Causal and Probabilistic Reasoning

Slides Set 2:

Rina Dechter

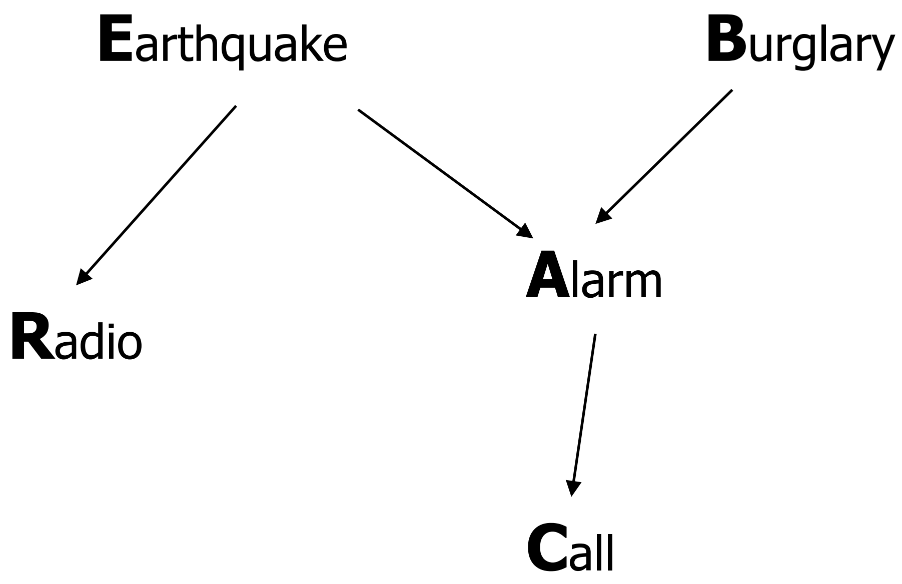
Reading:
Darwiche chapter 4
Pearl (probabilistic): chapter 3

Outline

- Basic of Probability Theory
- Bayesian Networks, DAGS, Markov(G)
- Graphoids axioms for Conditional Independence
- d-separation: Inferring CIs in graphs

Basics of Probabilistic Calculus (Chapter 3)

The Burglary Example



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:

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

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:

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

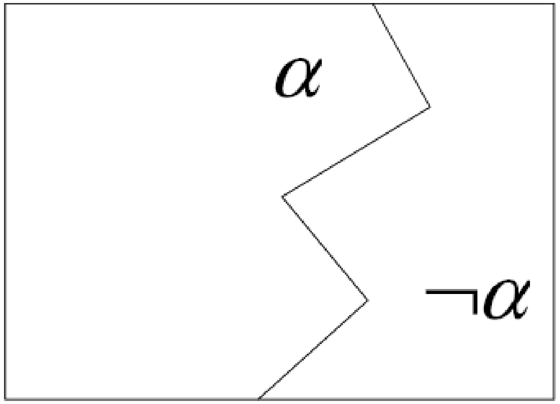
- A baseline for inconsistent sentences:

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

- A baseline for valid sentences:

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

Properties of Beliefs



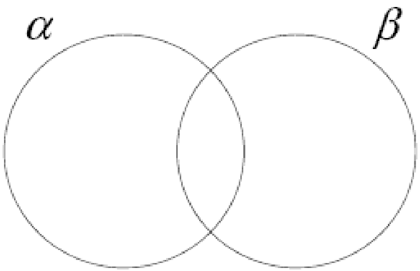
- The belief in a sentence given the belief in its negation:

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

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



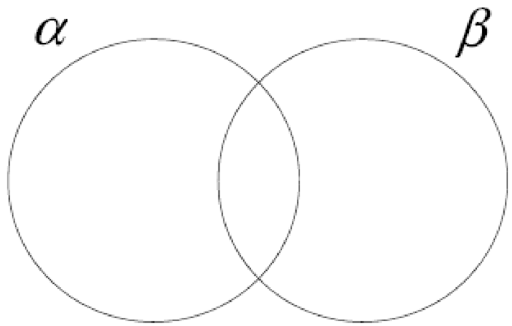
- The belief in a disjunction:

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

- 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

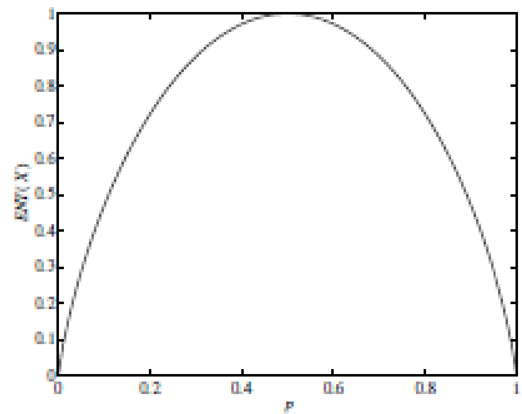
Quantify uncertainty about a variable X using the notion of entropy:

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

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

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558
$\mathrm{ENT}(.)$	.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:

$$
\begin{array}{l}
A \wedge B \text{ is independent of } C \text{ given } D \wedge E; \\
A \wedge \neg B \text{ is independent of } C \text{ given } D \wedge E; \\
\vdots \\
\neg A \wedge \neg B \text{ is independent of } \neg C \text{ given } \neg D \wedge \neg E;
\end{array}
$$

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

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

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, $\Pr(\alpha|\beta)$.

