{b}$	.2

B	C	$\theta_{C\|B}$
b	c	.7
$\bar{b}$	$\bar{c}$	.3
b	c	.1
$\bar{b}$	$\bar{c}$	.9



D-Separation: Soundness and Completeness

- We can always parametrize a DAG G in such a way to ensure the completeness of d-separation

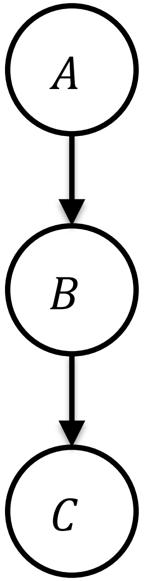
- d-separation satisfies the following weak notion of completeness
  - For every DAG G, there is a parametrization Θ such that $dsep_G(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ if and only if $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$

- This weaker notion of completeness implies that one cannot improve on the d-separation test
  - There is no other graphical test that can derive more independencies from G

A	θ_A
a	.6
$\bar{a}$	.4

A	B	$\theta_{B\|A}$
a	b	.8
a	$\bar{b}$	.2
$\bar{a}$	b	.8
$\bar{a}$	$\bar{b}$	.2

B	C	$\theta_{C\|B}$
b	c	.7
b	$\bar{c}$	.3
$\bar{b}$	c	.1
$\bar{b}$	$\bar{c}$	.9



Independence Maps: I-MAPs

- Independence maps describe the relationship between independence in a DAG and in a probability distribution
  - They are useful to understand the expressive power of DAGs as a language for independence statements
- Let G be a DAG and P a probability distribution over the same variables
  - G is an independence map (I-MAP) of P iff
  - It means that every independence declared by d-separation holds in P
- An I-MAP is *minimal* if G ceases to be an I-MAP if we delete any edges from G
  - If P is induced by a Bayesian network (G, Θ), then G must be an I-MAP of P
  - But it may not be minimal

$dsep_G(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ only if $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$

Independence Maps: D-MAPs

- G is a dependency map (D-MAP) of P iff
  - It means that the lack of d-separation in G implies a dependence in P
  - If P is induced by the Bayesian network (G, Θ), then G is not necessarily a D-MAP of P
  - G can be made a D-MAP of P if we choose the parametrization Θ carefully

$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ only if $dsep_G(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$

Independence Maps: Perfect MAPs

- If a DAG G is both an I-MAP and a D-MAP of P, then G is a *perfect map*
  - We want G to be a P-MAP of the induced distribution to make all independences of P accessible to d-separation
  - However, there are probability distributions for which there are no P-MAPs
- Suppose we have four variables and a distribution P that *only* satisfies these dependencies
  - There is no DAG that is a P-MAP of P in this case

