Algorithms for Reasoning with graphical models

Slides Set 5: Probabilistic Networks

Rina Dechter

Darwiche chapter 3,4,
Pearl: chapters 3

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- Capturing CIs by graphs
- D-separation: Inferring CIs in graphs

Examples: Common Sense Reasoning

- **Zebra on Pajama**: (7:30 pm): I told Susannah: you have a nice pajama, but it was just a dress. Why jump to that conclusion?: 1. because time is night time. 2. certain designs look like pajama.
- **Cars going out of a parking lot:** You enter a parking lot which is quite full (UCI), you see a car coming : you think ah… now there is a space (vacated), OR… there is no space and this guy is looking and leaving to another parking lot. What other clues can we have?
- **Robot gets out at a wrong level:** A robot goes down the elevator. stops at 2^{nd} floor instead of ground floor. It steps out and should immediately recognize not being in the right level, and go back inside.
- **Turing quotes**
  - If machines will not be allowed to be fallible they cannot be intelligent
  - (Mathematicians are wrong from time to time so a machine should also be allowed)

Why/What/How Uncertainty?

- Why Uncertainty?
  - Answer: It is abandant
- What formalism to use?
  - Answer: Probability theory
- How to overcome exponential representation?
  - Answer: Graphs, graphs, graphs… to capture irrelevance, independence

Why Uncertainty?

- AI goal: to have a declarative, model-based, framework that allows computer system to reason.
- People reason with partial information
- Sources of uncertainty:
  - Limitation in observing the world: e.g., a physician see symptoms and not exactly what goes in the body when he performs diagnosis. Observations are noisy (test results are inaccurate)
  - Limitation in modeling the world,
  - maybe the world is not deterministic.

Degrees of Belief

- Assign a degree of belief or probability in $[0, 1]$ to each world ω and denote it by $\Pr(\omega)$.
- The belief in, or probability of, a sentence α:

$$\Pr(\alpha) \stackrel{\textit{def}}{=} \sum_{\omega \models \alpha} \Pr(\omega).$$

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

Properties of Beliefs

- A bound on the belief in any sentence:

$$0 \leq \Pr(\alpha) \leq 1 \qquad \text{for any sentence } \alpha.$$

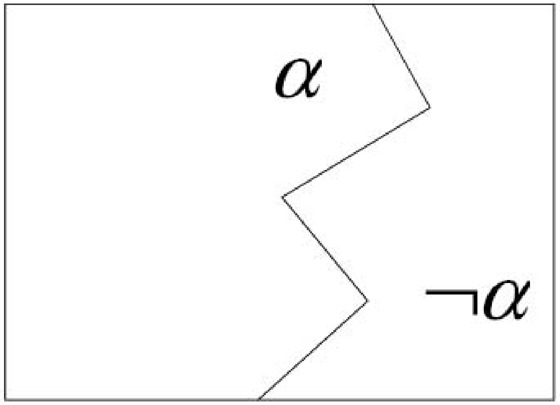
- A baseline for inconsistent sentences:

$$\Pr(\alpha) = 0 \qquad \text{when } \alpha \text{ is inconsistent.}$$

- A baseline for valid sentences:

$$\Pr(\alpha) = 1 \qquad \text{when } \alpha \text{ is valid.}$$

Properties of Beliefs



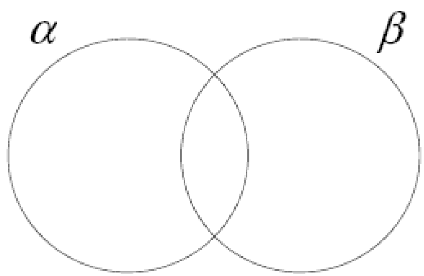
- The belief in a sentence given the belief in its negation:

$$\Pr(\alpha) + \Pr(\neg\alpha) = 1.$$

Example

$$\begin{aligned}\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\ \Pr(\neg\text{Burglary}) &= \Pr(\omega_3) + \Pr(\omega_4) + \Pr(\omega_7) + \Pr(\omega_8) = .8\end{aligned}$$

Properties of Beliefs



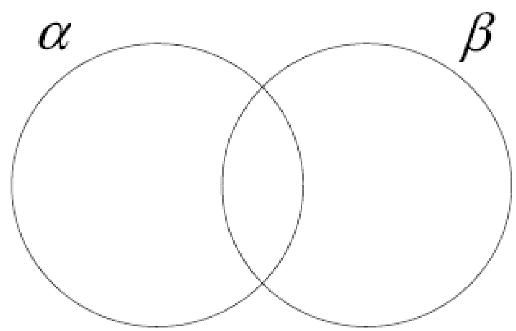
- The belief in a disjunction:

$$\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \wedge \beta).$$

- Example:

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
\Pr(\text{Earthquake} \wedge \text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) = .02 \\
\Pr(\text{Earthquake} \vee \text{Burglary}) &= .1 + .2 - .02 = .28
\end{aligned}$$

Properties of Beliefs



- The belief in a disjunction:

  $\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta)$ when α and β are mutually exclusive.

Entropy

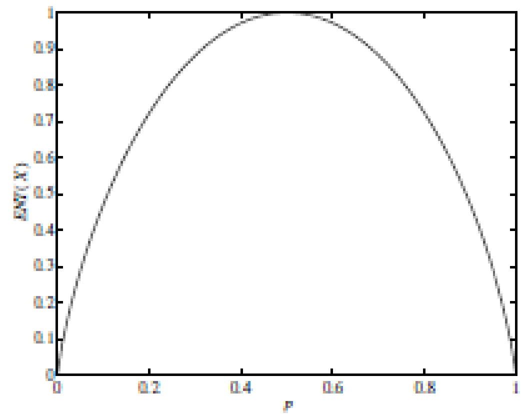
Quantify uncertainty about a variable X using the notion of entropy:

$$\mathrm{ENT}(X) \stackrel{def}{=} -\sum_x \Pr(x) \log_2 \Pr(x),$$

where $0 \log 0 = 0$ by convention.

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558
ENT(.)	.469	.722	.802

Entropy



- The entropy for a binary variable X and varying $p = \Pr(X)$.
- Entropy is non-negative.
- When $p = 0$ or $p = 1$, the entropy of X is zero and at a minimum, indicating no uncertainty about the value of X.
- When $p = \frac{1}{2}$, we have $\Pr(X) = \Pr(\neg X)$ and the entropy is at a maximum (indicating complete uncertainty).

Bayes Conditioning

Alpha and beta are events

Closed form for Bayes conditioning:

$$\Pr(\alpha|\beta) = \frac{\Pr(\alpha \wedge \beta)}{\Pr(\beta)}.$$

Defined only when $\Pr(\beta) \neq 0$.

Degrees of Belief

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= .2 \\
\Pr(\neg\text{Burglary}) &= .8 \\
\Pr(\text{Alarm}) &= .2442
\end{aligned}$$

Belief Change

Burglary is independent of Earthquake

Conditioning on evidence Earthquake:

$$\begin{aligned}\Pr(\text{Burglary}) &= .2\\ \Pr(\text{Burglary}|\text{Earthquake}) &= .2\\ \\ \Pr(\text{Alarm}) &= .2442\\ \Pr(\text{Alarm}|\text{Earthquake}) &\approx .75 \uparrow\end{aligned}$$

The belief in Burglary is not changed, but the belief in Alarm increases.

Belief Change

Earthquake is independent of burglary

Conditioning on evidence Burglary:

$$\begin{aligned}
\Pr(\text{Alarm}) &= .2442 \\
\Pr(\text{Alarm}|\text{Burglary}) &\approx .905 \uparrow \\
\\
\Pr(\text{Earthquake}) &= .1 \\
\Pr(\text{Earthquake}|\text{Burglary}) &= .1
\end{aligned}$$

The belief in Alarm increases in this case, but the belief in Earthquake stays the same.

Belief Change

The belief in Burglary increases when accepting the evidence Alarm. How would such a belief change further upon obtaining more evidence?

- Confirming that an Earthquake took place:

$$\begin{aligned}\Pr(\text{Burglary}|\text{Alarm}) &\approx .741\\ \Pr(\text{Burglary}|\text{Alarm} \wedge \text{Earthquake}) &\approx .253 \downarrow\end{aligned}$$

  We now have an explanation of Alarm.

- Confirming that there was no Earthquake:

$$\begin{aligned}\Pr(\text{Burglary}|\text{Alarm}) &\approx .741\\ \Pr(\text{Burglary}|\text{Alarm} \wedge \neg\text{Earthquake}) &\approx .957 \uparrow\end{aligned}$$

  New evidence will further establish burglary as an explanation.

Conditional Independence

$\Pr$ finds α conditionally independent of β given γ iff

$$\Pr(\alpha|\beta \wedge \gamma) = \Pr(\alpha|\gamma) \quad \text{or } \Pr(\beta \wedge \gamma) = 0.$$

Another definition

$$\Pr(\alpha \wedge \beta|\gamma) = \Pr(\alpha|\gamma)\Pr(\beta|\gamma) \quad \text{or } \Pr(\gamma) = 0.$$

Variable Independence

$\Pr$ finds $\mathbf{X}$ independent of $\mathbf{Y}$ given $\mathbf{Z}$, denoted $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, means that $\Pr$ finds $\mathbf{x}$ independent of $\mathbf{y}$ given $\mathbf{z}$ for all instantiations $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Example

$\mathbf{X} = \{A, B\}$, $\mathbf{Y} = \{C\}$ and $\mathbf{Z} = \{D, E\}$, where A, B, C, D and E are all propositional variables. The statement $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is then a compact notation for a number of statements about independence:

$A \wedge B$ is independent of C given $D \wedge E$;

$A \wedge \neg B$ is independent of C given $D \wedge E$;

$\vdots$

$\neg A \wedge \neg B$ is independent of $\neg C$ given $\neg D \wedge \neg E$;

That is, $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is a compact notation for $4 \times 2 \times 4 = 32$ independence statements of the above form.

Further Properties of Beliefs

Chain rule

$$\begin{aligned} & \Pr(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) \\ = \; & \Pr(\alpha_1 | \alpha_2 \wedge \ldots \wedge \alpha_n)\Pr(\alpha_2 | \alpha_3 \wedge \ldots \wedge \alpha_n) \ldots \Pr(\alpha_n). \end{aligned}$$

Case analysis (law of total probability)

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha \wedge \beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Further Properties of Beliefs

Another version of case analysis

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha|\beta_i)\Pr(\beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Two simple and useful forms of case analysis are these:

$$\begin{aligned}\Pr(\alpha) &= \Pr(\alpha \wedge \beta) + \Pr(\alpha \wedge \neg\beta)\\ \Pr(\alpha) &= \Pr(\alpha|\beta)\Pr(\beta) + \Pr(\alpha|\neg\beta)\Pr(\neg\beta).\end{aligned}$$

The main value of case analysis is that, in many situations, computing our beliefs in the cases is easier than computing our beliefs in α. We shall see many examples of this phenomena in later chapters.

Further Properties of Beliefs

Bayes rule

$$\Pr(\alpha|\beta) = \frac{\Pr(\beta|\alpha)\Pr(\alpha)}{\Pr(\beta)}.$$

- Classical usage: α is perceived to be a cause of β.
- Example: α is a disease and β is a symptom–
- Assess our belief in the cause given the effect.
- Belief in an effect given its cause, $\Pr(\beta|\alpha)$, is usually more readily available than the belief in a cause given one of its effects, $\Pr(\alpha|\beta)$.

Difficulty: Complexity in model construction and inference

- In Alarm example:
  - 31 numbers needed,
  - Quite unnatural to assess: e.g.

$$P(B=y, E=y, A=y, J=y, M=y)$$

  - Computing $P(B=y|M=y)$ takes 29 additions.
- In general,
  - $P(X_1, X_2, \ldots, X_n)$ needs at least $2^n - 1$ numbers to specify the joint probability. Exponential model size.
  - Knowledge acquisition difficult (complex, unnatural),
  - Exponential storage and inference.

Chain Rule and Factorization

Overcome the problem of exponential size by exploiting conditional independence

- The chain rule of probabilities:

$$\begin{aligned} P(X_1, X_2) &= P(X_1)P(X_2|X_1) \\ P(X_1, X_2, X_3) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \\ &\dots \\ P(X_1, X_2, \dots, X_n) &= P(X_1)P(X_2|X_1)\dots P(X_n|X_1, \dots, X_{n-1}) \\ &= \prod_{i=1}^{n} P(X_i|X_1, \dots, X_{i-1}). \end{aligned}$$

- No gains yet. The number of parameters required by the factors is: $2^{n-1} + 2^{n-1} + \ldots + 1 = 2^n - 1$.

Conditional Independence

- About $P(X_i|X_1, \ldots, X_{i-1})$:
  - Domain knowledge usually allows one to identify a subset $pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that
    - Given $pa(X_i)$, X_i is independent of all variables in $\{X_1, \ldots, X_{i-1}\} \setminus pa(X_i)$, i.e.

$$P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|pa(X_i))$$

- Then

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

- Joint distribution factorized.
- The number of parameters might have been substantially reduced.

Example

P(B,E,A,J,M)=?