Outline

- Basic of Probability Theory
- Bayesian Networks, DAGS, Markov(G)
  - From a distribution to a BN
  - From BN to distributions, DAGs, Markov(G)
  - Parameterization
- Graphoids axioms for Conditional Independence
- D-separation: Inferring CIs in graphs

Outline

- Bayesian Networks, DAGS, Markov(G)
  - From a distribution to a BN
  - From BN to distributions, DAGs, Markov(G)
  - Parameterization
- Graphoids axioms for Conditional Independence
- D-separation: Inferring CIs in graphs

Bayesian Networks (BNs) in 2 ways:

From a distribution to a BN:

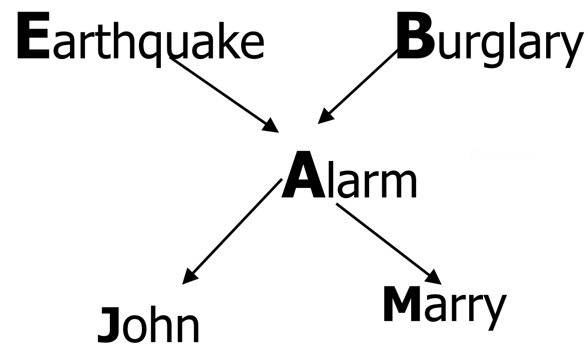
- A Bayesian network is factorize probability distribution along an ordering.
- The DAG emerging is a Bayesian network of the distribution
- The factorization is guided by a set of Markov assumption that transform the chain product formula into a Bayesian network.

From a BN to a distribution:

- Generate a DAG with its Markov assumptions.
- Parameterize the DAG yielding a Bayesian network which corresponds to a single probability distribution obtained by product.

- The BN distribution obeys additional independence assumption read from the DAG and can be proved using the Graphoid axioms.

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: $2^{n-1} + 2^{n-1} + \dots + 1 = 2^n - 1$.

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 continued

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

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

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

pa(B) = {}, pa(E)={B},P(A)= {B,E}, pa(J) = {A}, pa(M) = {A}

Example continued

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

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

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

pa(B) = {}, pa(E)={B},P(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.

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

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

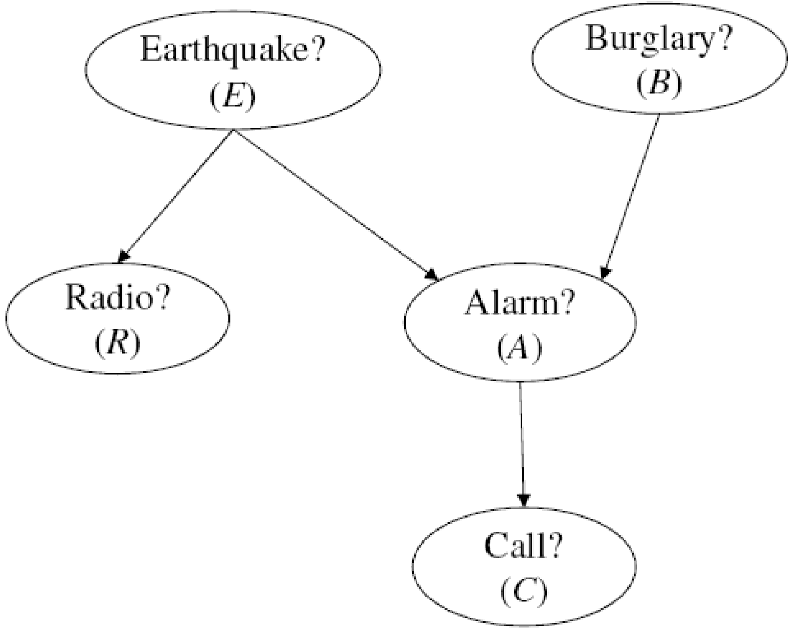
- Inference easier.Will see this later.