$$X_1 \perp X_2 \mid Y_1, Y_2$$

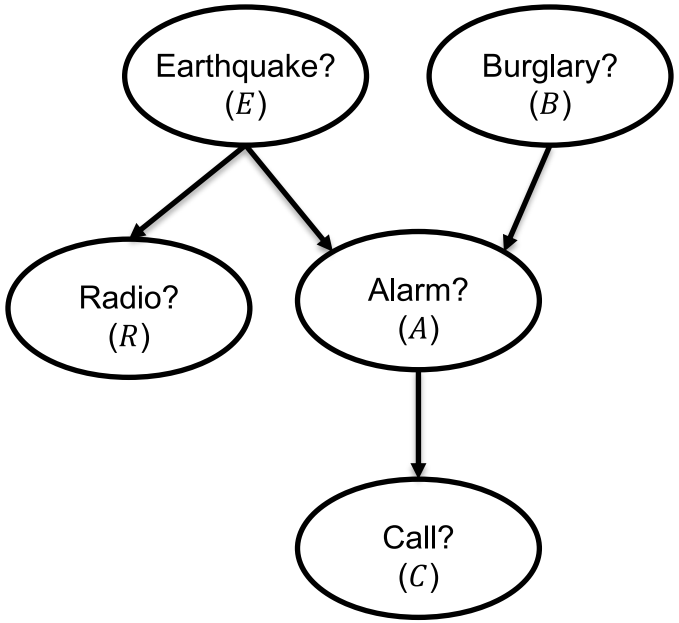
$$X_2 \perp X_1 \mid Y_1, Y_2$$

$$Y_1 \perp Y_2 \mid X_1, X_2$$

$$Y_2 \perp Y_1 \mid X_1, X_2$$

Independence Maps

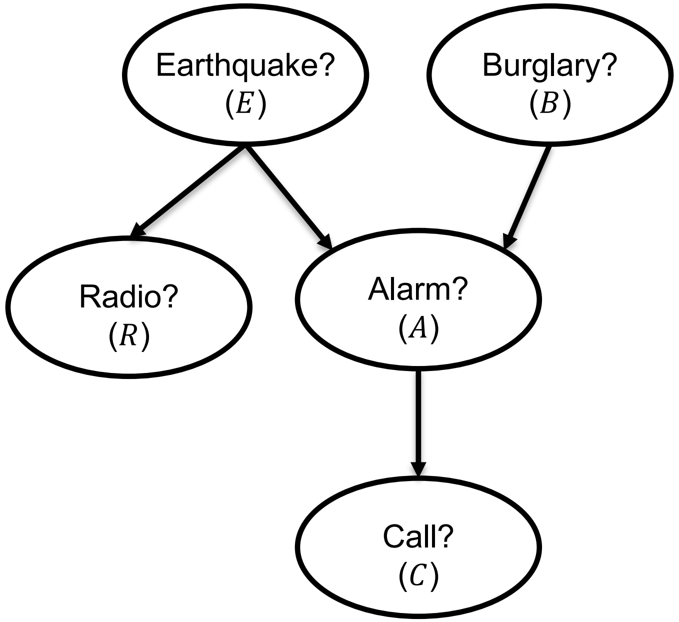
- Given a distribution P, how can we construct a DAG that is guaranteed to be a minimal I-MAP of P
  - Minimal I-MAPs tend to exhibit more independences
  - Therefore, requiring fewer parameters and leading to more compact networks
- Procedure to build a minimal I-MAP
  - Given ordering $X_1, \ldots, X_n$ of variables in P
  - Start with an empty DAG G and consider the variable X_i for $i = 1 \ldots n$
  - For each X_i, identify a minimal subset $\boldsymbol{P}$ of variables $X_1, \ldots, X_{i-1}$ such that $X_i \perp \mathrm{X}_1 \ldots, X_{i-1} \setminus \boldsymbol{P} \mid \boldsymbol{P}$
  - Make $\boldsymbol{P}$ the parents of X_i in G



A, B, C, E, R

Independence Maps

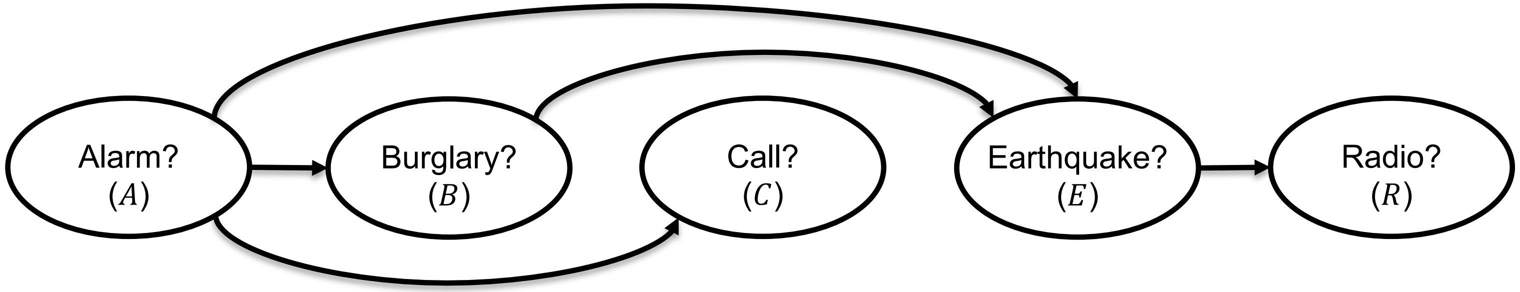
- Suppose this graph is a P-MAP of some distribution P