Example continued

$$
\begin{aligned}
& P(B, E, A, J, M) \\
= \quad & P(B)P(E|B)P(A|B, E)P(J|B, E, A)P(M|B, E, A, J) \\
= \quad & P(B)P(E)P(A|B, E)P(J|A)P(M|A)(Factorization)
\end{aligned}
$$

- $pa(B) = \{\}$, $pa(E) = \{\}$,$pa(A) = \{B, E\}$, $pa(J) = \{A\}$,$pa(M) = \{A\}$.
- Conditional probabilities tables (CPT)

B	P(B)
Y	.01
N	.99

E	P(E)
Y	.02
N	.98

A	B	E	P(A|B, E)
Y	Y	Y	.95
N	Y	Y	.05
Y	Y	N	.94
N	Y	N	.06
Y	N	Y	.29
N	N	Y	.71
Y	N	N	.001
N	N	N	.999

M	A	P(M|A)
Y	Y	.9
N	Y	.1
Y	N	.05
N	N	.95

J	A	P(J|A)
Y	Y	.7
N	Y	.3
Y	N	.01
N	N	.99

Example continued

- Model size reduced from 31 to 1+1+4+2+2=10
- Model construction easier
  - Fewer parameters to assess.
  - Parameters more natural to assess:e.g.

$$P(B = Y), P(E = Y), P(A = Y|B = Y, E = Y),$$

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

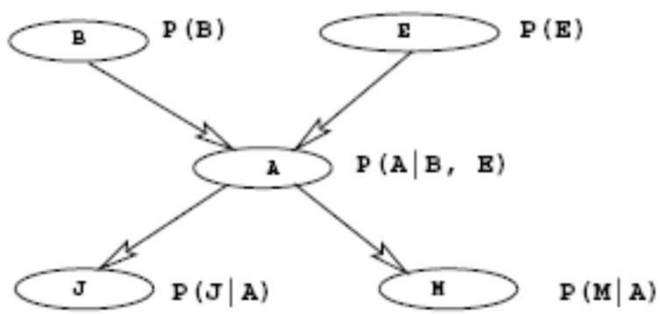
- Inference easier.Will see this later.

From Factorizations to Bayesian Networks

Graphically represent the conditional independency relationships:

- construct a directed graph by drawing an arc from X_j to X_i iff $X_j \in pa(X_i)$

$$pa(B) = \{\}, \; pa(E) = \{\}, \; pa(A) = \{B, E\}, \; pa(J) = \{A\}, \; pa(M) = \{A\}.$$



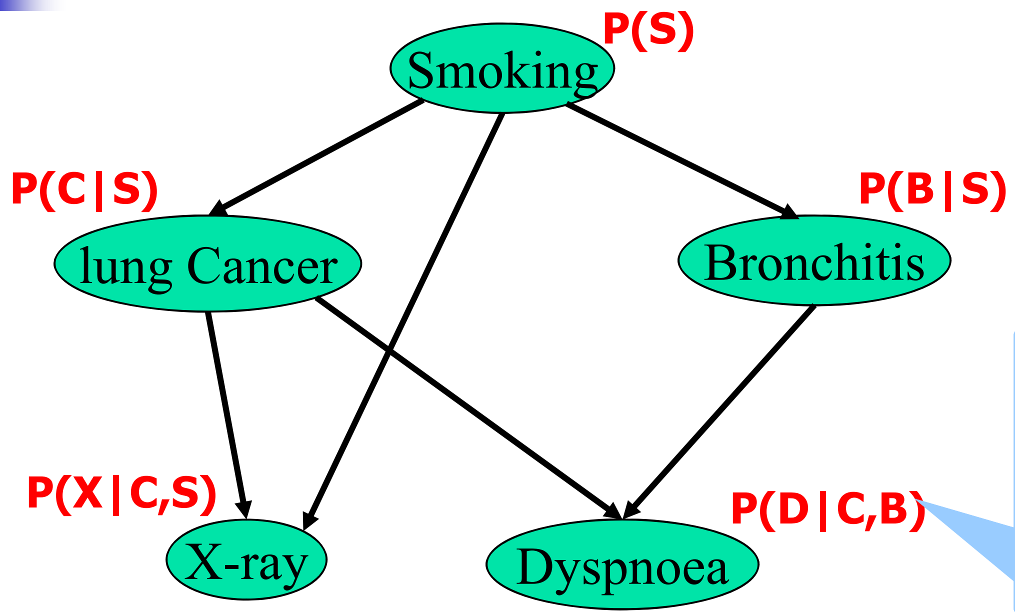
- Also attach the conditional probability (table) $P(X_i|pa(X_i))$ to node X_i.
- What results in is a **Bayesian network**. Also known as **belief network, probabilistic network**.

Formal Definition

A **Bayesian network** is:

- An directed acyclic graph (DAG), where
- Each node represents a random variable
- And is associated with the conditional probability of the node given its parents.

Bayesian Networks: Representation



BN = (G, Θ)

CPD:

C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

$$P(S, C, B, X, D) = P(S)\, P(C|S)\, P(B|S)\, P(X|C,S)\, P(D|C,B)$$

Conditional Independencies ➡ Efficient Representation

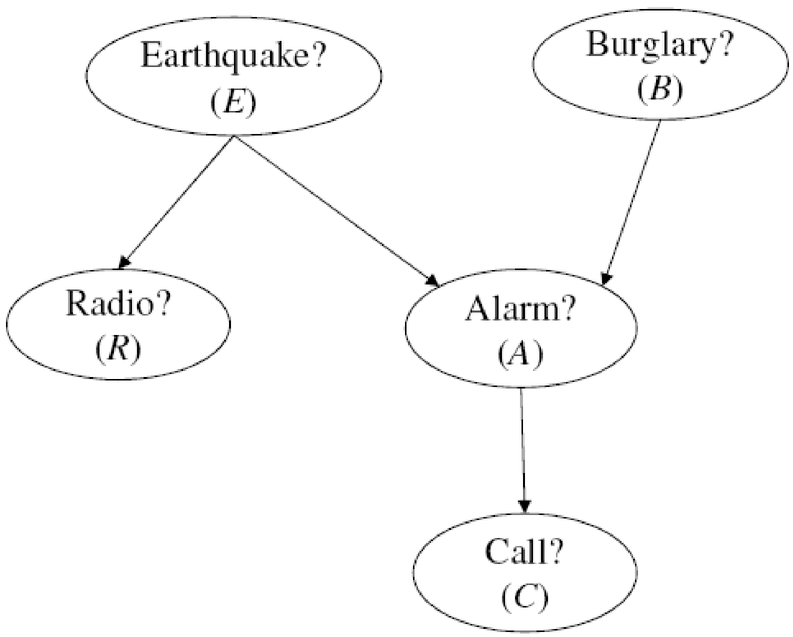
Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs

(Darwiche chapter 4)

Capturing Independence Graphically

The causal interpretation

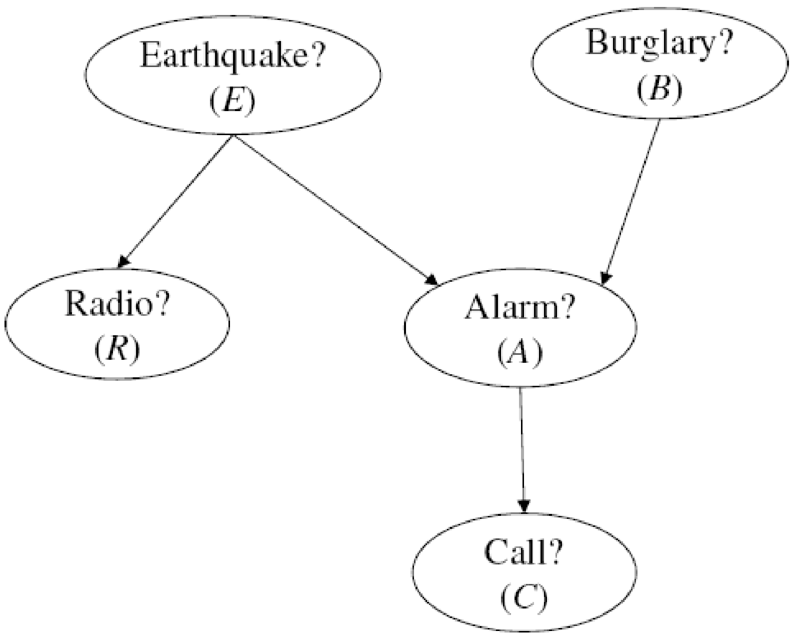


Assume that edges in this graph represent direct causal influences among these variables.

Example

The alarm triggering (A) is a direct cause of receiving a call from a neighbor (C).

Capturing Independence Graphically

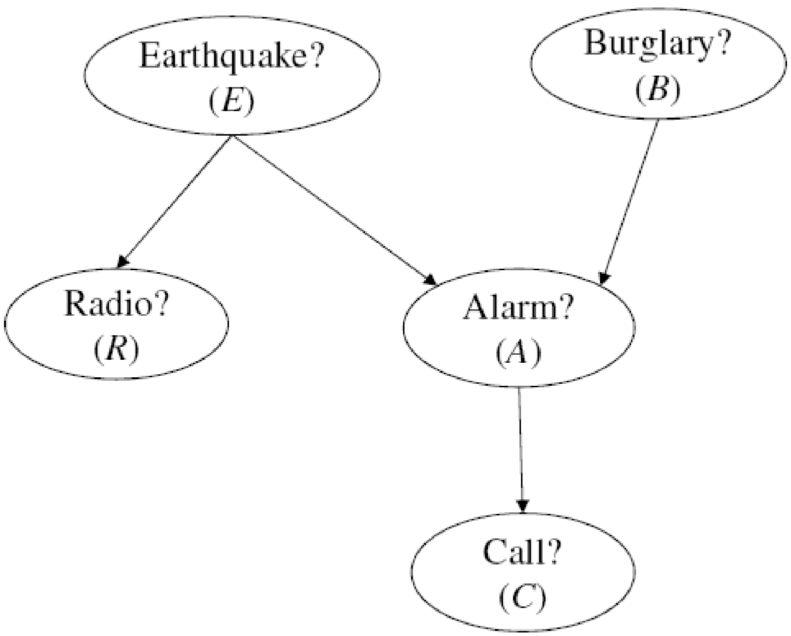


We expect our belief in C to be influenced by evidence on R.

Example

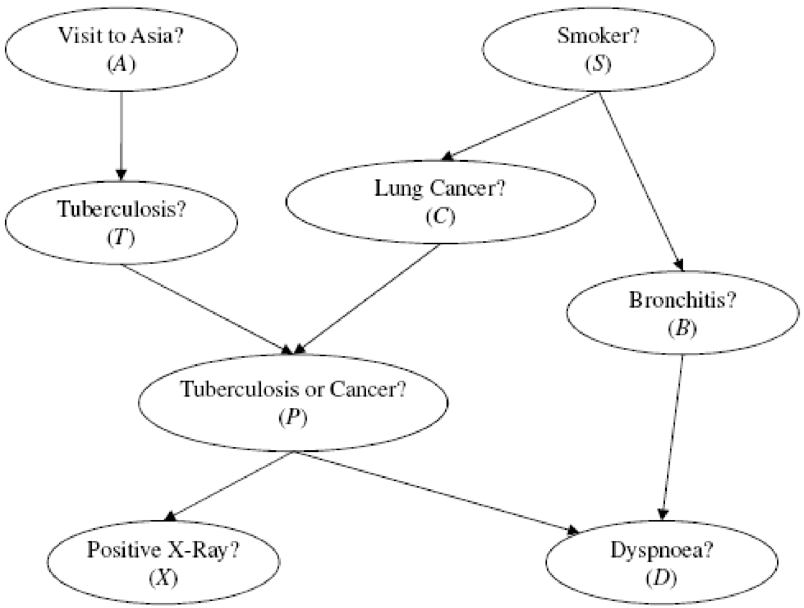
If we get a radio report that an earthquake took place in our neighborhood, our belief in the alarm triggering would probably increase, which would also increase our belief in receiving a call from our neighbor.

Capturing Independence Graphically



We would not change this belief, however, if we knew for sure that the alarm did not trigger. That is, we would find C independent of R given $\neg A$ in the context of this causal structure.

Capturing Independence Graphically



We would clearly find a visit to Asia relevant to our belief in the X-Ray test coming out positive, but we would find the visit irrelevant if we know for sure that the patient does not have Tuberculosis. That is, X is dependent on A, but is independent of A given $\neg T$.

Capturing Independence Graphically

These examples of independence are all implied by a formal interpretation of each DAG as a set of conditional independence statements.

Given a variable V in a DAG G:

$\text{Parents}(V)$ are the parents of V in DAG G, that is, the set of variables N with an edge from N to V.

$\text{Descendants}(V)$ are the descendants of V in DAG G, that is, the set of variables N with a directed path from V to N (we also say that V is an ancestor of N in this case).

$\text{Non\_Descendants}(V)$ are all variables in DAG G other than V, $\text{Parents}(V)$ and $\text{Descendants}(V)$. We will call these variables the non-descendants of V in DAG G.

Capturing Independence Graphically

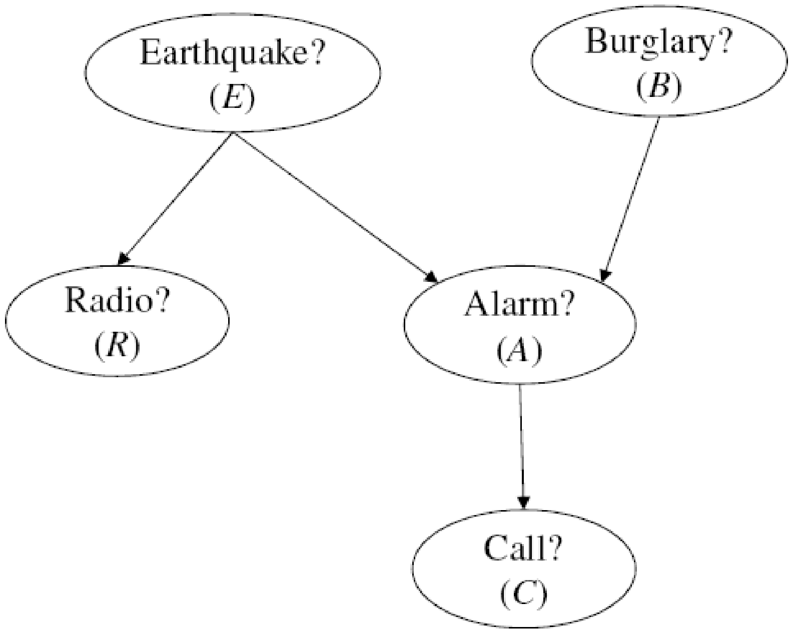
We will formally interpret each DAG G as a compact representation of the following independence statements (Markovian assumptions):

$$I(V, \text{Parents}(V), \text{Non\_Descendants}(V)),$$

for all variables V in DAG G.