Outline

- Bayesian Networks, DAGS, Markov(G)
  - From a distribution to a BN
  - From BN to distributions, DAGs, Markov(G)
  - Parameterization
- Graphoids axioms for Conditional Independence
- D-separation: Inferring CIs in graphs

Capturing Independence Graphically

The causal interpretation



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

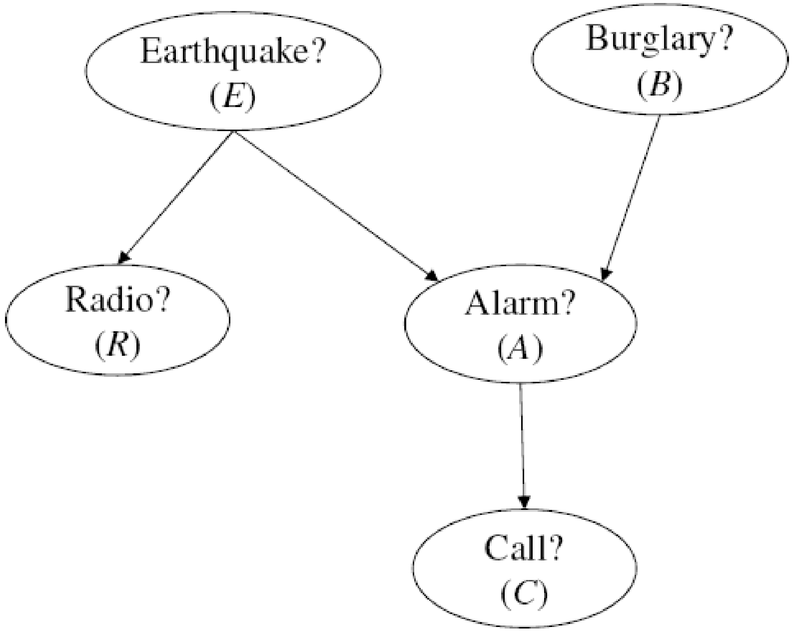
Example

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

Capturing Independence Graphically

But influences can be indirect as well.
For example…

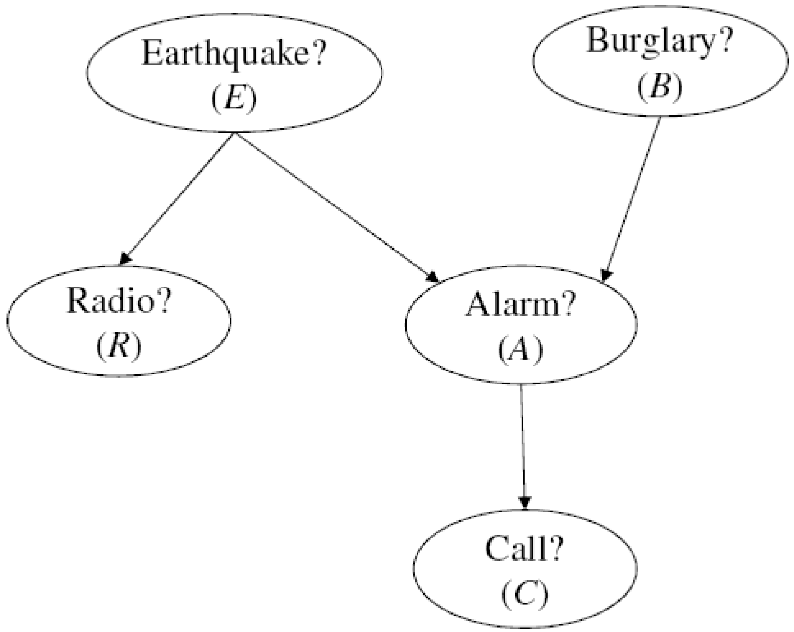
We expect our belief in C to be influenced by evidence on R.



Example

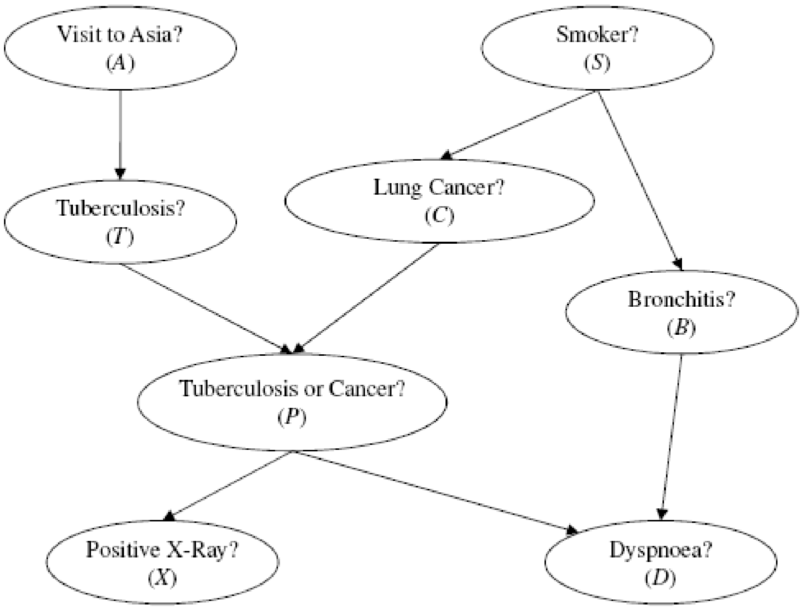
If we get a radio report that an earthquake took place in our neighborhood, our belief in the alarm triggering would probably increase, which would also increase our belief in receiving a call from our neighbor.