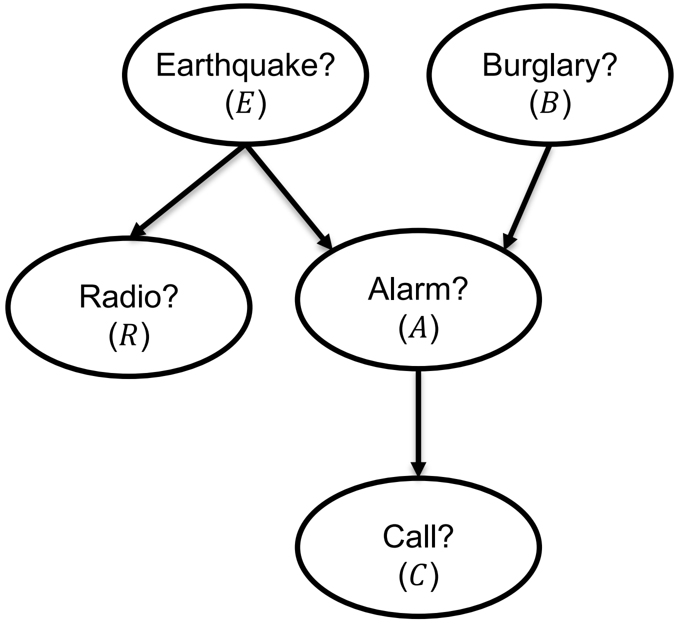


A, B, C, E, R

Independence Maps

- Suppose this graph is a P-MAP of some distribution P



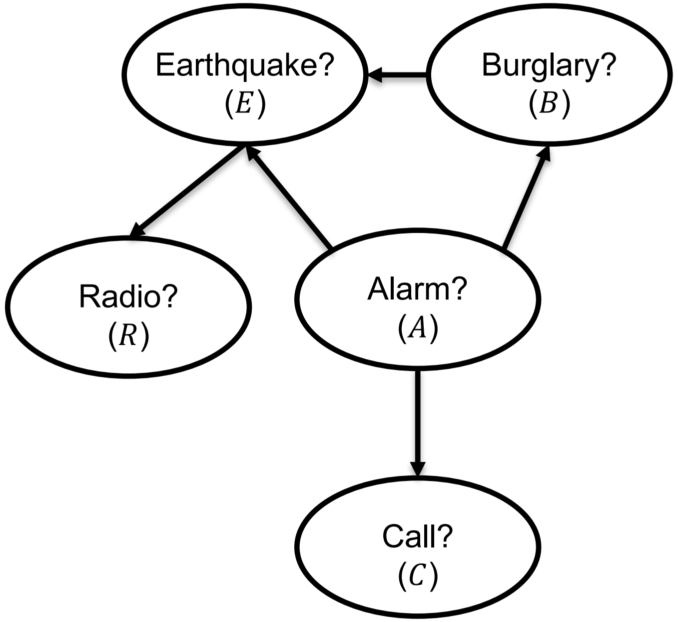


A, B, C, E, R

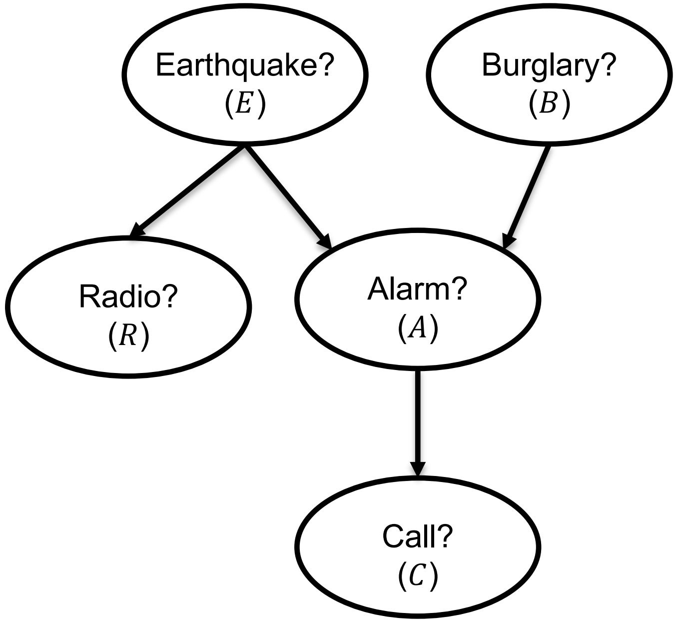
$\boldsymbol{P} = \emptyset$ $\quad$ $\boldsymbol{P} = A$ $\quad$ $\boldsymbol{P} = A$ $\quad$ $\boldsymbol{P} = A, B$ $\quad$ $\boldsymbol{P} = E$