- If we view the DAG as a causal structure, then $\text{Parents}(V)$ denotes the direct causes of V and $\text{Descendants}(V)$ 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.

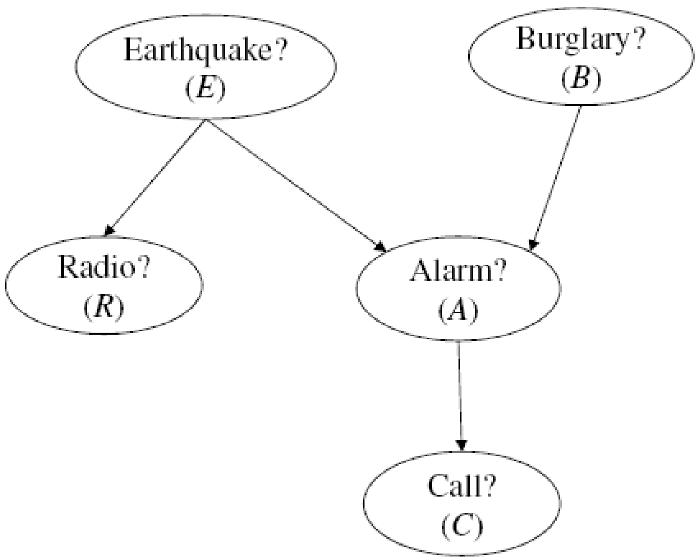
Capturing Independence Graphically

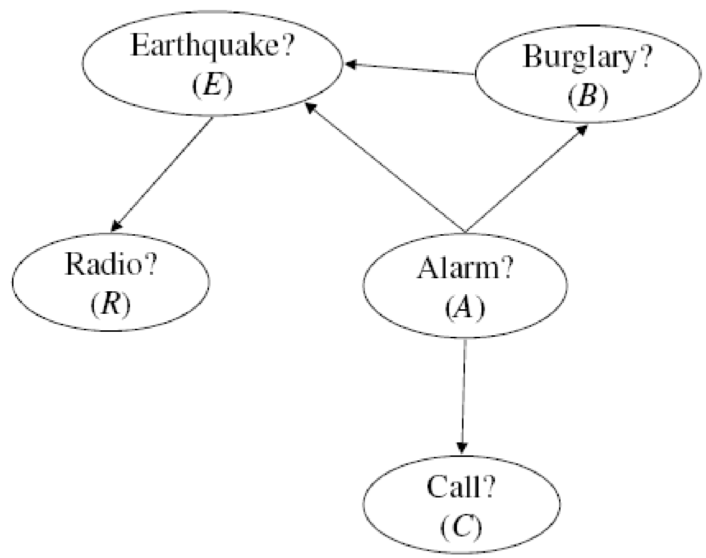


$I(C, A, \{B, E, R\})$
$I(R, E, \{A, B, C\})$
$I(A, \{B, E\}, R)$
$I(B, \emptyset, \{E, R\})$
$I(E, \emptyset, B)$

Note that variables B and E have no parents, hence, they are marginally independent of their non-descendants.

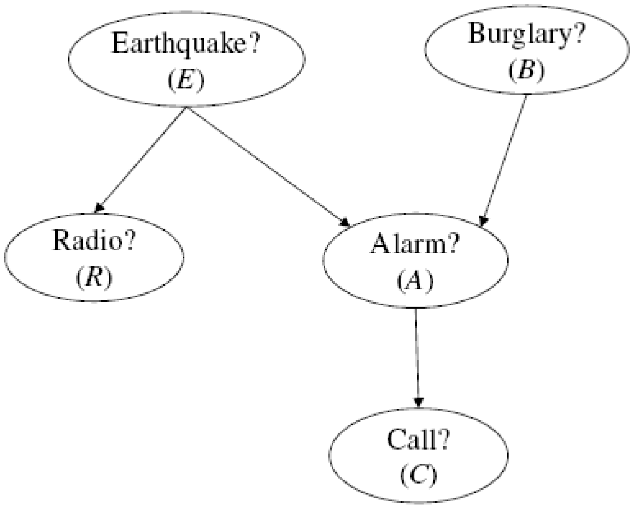
Capturing Independence Graphically





Every independence which is declared (or implied) by the second DAG is also declared (or implied) by the first one. Hence, if we accept the first DAG, then we must also accept the second.

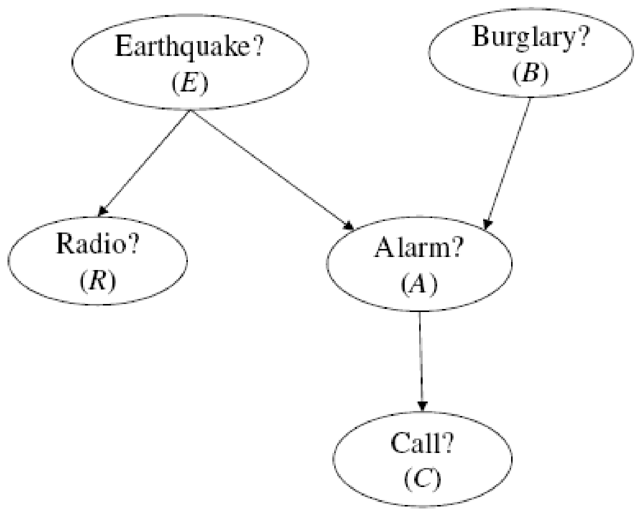
Parameterizing the Independence Structure



- The DAG G is a partial specification of our state of belief Pr.
- By constructing G, we are saying that the distribution Pr must satisfy the independence assumptions in $\mathrm{Markov}(G)$.
- This clearly constrains the possible choices for the distribution Pr, but does not uniquely define it.

We can augment the DAG G by a set of conditional probabilities that together with $\mathrm{Markov}(G)$ are guaranteed to define the distribution Pr uniquely.

Parameterizing the Independence Structure



For every variable X in the DAG G, and its parents $\mathbf{U}$, we need to provide the probability $\Pr(x|\mathbf{u})$ for every value x of variable X and every instantiation $\mathbf{u}$ of parents $\mathbf{U}$.

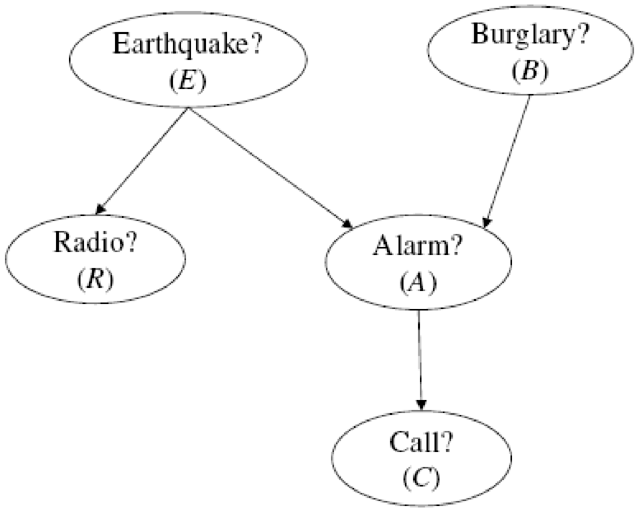
Example

We need to provide the following conditional probabilities:

$$\Pr(c|a),\ \ \Pr(r|e),\ \ \Pr(a|b,e),\ \ \Pr(e),\ \ \Pr(b),$$

where a, b, c, e and r are values of variables A, B, C, E and R.

Parameterizing the Independence Structure



The conditional probabilities required for variable C:

A	C	$\Pr(c\|a)$
true	true	.80
true	false	.20
false	true	.001
false	false	.999

The above table is known as a Conditional Probability Table (CPT) for variable C.

$$\Pr(c|a) + \Pr(\bar{c}|a) = 1 \text{ and } \Pr(c|\bar{a}) + \Pr(\bar{c}|\bar{a}) = 1.$$

Two of the probabilities in the above CPT are redundant and can be inferred from the other two. We only need 10 independent probabilities to completely specify the CPTs for this DAG.

Parameterizing the Independence Structure

Definition

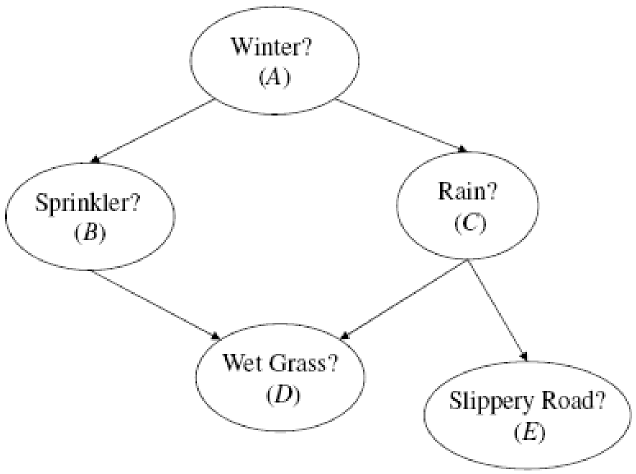
A Bayesian network for variables $\mathbf{Z}$ is a pair (G, Θ), where

- G is a directed acyclic graph over variables $\mathbf{Z}$, called the network structure.
- Θ is a set of conditional probability tables (CPTs), one for each variable in $\mathbf{Z}$, called the network parametrization.

- $\Theta_{X|\mathbf{U}}$: the CPT for variable X and its parents $\mathbf{U}$.
- $X\mathbf{U}$: a network family.
- $\theta_{x|\mathbf{u}}$: the value assigned by CPT $\Theta_{X|\mathbf{U}}$ to the conditional probability $\Pr(x|\mathbf{u})$. Called a network parameter.

We must have $\sum_x \theta_{x|\mathbf{u}} = 1$ for every parent instantiation $\mathbf{u}$.

Parameterizing the Independence Structure



A	B	$\Theta_{B\|A}$
true	true	.2
true	false	.8
false	true	.75
false	false	.25

A	C	$\Theta_{C\|A}$
true	true	.8
true	false	.2
false	true	.1
false	false	.9

A	Θ_A
true	.6
false	.4

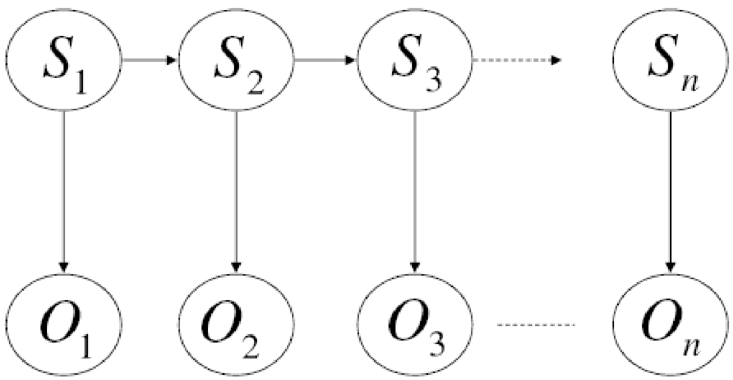
B	C	D	$\Theta_{D\|B,C}$
true	true	true	.95
true	true	false	.05
true	false	true	.9
true	false	false	.1
false	true	true	.8
false	true	false	.2
false	false	true	0
false	false	false	1

C	E	$\Theta_{E\|C}$
true	true	.7
true	false	.3
false	true	0
false	false	1

Parameterizing the Independence Structure

Each state variable S_i has m values and similarly for sensor variables O_i.

Hidden Markov Model



- The CPT for any state variable S_i, $i > 1$, will then contain m^2 parameters, which are usually known as transition probabilities.
- The CPT for any sensor variable O_i will have m^2 parameters, which are usually known as emission or sensor probabilities.

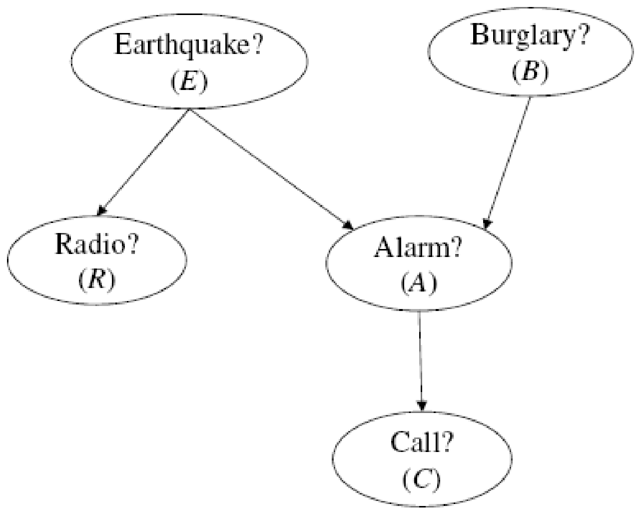
The CPT for S_1 will only have m parameters. The CPTs for S_i, $i > 1$, are all identical, and so are the CPTs for all sensor variables O_i. The HMM is said to be homogeneous in this case.