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 $\neg T$.

Graphs Convey Independence Statements

- Directed graphs by graph's d-separation
- Undirected graphs by graph separation
- Goal: capture probabilistic conditional independence by graphs.
- We focus on directed graphs.

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:

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

$\mathrm{Descendants}(V)$ 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).

$\mathrm{Non\_Descendants}(V)$ are all variables in DAG G other than V, $\mathrm{Parents}(V)$ and $\mathrm{Descendants}(V)$. 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):

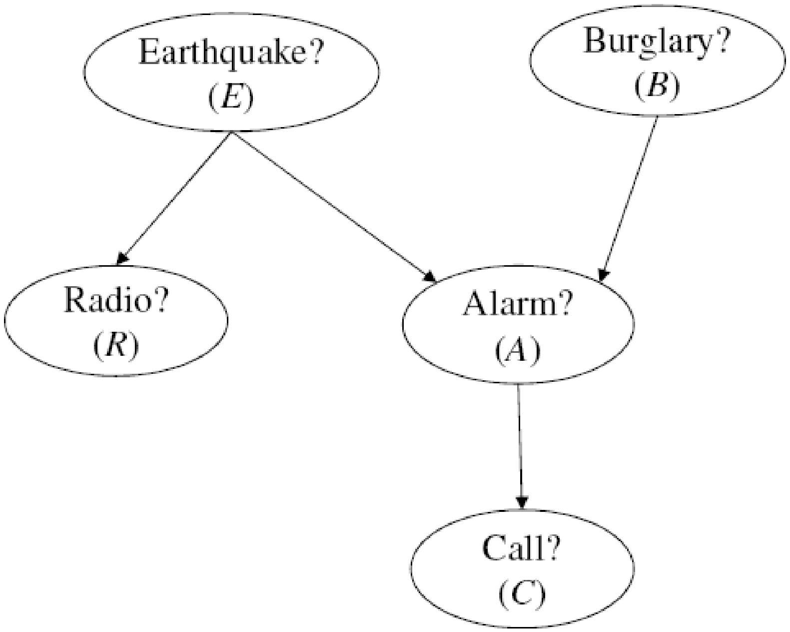
$$I(V, \mathrm{Parents}(V), \mathrm{Non\_Descendants}(V)),$$

for all variables V in DAG G.

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

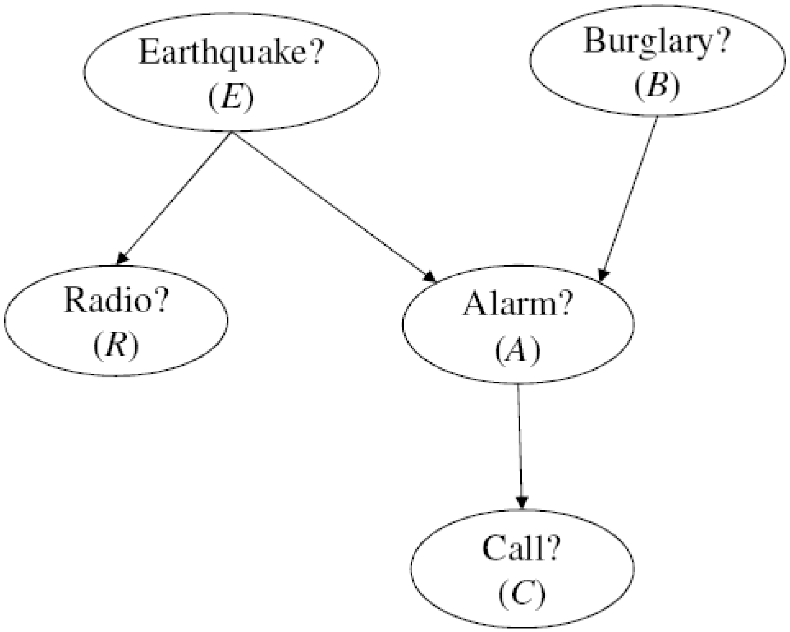
What are the Markov assumptions here?



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

Capturing Independence Graphically

What are the Markov assumptions here?