Independence Maps

- Suppose this graph is a P-MAP of some distribution P



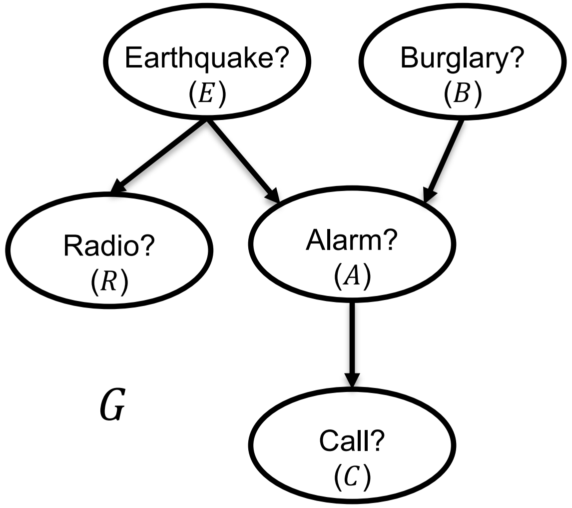
I-MAP procedure

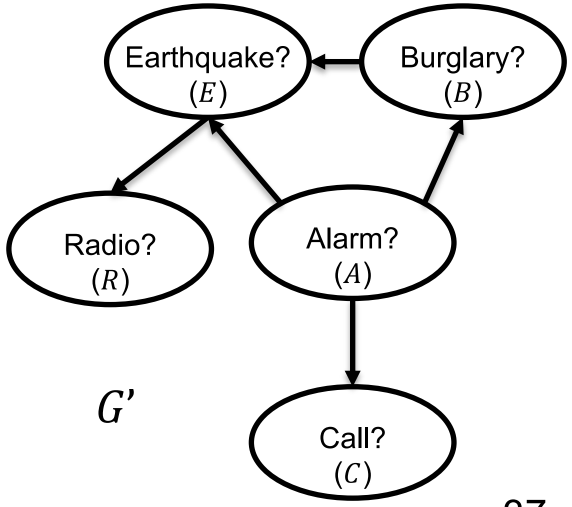


Causal

Independence Maps

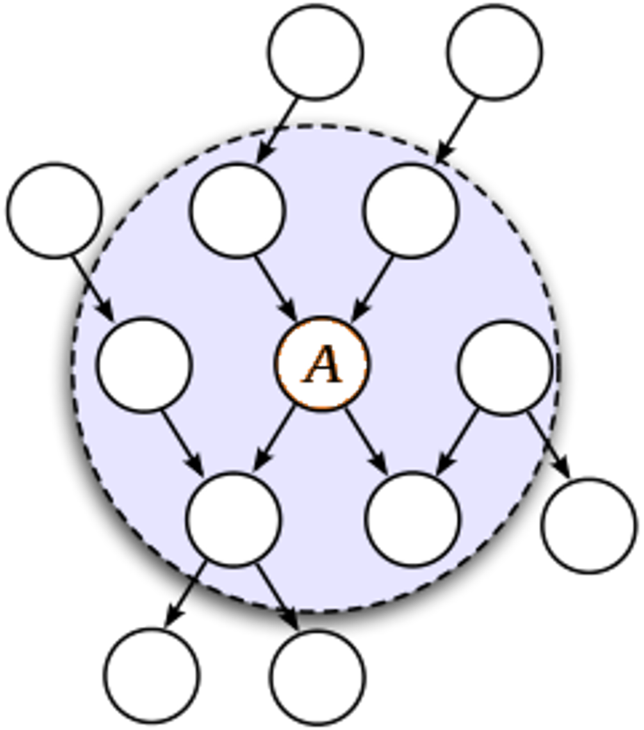
- The resulting DAG G' is guaranteed to be minimal
  - d-separation in G' leads to d-separation in G and independence in P
  - This ceases to hold if we delete any edges of G'
- G' is incompatible with causal relationships
  - Yet it is sound from an independence viewpoint
  - A person that agrees with G cannot disagree with the independences in G'
- Minimal I-MAP is not unique
  - It depends of the variable ordering
  - But also we may have multiple I-MAPs for a single ordering
  - Since we may find multiple minimal sets $\boldsymbol{P}$ for the same variable X_i





Blankets and Boundaries

- An important notion for independence is the *Markov blanket*
  - Let P be a distribution over variables $\boldsymbol{X}$. A *Markov blanket* for a variable $X \in \boldsymbol{X}$ is the set of variables $\boldsymbol{B} \subseteq \boldsymbol{X}$ such that $X \notin \boldsymbol{B}$ and $X \perp \boldsymbol{X} \setminus (\boldsymbol{B} \cup \{X\}) \mid \boldsymbol{B}$
  - A Markov blanket for X will render every other variable irrelevant to X
  - A minimal Markov blanket is known as a *Markov boundary.* A blanket is minimal iff no strict subset of $\boldsymbol{B}$ is also a Markov blanket
- If P is a distribution induced by a DAG G, then a Markov blanket for X can be constructed with its parents, children, and spouses in G.
  - A variable Y is a spouse of X if the two variables have a common child in G



Conclusion

- Bayesian networks are a graphical model with a DAG
  - The graph represents the independencies between variables
  - The parametrisation expresses the strength of the dependencies
- D-separation provides a convenient and efficient approach to detect independencies
  - However, additional independencies may be hidden in the graph parametrisation
  - We also discussed the concepts of I-MAP, D-MAP and P-MAP
- Tasks
  - Read Chapter 4 from the textbook (Darwiche)