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs

Properties of Probabilistic Independence

This independence follows from the Markov assumption

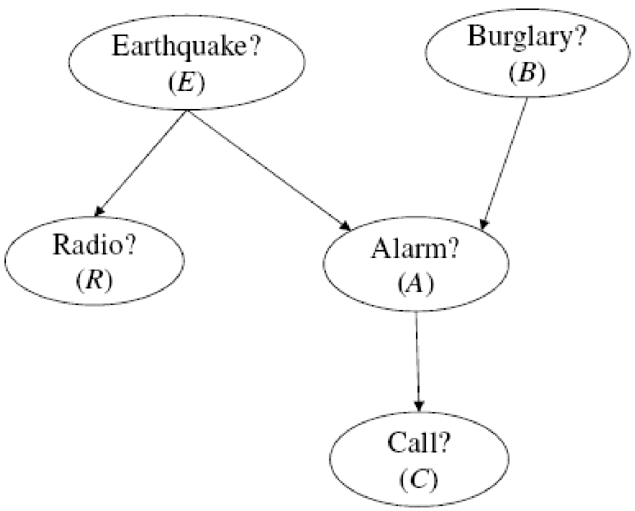


The distribution $\Pr$ specified by a Bayesian network (G, Θ) is guaranteed to satisfy every independence assumption in $\text{Markov}(G)$.

These, however, are not the only independencies satisfied by the distribution $\Pr$.

R and C are independent given A

Properties of Probabilistic Independence



D and E are independent given A, C

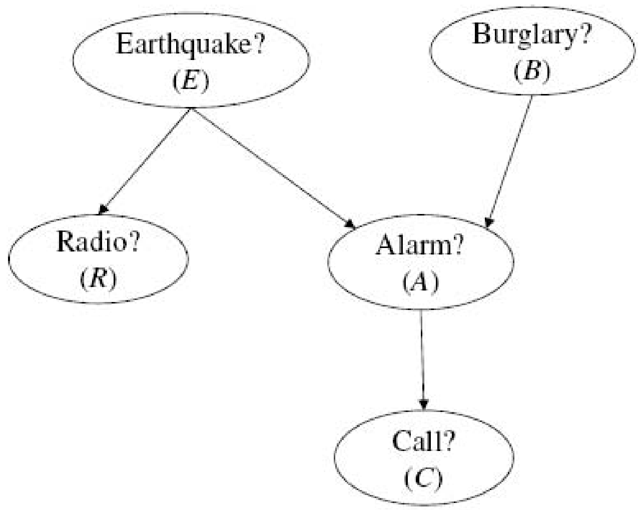
This independence, and additional ones, follow from the ones in $\mathrm{Markov}(G)$ using a set of properties for probabilistic independence, known as the graphoid axioms, which include Symmetry, Decomposition, Weak Union, and Contraction.

Properties of Probabilistic independence

THEOREM 1: Let X, Y, and Z be three disjoint subsets of variables from U. If $I(X, Z, Y)$ stands for the relation "X is independent of Y, given Z" in some probabilistic model P, then I must satisfy the following four independent conditions:

- Symmetry:
  - **I(X,Z,Y) → I(Y,Z,X)**
- Decomposition:
  - **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)→I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)→I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) → I(X,Z,YW)**

Symmetry



$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ iff $I_{\Pr}(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$

If learning **y** does not influence our belief in **x**, then learning **x** does not influence our belief in **y** either.

Example

From the independencies declared by $\mathrm{Markov}(G)$, we know that $I_{\Pr}(A, \{B, E\}, R)$. Using Symmetry, we can then conclude that $I_{\Pr}(R, \{B, E\}, A)$, which is not part of the independencies declared by $\mathrm{Markov}(G)$.

Decomposition

If some information is irrelevant, then any part of it is also irrelevant.

$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W}) \Longrightarrow I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{W}).$$

If learning **yw** does not influence our belief in **x**, then learning **y** alone, or learning **w** alone, will not influence our belief in **x** either.

Pearl language:
If two pieces of information are irrelevant to X then each one is irrelevant to X

Decomposition

The opposite of Decomposition, called Composition:

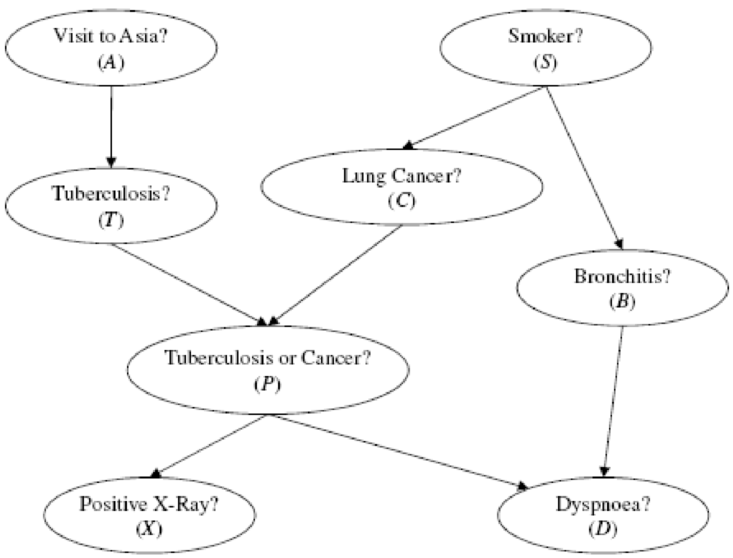
$$I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{W}) \text{ ~~only if~~ } \Longrightarrow I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

does not hold in general.

Two pieces of information may each be irrelevant on their own, yet their combination may be relevant.

Example: Two coins and a bell

Decomposition



Example

From $\mathrm{Markov}(G)$ we have
$I(B, S, \{A, C, P, T, X\})$.
By Decomposition, we get $I(B, S, C)$.

Once we know whether the person is a smoker, our belief in developing bronchitis is no longer influenced by information about developing cancer.

This independence holds in any probability distribution induced by a parametrization of DAG G. Yet, this independence is not part of the independencies declared by $\mathrm{Markov}(G)$.

Decomposition

More generally...

Decomposition allows us to state the following:

$$I_{\mathrm{Pr}}(X, \mathrm{Parents}(X), \mathbf{W}) \quad \text{for every } \mathbf{W} \subseteq \mathrm{Non\_Descendants}(X).$$

Every variable X is conditionally independent of any subset of its non-descendants given its parents.

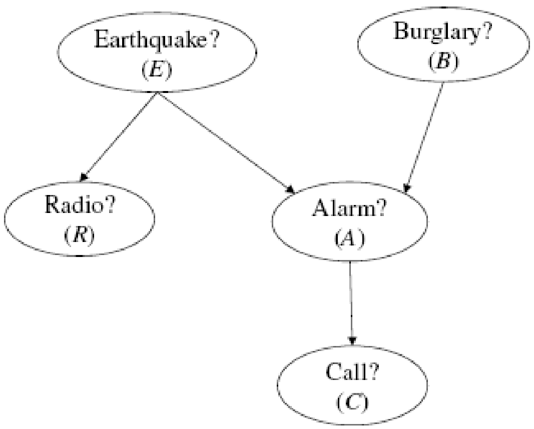
This is a strengthening of the independence statements declared by $\mathrm{Markov}(G)$, which is a special case when $\mathbf{W}$ contains all non-descendants of X.

Decomposition

Decomposition proves the chain rule for Bayesian networks.

By the chain rule of probability calculus:

$$\Pr(r, c, a, e, b) = \Pr(r|c, a, e, b)\Pr(c|a, e, b)\Pr(a|e, b)\Pr(e|b)\Pr(b).$$



By Decomposition:

$$\begin{aligned}\Pr(r|c, a, e, b) &= \Pr(r|e)\\ \Pr(c|a, e, b) &= \Pr(c|a)\\ \Pr(e|b) &= \Pr(e).\end{aligned}$$

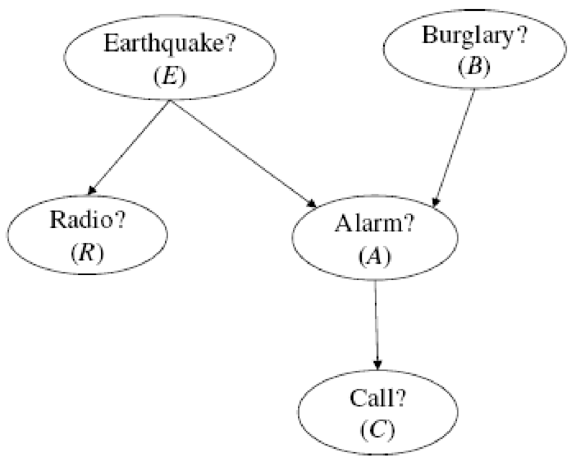
This leads to the chain rule of Bayesian networks:

$$\begin{aligned}\Pr(r, c, a, e, b) &= \Pr(r|e)\Pr(c|a)\Pr(a|e, b)\Pr(e)\Pr(b)\\ &= \theta_{r|e}\ \theta_{c|a}\ \theta_{a|e,b}\ \theta_e\ \theta_b.\end{aligned}$$

Decomposition

The variable ordering c, a, r, b, e gives

$$\Pr(c, a, r, b, e) = \Pr(c|a, r, b, e)\Pr(a|r, b, e)\Pr(r|b, e)\Pr(b|e)\Pr(e)$$



By Decomposition:

$$\begin{aligned}\Pr(c|a, r, b, e) &= \Pr(c|a)\\ \Pr(a|r, b, e) &= \Pr(a|b, e)\\ \Pr(r|b, e) &= \Pr(r|e)\\ \Pr(b|e) &= \Pr(b)\end{aligned}$$

Hence,

$$\begin{aligned}\Pr(c, a, r, b, e) &= \Pr(c|a)\Pr(a|b, e)\Pr(r|e)\Pr(b)\Pr(e)\\ &= \theta_{c|a}\ \theta_{a|b,e}\ \theta_{r|e}\ \theta_b\ \theta_e.\end{aligned}$$

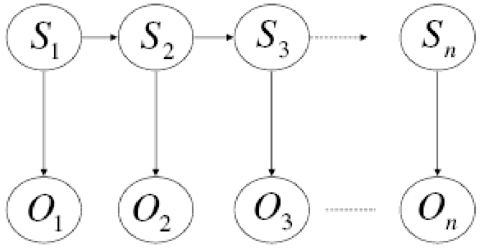
Decomposition

The variable ordering $o_n, \ldots, o_1, s_n, \ldots, s_1$ gives

$\Pr(o_n, \ldots, o_1, s_n, \ldots, s_1) =$
$\Pr(o_n | o_{n-1} \ldots, o_1, s_n, \ldots, s_1) \ldots \Pr(o_1 | s_n, \ldots, s_1) \Pr(s_n | s_{n-1} \ldots, s_1) \ldots \Pr(s_1)$

By Decomposition:

$$\begin{aligned} &\Pr(o_n, \ldots, o_1, s_n, \ldots, s_1) \\ &= \Pr(o_n | s_n) \ldots \Pr(o_1 | s_1) \Pr(s_n | s_{n-1}) \ldots \Pr(s_1) \\ &= \theta_{o_n | s_n} \ldots \theta_{o_1 | s_1} \, \theta_{s_n | s_{n-1}} \ldots \theta_{s_1}. \end{aligned}$$

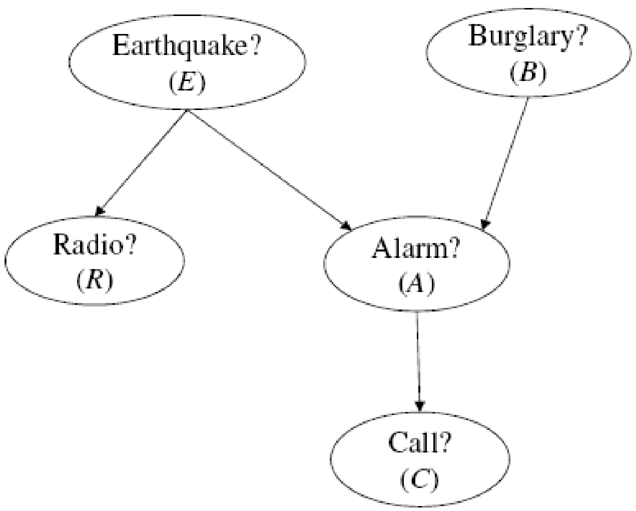


Hence, we were able again to express $\Pr(o_n, \ldots, o_1, s_n, \ldots, s_1)$ as a product of network parameters.

Weak Union

$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W}) \text{ only if } I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$$

If the information **yw** is not relevant to our belief in **x**, then the partial information **y** will not make the rest of the information, **w**, relevant.



$I(C, A, \{B, E, R\})$ is part of $\mathrm{Markov}(G)$. By Weak Union: $I_{\Pr}(C, \{A, B, E\}, R)$, which is not part of the independencies declared by $\mathrm{Markov}(G)$.

Weak Union

An implication of Weak Union

$$I_{\mathrm{Pr}}(X, \mathrm{Parents}(X) \cup \mathbf{W}, \mathrm{Non\_Descendants}(X) \setminus \mathbf{W}),$$

for any $\mathbf{W} \subseteq \mathrm{Non\_Descendants}(X)$.

- Each variable X in DAG G is independent of any of its non-descendants given its parents and the remaining non-descendants.
- This can be viewed as a strengthening of the independencies declared by $\mathrm{Markov}(G)$, which fall as a special case when the set $\mathbf{W}$ is empty.

Contraction

$$I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W}) \text{ ~~only if~~ } \Longrightarrow I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

If after learning the irrelevant information **y**, the information **w** is found to be irrelevant to our belief in **x**, then the combined information **yw** must have been irrelevant from the beginning.

Compare Contraction with Composition:

$$I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{W}) \text{ ~~only if~~ } \Longrightarrow I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

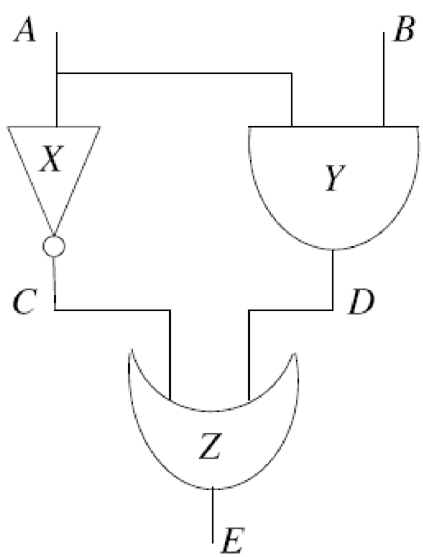
One can view Contraction as a weaker version of Composition. Recall that Composition does not hold for probability distributions.