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

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

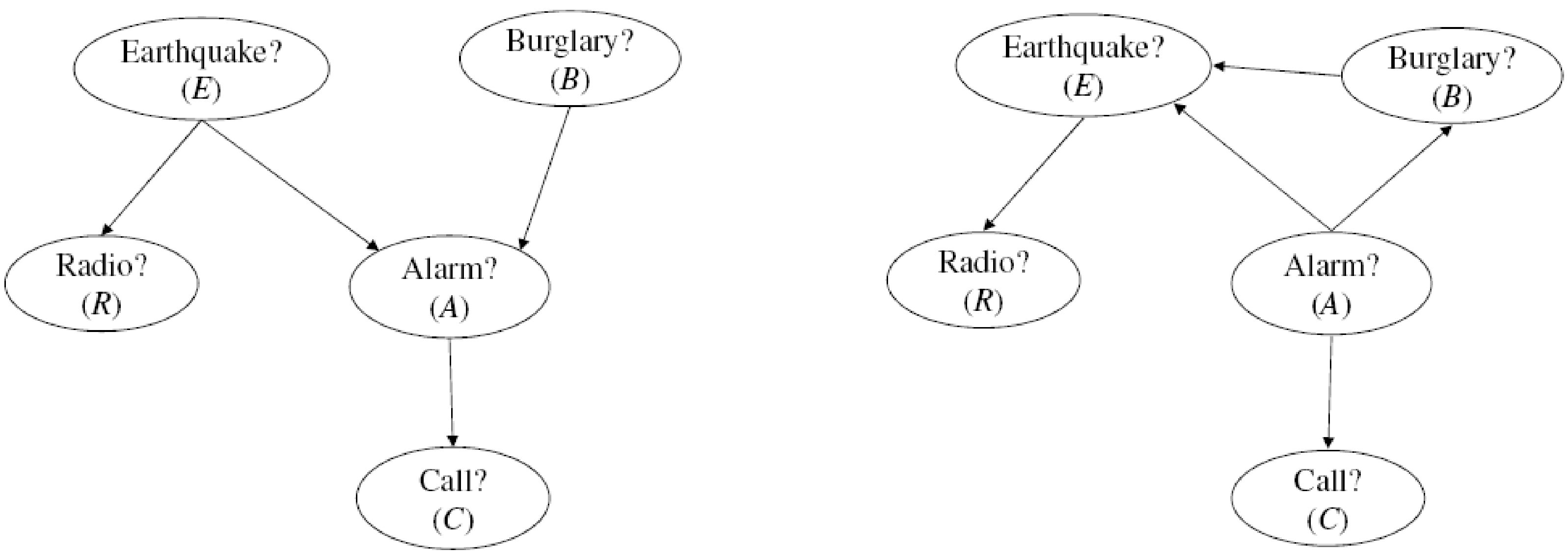
Capturing Independence Graphically

The formal interpretation of a DAG as a set of conditional independence statements makes no reference to the notion of causality, even though we have used causality to motivate this interpretation.

If one constructs the DAG based on causal perceptions, then one would tend to agree with the independencies declared by the DAG.

It is perfectly possible to have a DAG that does not match our causal perceptions, yet we agree with the independencies declared by the DAG.

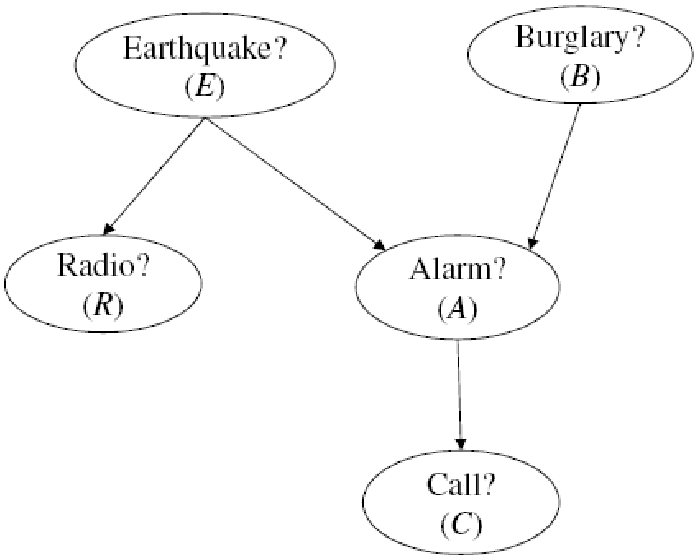
Capturing Independence Graphically



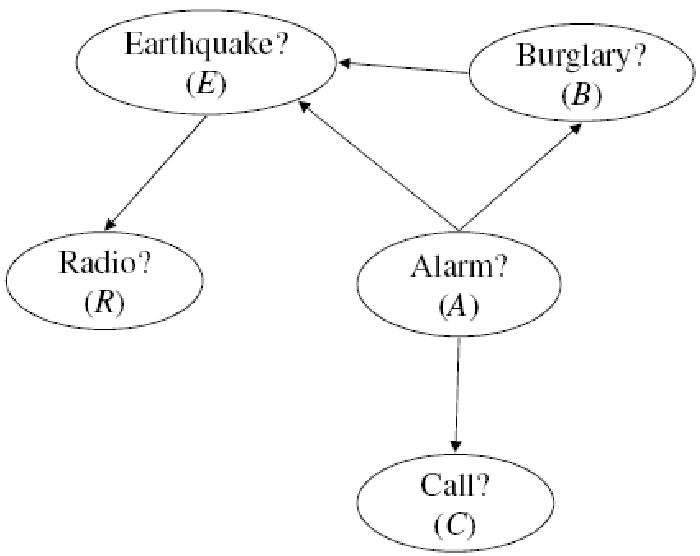
Every independence which is declared (or implied) by the second DAG is also declared (or implied) by the first one. Hence, if we accept the first DAG, then we must also accept the second.

Capturing Independence Graphically

A **Causal** Bayesian Network



A **non-causal** Bayesian Network

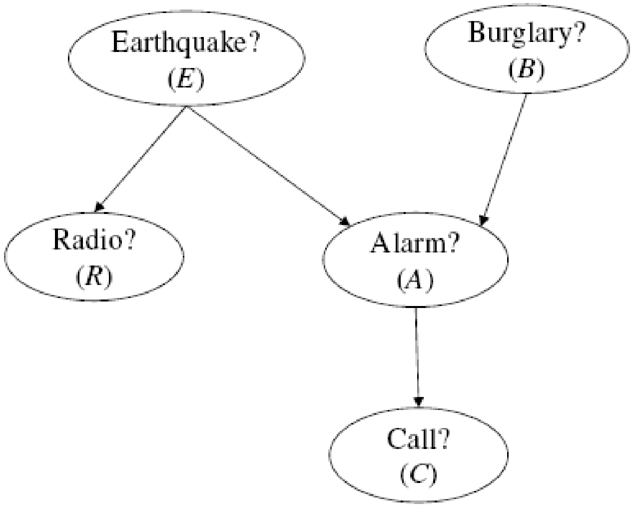


Every independence which is declared (or implied) by the second DAG is also declared (or implied) by the first one. Hence, if we accept the first DAG, then we must also accept the second.

Outline

- Bayesian Networks, DAGS, Markov(G)
  - From a distribution to a BN
  - From BN to distributions, DAGs, Markov(G)
  - Parameterization
- Graphoids axioms for Conditional Independence
- D-separation: Inferring CIs in graphs

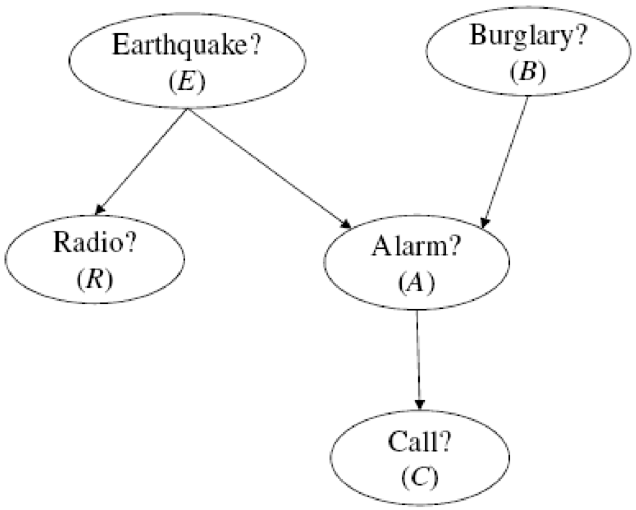
Parameterizing the Independence Structure



- 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 $\mathbf{U}$.