Strictly Positive Distributions

When there are no constraints

Definition

A strictly positive distribution assign a non-zero probability to every consistent event.



Example

A strictly positive distribution cannot represent the behavior of Inverter X as it will have to assign the probability zero to the event A=true, C=true.

A strictly positive distribution cannot capture logical constraints.

Intersection

Holds only for strictly positive distributions

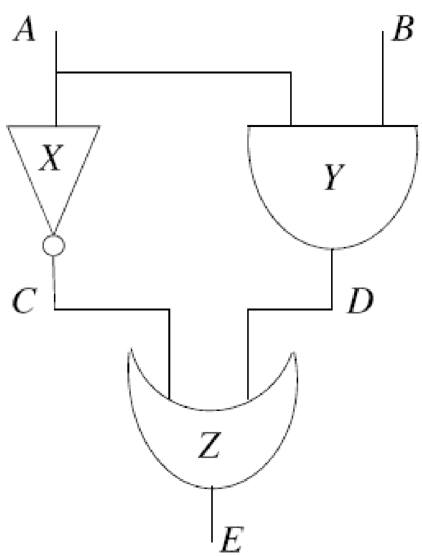
$I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$ and $I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$ only if $I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$
If information **w** is irrelevant given **y**, and **y** is irrelevant given **w**, then combined information **yw** is irrelevant to start with.

Intersection

Holds only for strictly positive distributions

$I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$ and $I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$ only if $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$

If information **w** is irrelevant given **y**, and **y** is irrelevant given **w**, then combined information **yw** is irrelevant to start with.



- If we know the input A of inverter X, its output C becomes irrelevant to our belief in the circuit output E.
- If we know the output C of inverter X, its input A becomes irrelevant to this belief.
- Yet, variables A and C are not irrelevant to our belief in the circuit output E.

Properties of Probabilistic Independence

Triviality: $I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \emptyset)$.

Symmetry, Decomposition, Weak Union, and Contraction, combined with Triviality, are known as the graphoid axioms.

With Intersection, the set is known as the positive graphoid axioms.

- Decomposition, Weak Union, and Contraction can be summarized tersely in one statement:

$$I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W}) \text{ iff } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$$

- The terms semi-graphoid and graphoid are sometimes used instead of graphoid and positive graphoid, respectively.

Properties of Probabilistic independence

THEOREM 1: Let X, Y, and Z be three disjoint subsets of variables from U. If $I(X, Z, Y)$ stands for the relation "X is independent of Y, given Z" in some probabilistic model P, then I must satisfy the following four independent conditions:

- Symmetry:
  - **I(X,Z,Y) → I(Y,Z,X)**
- Decomposition:
  - **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)→I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)→I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) → I(X,Z,YW)**

Graphoid axioms:
Symmetry, decomposition
Weak union and contraction

Positive graphoid:
+intersection

In Pearl: the 5 axioms
are called Graphids,
the 4, semi-graphois

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs
  - I-maps, D-maps, perfect maps
  - Markov boundary and blanket
  - Markov networks

A Graphical Test of Independence

The inferential power of the graphoid axioms can be tersely captured using a graphical test, known as d-separation, which allows one to mechanically, and efficiently, derive the independencies implied by these axioms.

- To test whether $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in DAG G, written $\text{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, we need to consider every path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$, and then ensure that the path is blocked by $\mathbf{Z}$.
- The definition of d-separation relies on the notion of blocking a path by a set of variables $\mathbf{Z}$.

$\text{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ implies $I_{\text{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ for every probability distribution Pr induced by G.

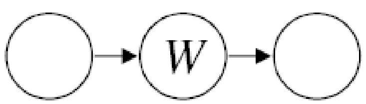
d-speration

- To test whether **X** and **Y** are d-separated by **Z** in dag G, we need to consider every path between a node in **X** and a node in **Y**, and then ensure that the path is blocked by **Z.**
- A path is blocked by **Z** if **at least** one valve (node) on the path is 'closed' given **Z**.
- A divergent valve or a sequential valve is closed if it is in **Z**
- A convergent valve is closed if it is not on **Z** nor any of its descendants are in **Z.**

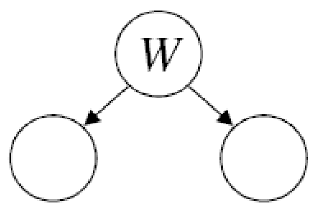
d-separation

> The type of a valve is determined by its relationship to its neighbors on the path.

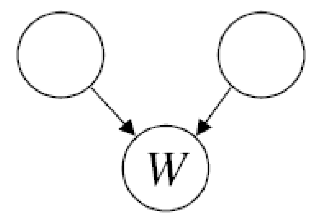
sequential



divergent

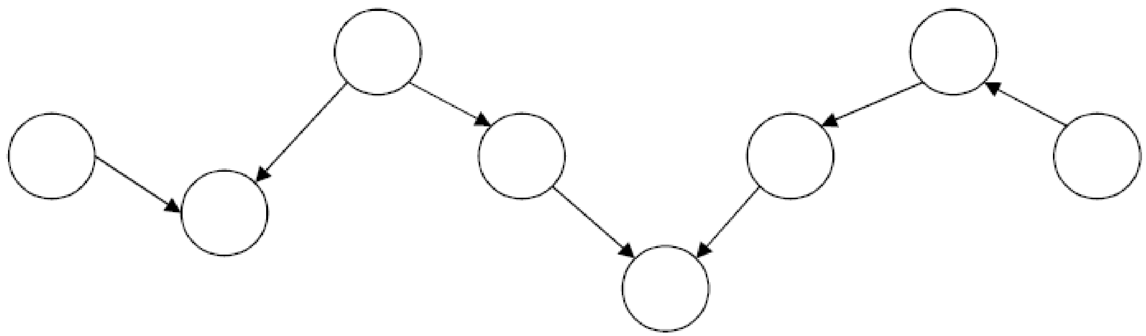


convergent



- A sequential valve $\rightarrow W \rightarrow$ arises when W is a parent of one of its neighbors and a child of the other.
- A divergent valve $\leftarrow W \rightarrow$ arises when W is a parent of both neighbors.
- A convergent valve $\rightarrow W \leftarrow$ arises when W is a child of both neighbors.

d-separation



Example

A path with 6 valves. From left to right, convergent, divergent, sequential, convergent, sequential, and sequential.

d-separation

Definition

Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint sets of nodes in a DAG G. We will say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$, written $\mathrm{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, iff every path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$ is blocked by $\mathbf{Z}$, where a path is blocked by $\mathbf{Z}$ iff at least one valve on the path is closed given $\mathbf{Z}$.

A path with no valves (i.e., $X \rightarrow Y$) is never blocked.

DEPENDENCE SEMANTICS FOR BAYESIAN NETWORKS

DEFINITION: If X, Y, and Z are three disjoint subsets of nodes in a DAG D, then Z is said to ***d-separate*** X from Y, denoted $< X \mid Z \mid Y >_D$, if there is no path between a node in X and a node in Y along which the following two conditions hold: (1) every node with converging arrows is in Z or has a descendent in Z and (2) every other node is outside Z.

No path
Is active =
Every path is
blocked

- If a path satisfies the condition above, it is said to be *active;* otherwise, it is said to be *blocked* by Z.

$$<2 \mid 1 \mid 3>_D \quad , \quad \neg <2 \mid 15 \mid 3>_D$$

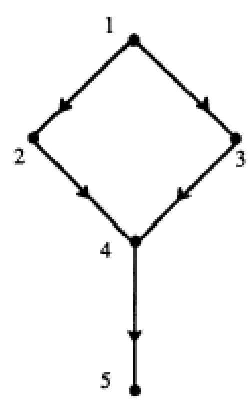
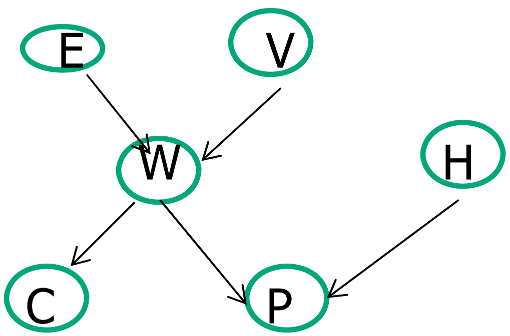


Figure 3.10. A DAG depicting *d-separation*; node 1 blocks the path 2-1-3, while node 5 activates the path 2-4-3.

Bayesian Networks as i-maps

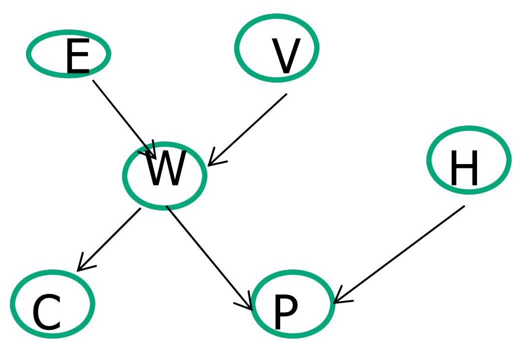
- E: Employment
- V: Investment
- H: Health
- W: Wealth
- C: Charitable contributions
- P: Happiness



Are C and V d-separated give E and P?
Are C and H d-separated?

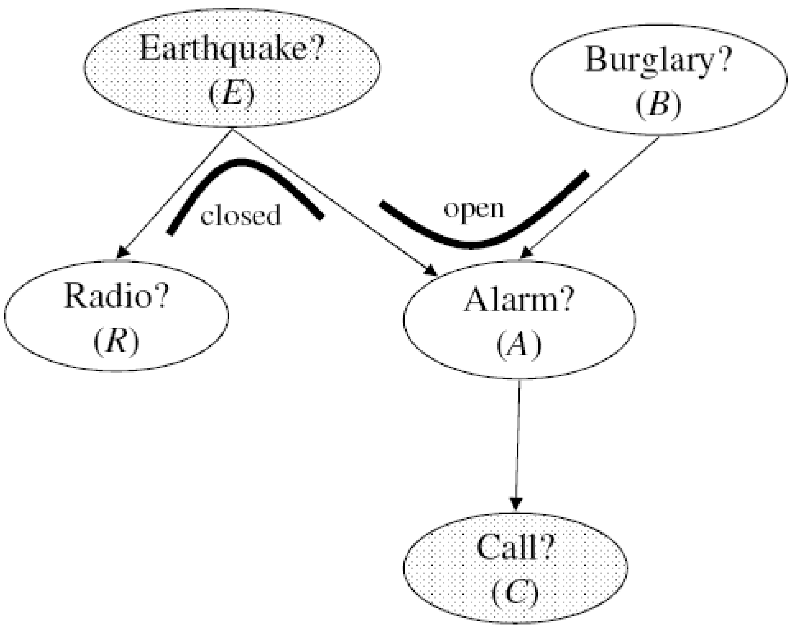
d-Seperation Using Ancestral Graph

- X is d-separated from Y given Z (<X,Z,Y>d) iff:
  - Take the ancestral graph that contains **X,Y,Z** and their ancestral subsets.
  - Moralized the obtained subgraph
  - Apply regular undirected graph separation
  - Check: (E,{},V),(E,P,H),(C,EW,P),(C,E,HP)?

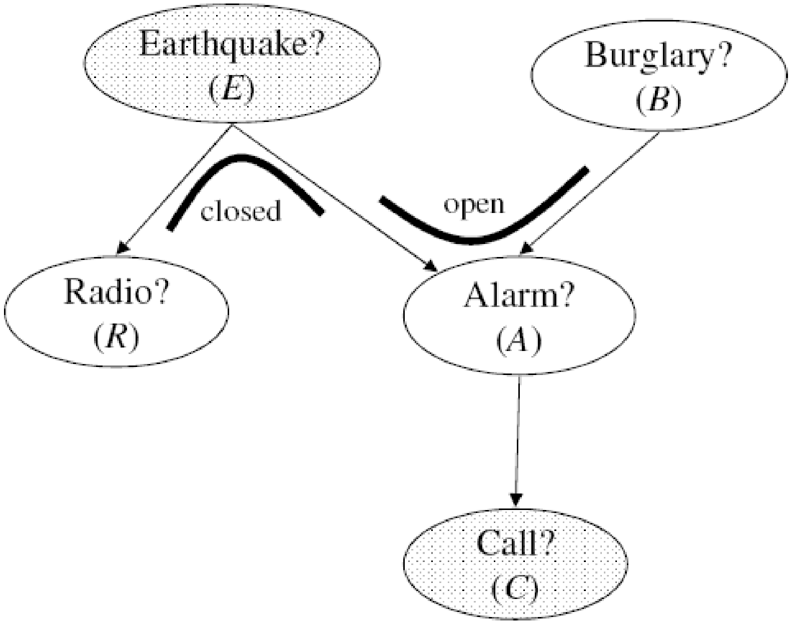


d-separation

$I_{dsep}(R,EC,B)$?