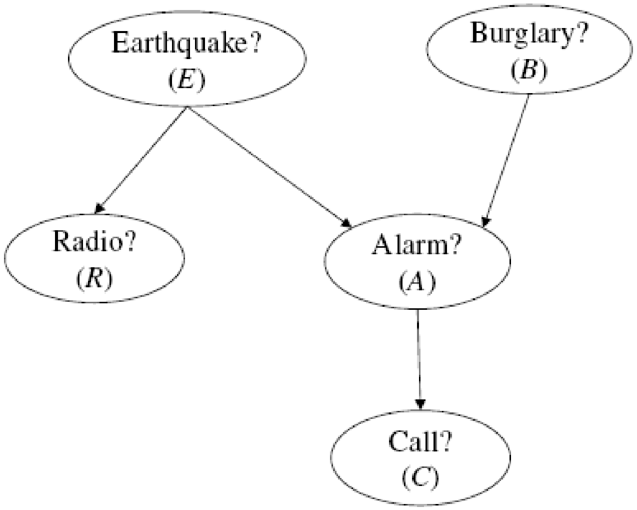
Example

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

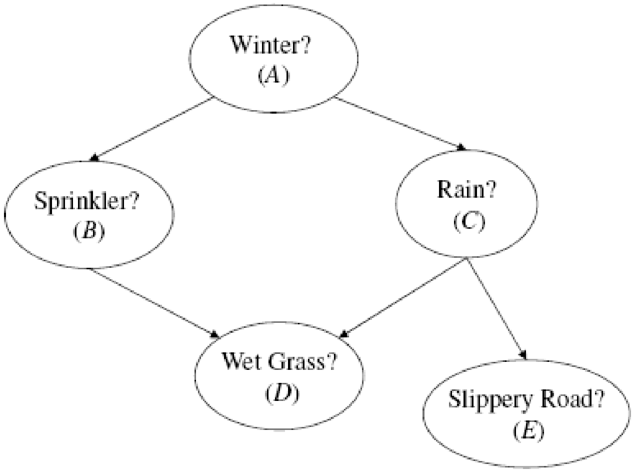
A Bayesian network for variables $\mathbf{Z}$ is a pair (G, Θ), where

- 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

Use GeNie/Smile
To create this network

Parameterizing the Independence Structure

Chain rule for Bayesian networks

A Bayesian network is an implicit representation of a unique probability distribution $\Pr$ given by

$$\Pr(\mathbf{z}) \stackrel{def}{=} \prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}}.$$

The probability assigned to a network instantiation $\mathbf{z}$ is simply the product of all network parameters that are compatible with $\mathbf{z}$.

Parameterizing the Independence Structure

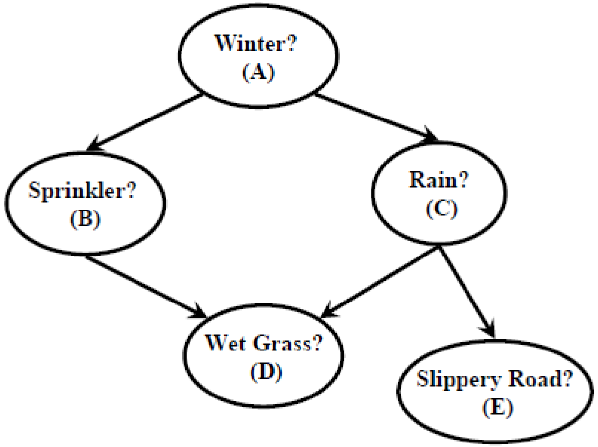
Chain rule for Bayesian networks

A Bayesian network is an implicit representation of a unique probability distribution $\Pr$ given by

$$\Pr(\mathbf{z}) \stackrel{def}{=} \prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}}.$$

The probability assigned to a network instantiation $\mathbf{z}$ is simply the product of all network parameters that are compatible with $\mathbf{z}$.

Parameterizing the Independence Structure



Example

$$
\begin{aligned}
&\Pr(a, b, \bar{c}, d, \bar{e}) \\
&= \theta_{a}\ \theta_{b|a}\ \theta_{\bar{c}|a}\ \theta_{d|b,\bar{c}}\ \theta_{\bar{e}|\bar{c}} \\
&= (.6)(.2)(.2)(.9)(1) \\
&= .0216
\end{aligned}
$$

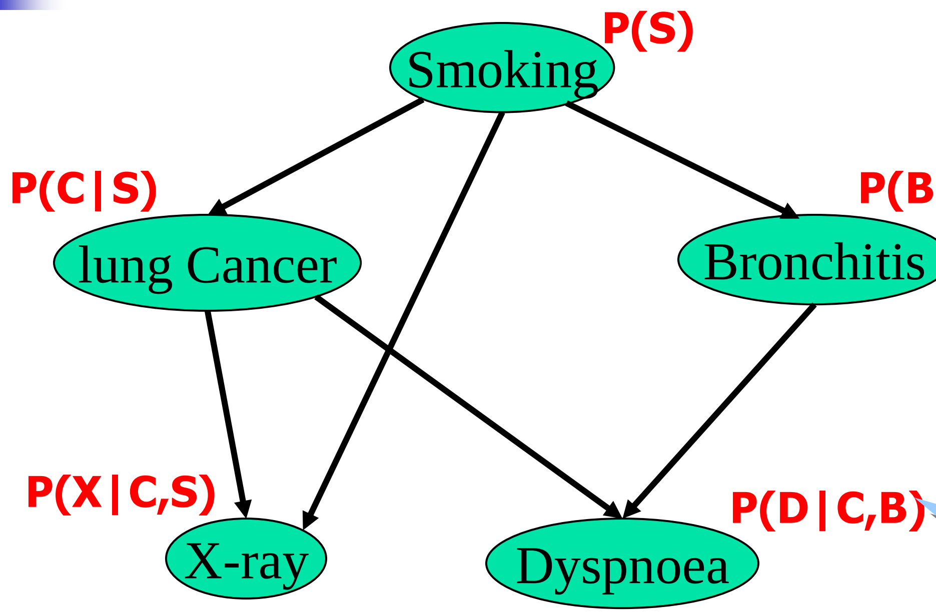
Example

$$
\begin{aligned}
&\Pr(\bar{a}, \bar{b}, \bar{c}, \bar{d}, \bar{e}) \\
&= \theta_{\bar{a}}\ \theta_{\bar{b}|\bar{a}}\ \theta_{\bar{c}|\bar{a}}\ \theta_{\bar{d}|\bar{b},\bar{c}}\ \theta_{\bar{e}|\bar{c}} \\
&= (.4)(.25)(.9)(1)(1) \\
&= .09
\end{aligned}
$$