d-separation

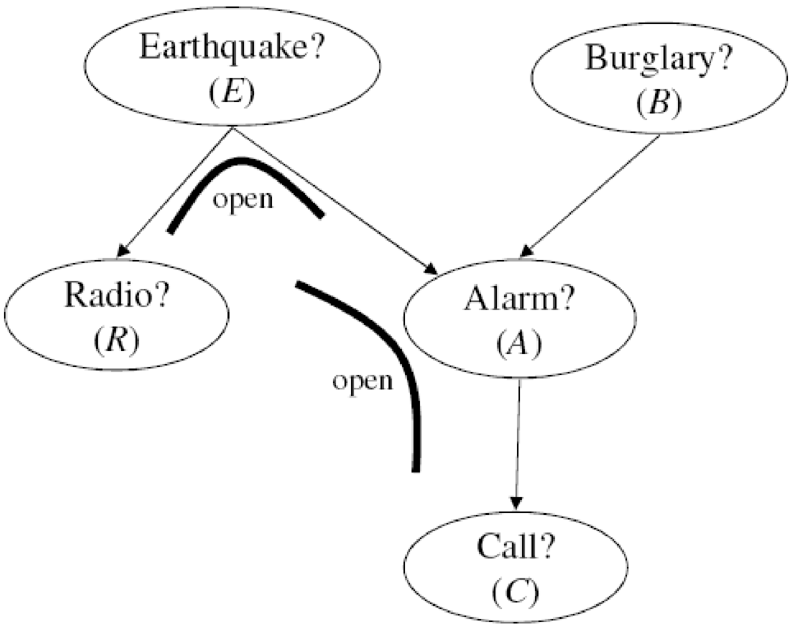


Example

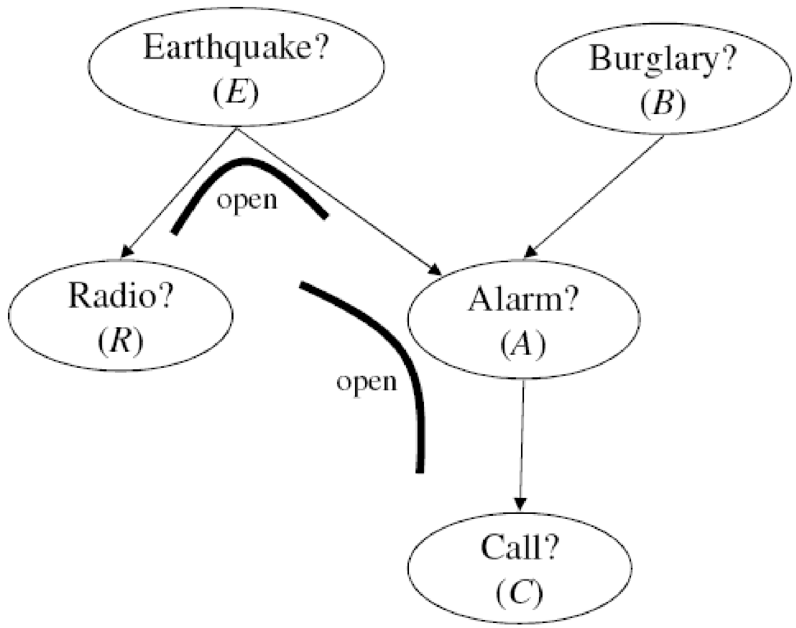
R and B are d-separated by E and C. The closure of only one valve is sufficient to block the path, therefore, establishing d-separation.

d-separation

$I_{dsep}(R,\emptyset,C)$?



d-separation

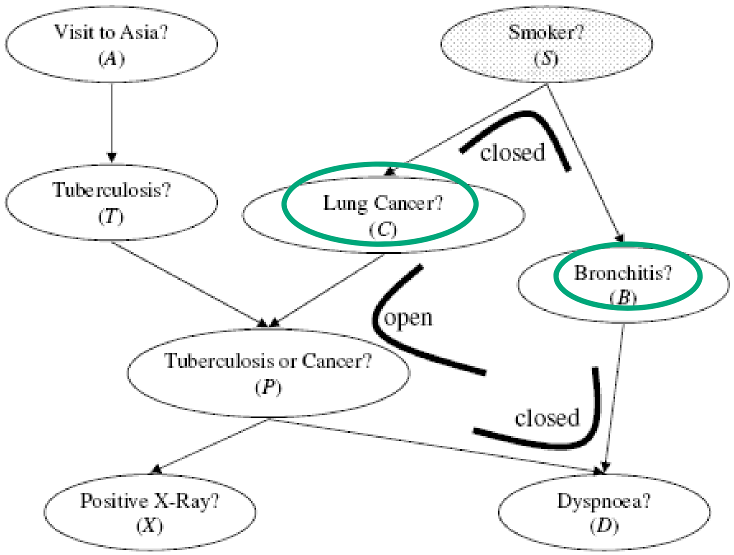


Example

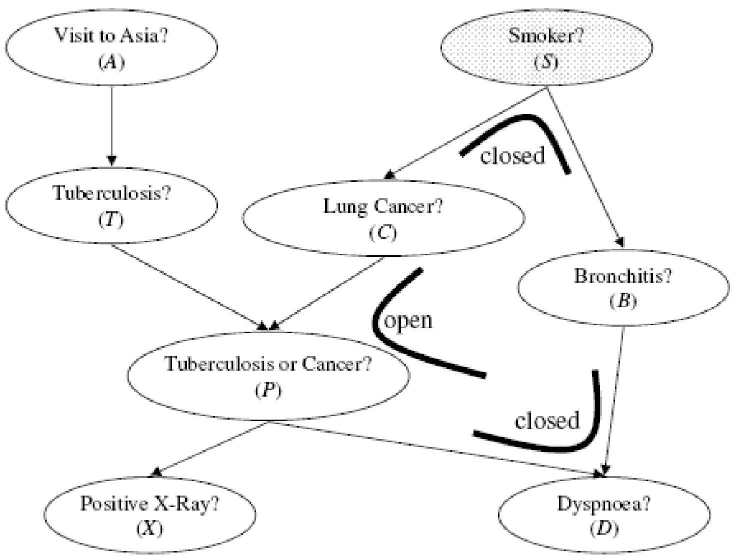
R and C are not d-separated since both valves are open. Hence, the path is not blocked and d-separation does not hold.

d-separation

$I_{dsep}(\mathbf{C},\mathbf{S},\mathbf{B})$=?



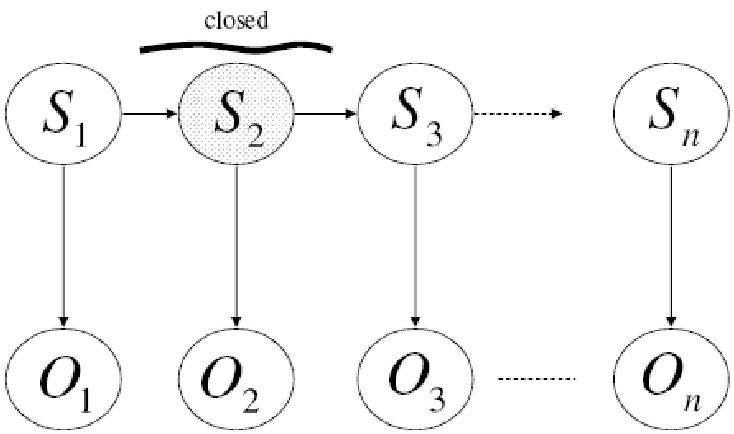
d-separation



Example

C and B are d-separated by S since both paths between them are blocked by S.

d-separation

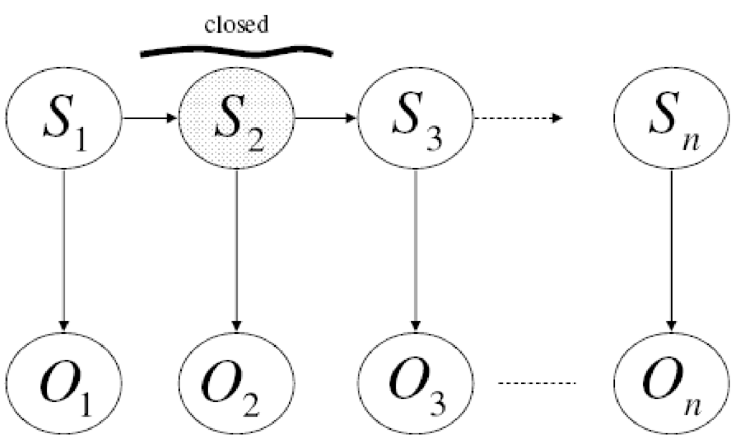


Exam

$I_{\Pr}$

$I_{\Pr}(S_1, S_2, \{S_3, S_4\})$ for any probability distribution $\Pr$ which is induced by the DAG.

d-separation

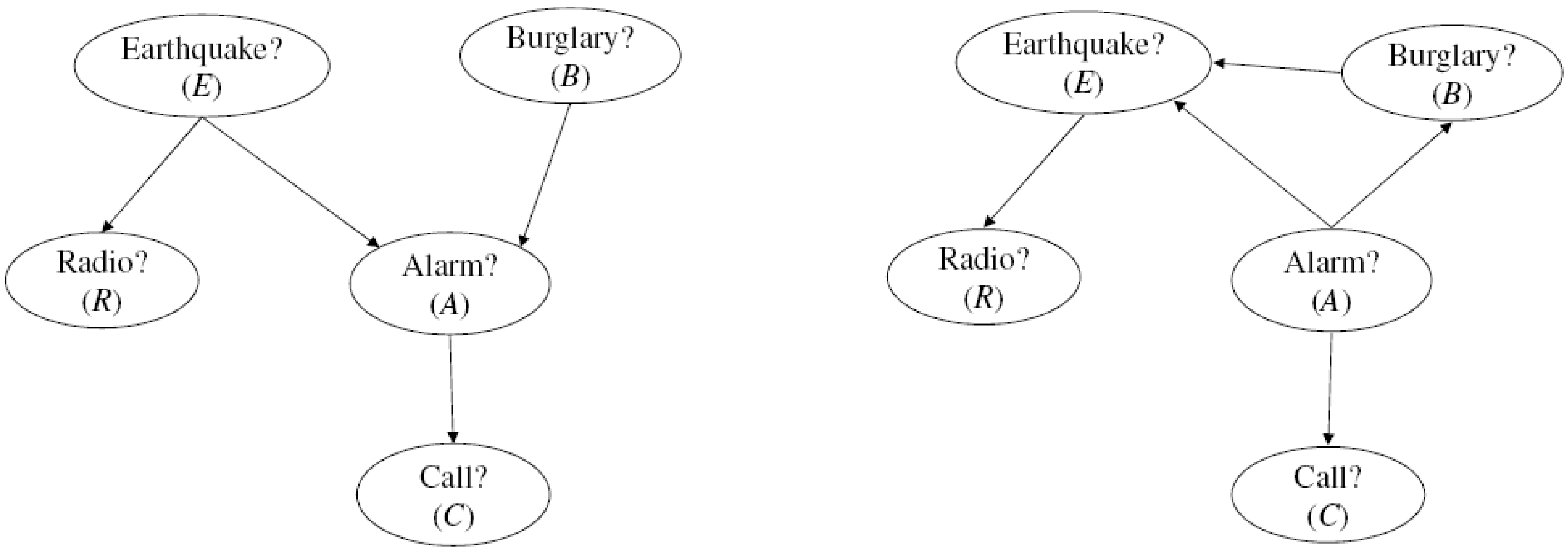


Example

Any path between S_1 and $\{S_3, S_4\}$ must have the valve $S_1 \rightarrow S_2 \rightarrow S_3$ on it, which is closed given S_2. Hence, every path from S_1 to $\{S_3, S_4\}$ is blocked by S_2, and we have $\mathrm{dsep}_G(S_1, S_2, \{S_3, S_4\})$, which leads to $I_{\mathrm{Pr}}(S_1, S_2, \{S_3, S_4\})$.

$I_{\mathrm{Pr}}(S_1, S_2, \{S_3, S_4\})$ for any probability distribution Pr which is induced by the DAG.

Capturing Independence Graphically



Every independence which is declared (or implied) by the second DAG is also declared (or implied) by the first one. Hence, if we accept the first DAG, then we must also accept the second.

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs
  - Soundness, completeness of d-seperation
  - I-maps, D-maps, perfect maps
  - Construction a minimal I-map of a distribution
  - Markov boundary and blanket

Soundness of d-separation

The d-separation test is sound in the following sense.

Theorem

If Pr *is a probability distribution induced by a Bayesian network* (G, Θ)*, then*

$$\mathrm{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ only if } I_{\mathrm{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}).$$

The proof of soundness is constructive, showing that every independence claimed by d-separation can indeed be derived using the graphoid axioms.

Completeness of d-separation

It is not a d-map

d-separation is not complete in the following sense:

- Consider a network with three binary variables $X \rightarrow Y \rightarrow Z$.
- Z is not d-separated from X.
- Z can be independent of X in a probability distribution induced by this network.

Example

Choose the CPT for variable Y so that $\theta_{y|x} = \theta_{y|\bar{x}}$.
Y independent of X since

- $\Pr(y) = \Pr(y|x) = \Pr(y|\bar{x})$ and
- $\Pr(\bar{y}) = \Pr(\bar{y}|x) = \Pr(\bar{y}|\bar{x})$.

Z is also independent of X.

BAYESIAN NETWORKS AS I-MAPS

DEFINITION: A DAG D is said to be an *I-map* of a dependency model M if every d-separation condition displayed in D corresponds to a valid conditional independence relationship in M, i.e., if for every three disjoint sets of vertices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ we have

$$<X \mid Z \mid Y>_D \;\Rightarrow\; I(X, Z, Y)_M .$$

- A DAG is a *minimal* I-map of M if none of its arrows can be deleted without destroying its I-mapness.

DEFINITION: Given a probability distribution P on a set of variables U, a DAG $D = (U, \vec{E})$ is called a *Bayesian network* of P iff D is a minimal I-map of P.

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs
  - Soundness, completeness of d-seperation
  - I-maps, D-maps, perfect maps
  - Construction a minimal I-map of a distribution
  - Markov boundary and blanket

More on DAGs and Independence

Definition

G is an Independence MAP (I-MAP) of $\Pr$ iff every independence declared by d-separation on DAG G holds in the distribution $\Pr$:

$$\text{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ only if } I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}).$$

Definition

An I-MAP G is minimal if G ceases to be an I-MAP when we delete any edge from G.

By the semantics of Bayesian networks, if $\Pr$ is induced by a Bayesian network (G, Θ), then G must be an I-MAP of $\Pr$, although it may not be minimal.

More on DAGs and Independence

Definition

G is a Dependency MAP (D-MAP) of $\Pr$ iff

$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ only if } \mathrm{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y}).$$

If G is a D-MAP of $\Pr$, then the lack of d-separation in G implies a dependence in $\Pr$.

Definition

If DAG G is both an I-MAP and a D-MAP of distribution $\Pr$, then G is called a Perfect MAP (P-MAP) of $\Pr$.

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs
  - Soundness, completeness of d-seperation
  - I-maps, D-maps, perfect maps
  - Construction a minimal I-map of a distribution
  - Markov boundary and blanket

CONSTRUCTING A BAYESIAN NETWORK FOR ANY GIVEN DISTRIBUTION P

DEFINITION: Let M be a dependency model defined on a set $\boldsymbol{U} = \{X_1, X_2, ..., X_n\}$ of elements, and let d be an ordering $(X_1, X_2, ..., X_i, ...)$ of the elements of $\boldsymbol{U}$.

- The **boundary strata** of M relative to d is an ordered set of subsets of $\boldsymbol{U}$, $(\boldsymbol{B}_1, \boldsymbol{B}_2, ..., \boldsymbol{B}_i, ...)$, such that each $\boldsymbol{B}_i$ is a Markov boundary of X_i with respect to the set $\boldsymbol{U}_{(i)} = \{X_1, X_2, ..., X_{i-1}\}$, i.e., $\boldsymbol{B}_i$ is a minimal set satisfying $\boldsymbol{B}_i \subseteq \boldsymbol{U}_{(i)}$ and $I(X_i, \boldsymbol{B}_i, \boldsymbol{U}_{(i)} - \boldsymbol{B}_i)$.

- The DAG created by designating each $\boldsymbol{B}_i$ as parents of vertex X_i is called a *boundary DAG* of M relative to d.

THEOREM 9: [Verma 1986]: *Let* M be any semi-graphoid (i.e., any dependency model satisfying the axioms of Eqs. (3.6*a*) through (3.6*d*)). If D is a boundary DAG of M relative to any ordering d, then D is a minimal I-map of M.

Independence MAPs

Given a distribution $\Pr$, how can we construct a DAG G which is guaranteed to be a minimal I-MAP of $\Pr$?

Given an ordering $X_1, \ldots, X_n$ of the variables in $\Pr$:

- Start with an empty DAG G (no edges)
- Consider the variables X_i one by one, for $i = 1, \ldots, n$.
- For each variable X_i, identify a minimal subset $\mathbf{P}$ of the variables in $X_1, \ldots, X_{i-1}$ such that

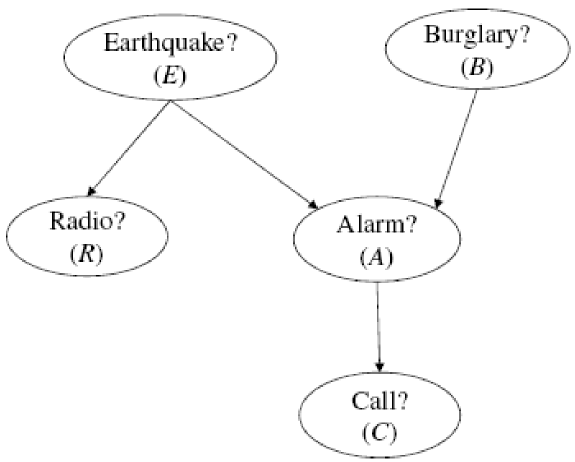
$$I_{\Pr}(X_i, \mathbf{P}, \{X_1, \ldots, X_{i-1}\} \setminus \mathbf{P}).$$

- Make $\mathbf{P}$ the parents of X_i in DAG G.

The resulting DAG is a minimal I-MAP of $\Pr$.

Independence MAPs

Construct a minimal I-MAP G for some distribution $\Pr$ using the previous procedure and variable order A, B, C, E, R.



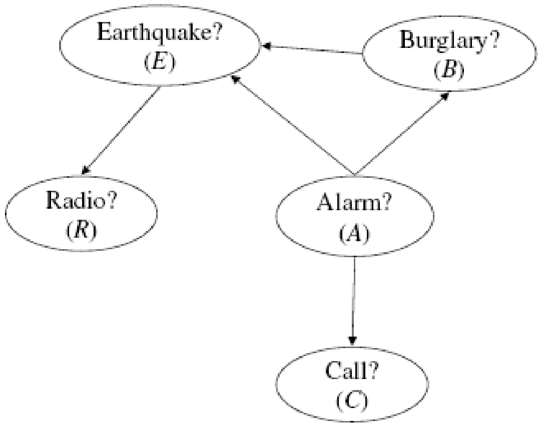
Suppose that DAG G' is a P-MAP of distribution $\Pr$

Independence tests on $\Pr$, $I_{\Pr}(X_i, \mathbf{P}, \{X_1, \ldots, X_{i-1}\} \setminus \mathbf{P})$, can now be reduced to equivalent d-separation tests on DAG G', $\text{dsep}_{G'}(X_i, \mathbf{P}, \{X_1, \ldots, X_{i-1}\} \setminus \mathbf{P})$.

Independence MAPs

This minimal I-MAP G is constructed according to the following details:

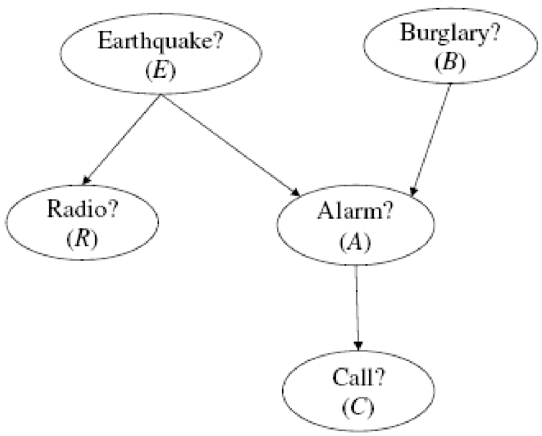
- Variable A added with $\mathbf{P} = \emptyset$.



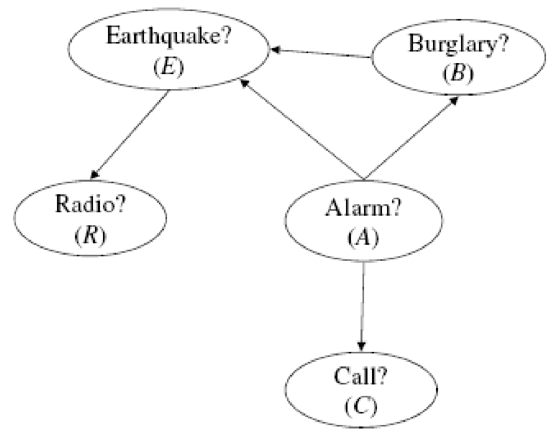
- Variable B added with $\mathbf{P} = A$, since $\mathrm{dsep}_{G'}(B, A, \emptyset)$ holds and $\mathrm{dsep}_{G'}(B, \emptyset, A)$ does not.
- Variable C added with $\mathbf{P} = A$, since $\mathrm{dsep}_{G'}(C, A, B)$ holds and $\mathrm{dsep}(C, \emptyset, \{A, B\})$ does not.
- Variable E added with $\mathbf{P} = A, B$ since this is the smallest subset of A, B, C such that $\mathrm{dsep}_{G'}(E, \mathbf{P}, \{A, B, C\} \setminus \mathbf{P})$ holds.
- Variable R added with $\mathbf{P} = E$ since this is the smallest subset of A, B, C, E such that $\mathrm{dsep}_{G'}(R, \mathbf{P}, \{A, B, C, E\} \setminus \mathbf{P})$ holds.

Independence MAPs

DAG G' and distribution $\Pr$



Minimal I-MAP G



- If $\text{dsep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, then $\text{dsep}_{G'}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$.
- This ceases to hold if we delete any of the five edges in G.

For example, if we delete the edge $E \leftarrow B$, we will have $\text{dsep}_G(E, A, B)$, yet $\text{dsep}_{G'}(E, A, B)$ does not hold.

Independence MAPs

- The minimal I-MAP of a distribution is not unique, as we may get different ones depending on which variable ordering we start with.
- Even when using the same variable ordering, it is possible to arrive at different minimal I-MAPs. This is possible since we may have multiple minimal subsets **P** of $\{X_1, \ldots, X_{i-1}\}$ for which $I_{\mathrm{Pr}}(X_i, \mathbf{P}, \{X_1, \ldots, X_{i-1}\} \setminus \mathbf{P})$ holds.
- This can only happen if the probability distribution Pr represents some logical constraints.
- We can ensure the uniqueness of a minimal I-MAP for a given variable ordering if we restrict ourselves to strictly positive distributions.

Perfect Maps for DAGs

- Theorem 10 [Geiger and Pearl 1988]: For any dag D there exists a P such that D is a perfect map of P relative to d-separation.

- Corollary 7: d-separation identifies any implied independency that follows logically from the set of independencies characterized by its dag.

Outline

- Basics of probability theory
- DAGS, Markov(G), Bayesian networks
- Graphoids: axioms of for inferring conditional independence (CI)
- D-separation: Inferring CIs in graphs
  - Soundness, completeness of d-seperation
  - I-maps, D-maps, perfect maps
  - Construction a minimal I-map of a distribution
  - Markov boundary and blanket

Blankets and Boundaries

Definition

Let $\Pr$ be a distribution over variables $\mathbf{X}$. A Markov blanket for a variable $X \in \mathbf{X}$ is a set of variables $\mathbf{B} \subseteq \mathbf{X}$ such that $X \notin \mathbf{B}$ and $I_{\Pr}(X, \mathbf{B}, \mathbf{X} \setminus \mathbf{B} \setminus \{X\})$.

A Markov blanket for X is a set of variables which, when known, will render every other variable irrelevant to X.

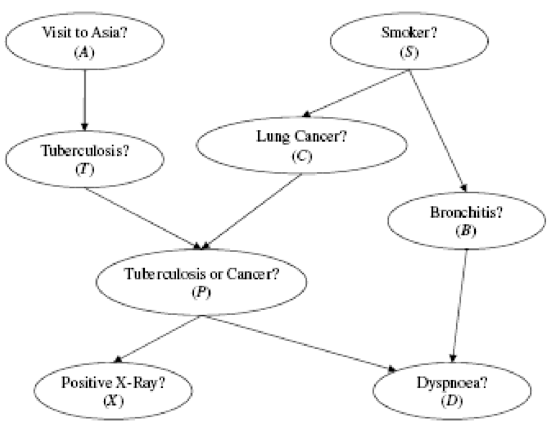
Definition

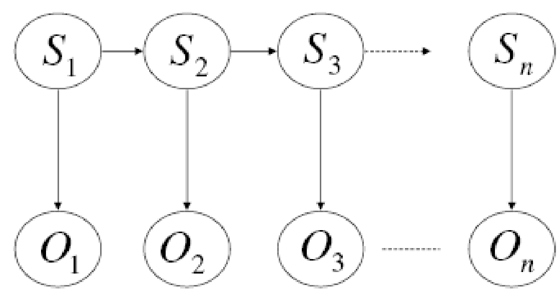
A Markov blanket $\mathbf{B}$ is minimal iff no strict subset of $\mathbf{B}$ is also a Markov blanket. A minimal Markov blanket is a Markov Boundary.

The Markov Boundary for a variable is not unique, unless the distribution is strictly positive.

Blanket Examples

If $\Pr$ is induced by DAG G, then a Markov blanket for variable X with respect to $\Pr$ can be constructed using its parents, children, and spouses in DAG G. Here, variable Y is a spouse of X if the two variables have a common child in DAG G.

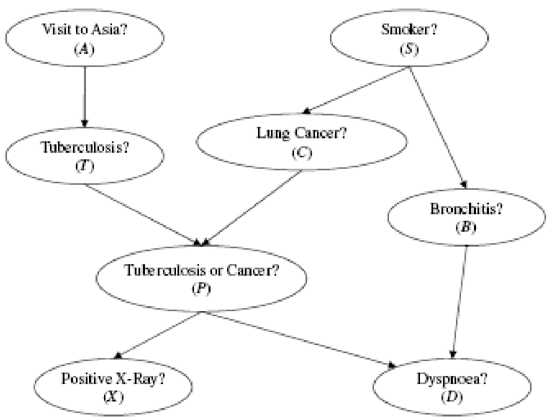




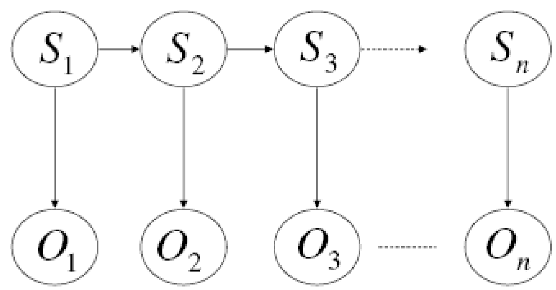
$\{S_{t-1}, S_{t+1}, O_t\}$ is a Markov blanket for every variable S_t, where $t > 1$

Blanket Examples

If $\Pr$ is induced by DAG G, then a Markov blanket for variable X with respect to $\Pr$ can be constructed using its parents, children, and spouses in DAG G. Here, variable Y is a spouse of X if the two variables have a common child in DAG G.



$\{S, P, T\}$ is a Markov blanket for variable C



$\{S_{t-1}, S_{t+1}, O_t\}$ is a Markov blanket for every variable S_t, where $t > 1$

Bayesian Networks as Knowledge-Bases

- Given any distribution, P, and an ordering we can construct a minimal i-map.
- The conditional probabilities of x given its parents is all we need.
- In practice we go in the opposite direction: the parents must be identified by human expert… they can be viewed as direct causes, or direct influences.

BAYESIAN NETWORK AS A KNOWLEDGE BASE

STRUCTURING THE NETWORK

- Given any joint distribution $P(x_1, ..., x_n)$ and an ordering d of the variables in U, Corollary 4 prescribes a simple recursive procedure for constructing a Bayesian network.

- Choose X_1 as a root and assign to it the marginal probability $P(x_1)$ dictated by $P(x_1, ..., x_n)$.

- If X_2 is dependent on X_1, a link from X_1 to X_2 is established and quantified by $P(x_2 | x_1)$. Otherwise, we leave X_1 and X_2 unconnected and assign the prior probability $P(x_2)$ to node X_2.

- At the i-th stage, we form the node X_i, draw a group of directed links to X_i from a parent set Π_{X_i} defined by Eq. (3.27), and quantify this group of links by the conditional probability $P(x_i | \pi_{X_i})$.

- The result is a directed acyclic graph that represents all the independencies that follow from the definitions of the parent sets.

- In practice, $P(x_1,\ldots,x_n)$ is not available.

- The parent sets Π_{X_i} must be identified by human judgment.

- To specify the strengths of influences, assess the conditional probabilities $P(x_i \mid \pi_{X_i})$ by some functions $F_i(x_i, \pi_{X_i})$ and make sure these assessments satisfy

$$\sum_{x_i} F_i(x_i, \pi_{X_i}) = 1 , \tag{3.30}$$

where $0 \le F_i(x_i, \pi_{X_i}) \le 1$

- This specification is complete and consistent because the product form

$$P_a(x_1, \ldots, x_n) = \prod_i F_i(x_i, \pi_{X_i}) \tag{3.31}$$

constitutes a joint probability distribution that supports the assessed quantities.

$$P_a(x_i \mid \pi_{X_i}) = \frac{P_a(x_i, \pi_{X_i})}{P_a(\pi_{X_i})} = \frac{\sum_{x_j \notin (x_i \cup \Pi_{X_i})} P_a(x_1,\ldots,x_n)}{\sum_{x_j \notin \Pi_{X_i}} P_a(x_1,\ldots,x_n)} = F_i(x_i, \pi_{X_i}) \tag{3.32}$$

- DAGs constructed by this method will be called *Bayesian belief networks* or *causal networks* interchangeably.

Markov Networks and Markov Random Fields (MRF)

Can we also capture conditional independence by undirected graphs?

Yes: using simple graph separation

Graphoids

- Symmetry:
  - **I(X,Z,Y) → I(Y,Z,X)**
- Decomposition:
  - **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)→I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)→I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) → I(X,Z,YW)**

Undirected Graphs as I-maps of Distributions

-
- We say $<X,Z,Y>_G$ iff once you remove Z from the graph X and Y are not connected
- Can we completely capture probabilistic independencies by the notion of separation in a graph?
- Example: 2 coins and a bell.

Axiomatic Characterization of Graphs

- **Graph separation satisfies:**
  - Symmetry: **I(X,Z,Y) → I(Y,Z,X)**
  - Decomposition: **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,Y)**
  - Intersection: **I(X,ZW,Y)** and **I(X,ZY,W)→I(X,Z,YW)**
  - Strong union: I(**X,Z,Y) →** I**(X,ZW, Y)**
  - Transitivity: I(**X,Z,Y**) → exists t s.t. I(**X,Z**,t) or I(t,**Z,Y**)

Graphoids vs Undirected graphs

- Symmetry:
  - **I(X,Z,Y) → I(Y,Z,X**)
- Decomposition:
  - **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)→I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)→I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) → I(X,Z,YW**)

Symmetry: **I(X,Z,Y) → I(Y,Z,X**)

Decomposition: **I(X,Z,YW)→ I(X,Z,Y) and I(X,Z,Y)**

Intersection: **I(X,ZW,Y)** and **I(X,ZY,W)→I(X,Z,YW)**

Strong union: I(**X,Z,Y) →** I**(X,ZW, Y)**

Transitivity: I(**X,Z,Y**) → exists t s.t. I(**X,Z**,t) or I(t,**Z,Y**)

Markov Networks

- An undirected graph G which is a minimal I-map of a probability distribution Pr, namely deleting any edge destroys its i-mappness relative to (undirected) seperation, is called a **Markov network of P**.

CONCEPTUAL DEPENDENCIES AND THEIR MARKOV NETWORKS

- An agent identifies the following variables as having influence on the main question of being late to a meeting:

    1. The time shown on the watch of Passerby 1.
    2. The time shown on the watch of Passerby 2.
    3. The correct time.
    4. The time it takes to travel to the meeting place.
    5. The arrival time at the meeting place.

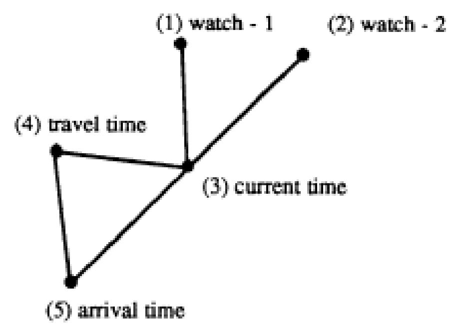
- The construction of G_0 can proceed by one of two methods:
    - The *edge-deletion* method.
    - The *Markov boundary* method.

- The first method requires that for every pair of variables (α, β) we determine whether fixing the values of all other variables in the system will render our belief in α sensitive to β.

- For example, the reading on Passerby 1's watch (1) will vary with the actual time (3) even if all other variables are known, so connect node 1 to node 3

- The Markov boundary method requires that for every variable α in the system, we identify a minimal set of variables sufficient to render the belief in α insensitive to all other variables in the system.

- For instance, once we know the current time (3), no other variable can affect what we expect to read on passerby 1's watch (1).



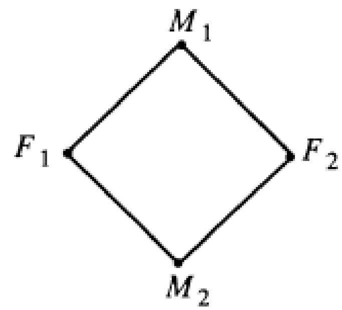
The unusual edge (3,4) reflects the reasoning that if we fix the arrival time (5) the travel time (4) must depends on current time (3)

Figure 3.6. The Markov network representing the prediction of A's arrival time.

- G_0 can be used as an inference instrument.

  - For example, knowing the current time (3) renders the time on Passerby 1's watch (1) irrelevant for estimating the travel time (4) (i.e., $I(1,3,4)$); 3 is a cutset in G_0, separating 1 from 4.

MARKOV NETWORK AS A KNOWLEDGE BASE



How can we construct a probability Distribution that will have all these independencies?

Figure 3.2. An undirected graph representing interactions among four individuals.

QUANTIFYING THE LINKS

- If couple (M_1, F_2) meet less frequently than the couple (M_1, F_1), then the first link should be weaker than the second
- The model must be consistent, complete and a Markov field of G.
- Arbitrary specification of $P(M_1, F_1)$, $P(F_1, M_2)$, $P(M_2, F_2)$, and $P(F_2, M_1)$ might lead to inconsistencies.
- If we specify the pairwise probabilities of only three pairs, incompleteness will result.

Markov Random Field (MRF)

- A safe method (called *Gibbs' potential*) for constructing a complete and consistent quantitative model while preserving the dependency structure of an arbitrary graph G.

1. Identify the cliques† of G, namely, the largest subgraphs whose nodes are all adjacent to each other.
2. For each clique C_i, assign a nonnegative compatibility function $g_i(c_i)$, which measures the relative degree of compatibility associated with the value assignment c_i to the variables included in C_i.
3. Form the product $\prod_i g_i(c_i)$ of the compatibility functions over all the cliques.
4. Normalize the product over all possible value combinations of the variables in the system

$$P(x_1, \ldots, x_n) = K \prod_i g_i(c_i), \tag{3.13}$$

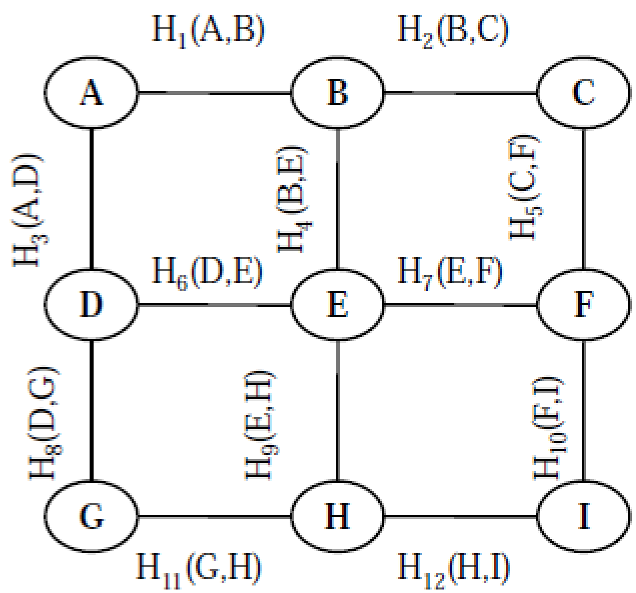
where

† We use the term *clique* for the more common term *maximal clique*.

So, How do we learn Markov networks From data?

Examples of Bayesian and and Markov Networks

Markov Networks



D	E	H_6(D,E)
0	0	20.2
0	1	12
1	0	23.4
1	1	11.7

(a) (b)

Figure 2.6: (a) An example 3×3 square Grid Markov network (ising model) and (b) An example potential $H_6(D, E)$

network represents a global joint distribution over the variables $\boldsymbol{X}$ *given by:*

$$P(\boldsymbol{x}) = \frac{1}{Z}\prod_{i=1}^{m} H_i(\boldsymbol{x}) \quad , \quad Z = \sum_{\boldsymbol{x} \in \boldsymbol{X}} \prod_{i=1}^{m} H_i(\boldsymbol{x})$$

Sample Applications for Graphical Models

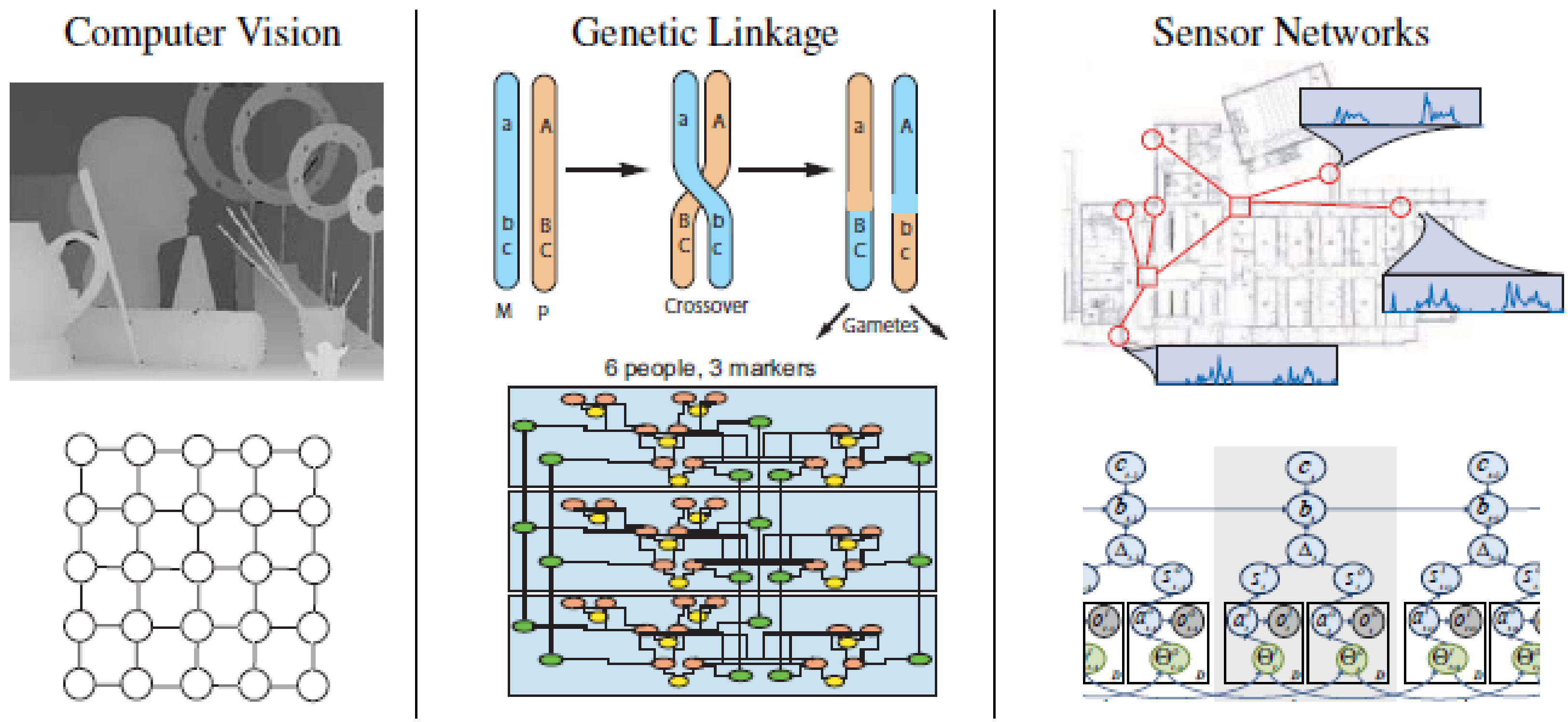


Figure 1: Application areas and graphical models used to represent their respective systems: (a) Finding correspondences between images, including depth estimation from stereo; (b) Genetic linkage analysis and pedigree data; (c) Understanding patterns of behavior in sensor measurements using spatio-temporal models.