Parameterizing the Independence Structure

- The CPT $\Theta_{X|\mathbf{U}}$ is exponential in the number of parents $\mathbf{U}$.
- If every variable can take up to d values, and has at most k parents, the size of any CPT is bounded by $O(d^{k+1})$.
- If we have n network variables, the total number of Bayesian network parameters is bounded by $O(n \cdot d^{k+1})$.
- This number is quite reasonable as long as the number of parents per variable is relatively small.

Bayesian Networks: Representation



$\mathbf{BN} = (\mathbf{G}, \mathbf{\Theta})$

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

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

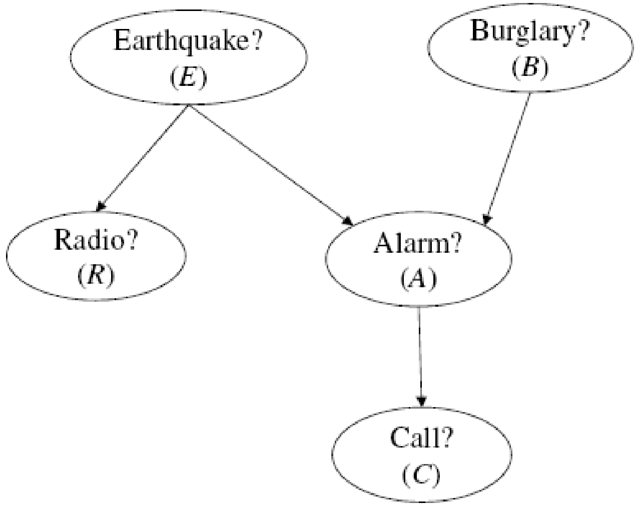
Conditional Independencies ➡ Efficient Representation

Outline

- Basic of Probability Theory
- Bayesian Networks, DAGS, Markov(G)
- Graphoids axioms for Conditional Independence
- 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 $\mathrm{Markov}(G)$.

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

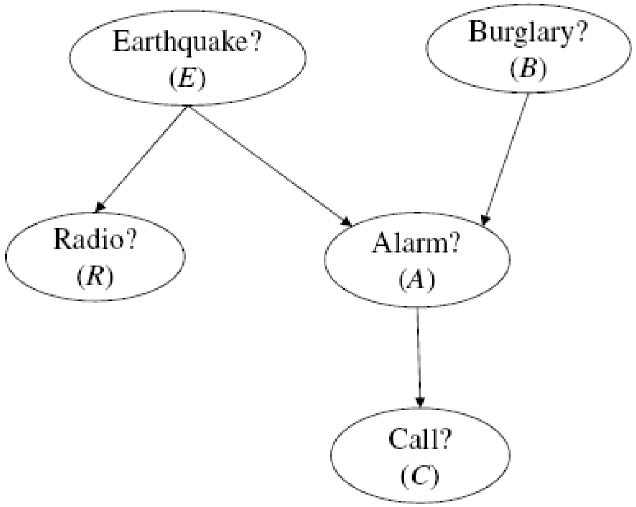
R and C are independent given A

Properties of Probabilistic Independence
(Pearl ch 3)

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.

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

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

Decomposition

The opposite of Decomposition, called Composition:

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

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 (C1,C2,) and a bell (B)

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.

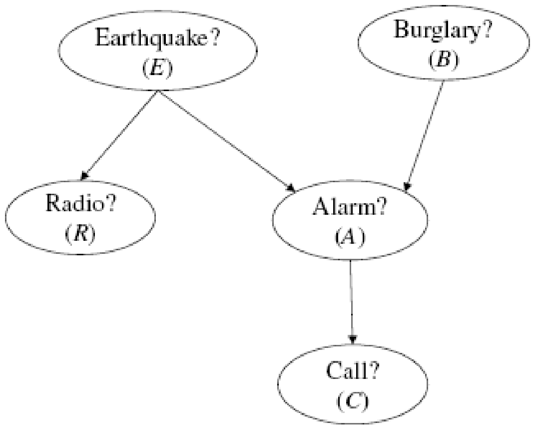
This is a strengthening of the independence statements declared by $\mathrm{Markov}(G)$, which is a special case when $\mathbf{W}$ contains all non-descendants of X.

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:

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

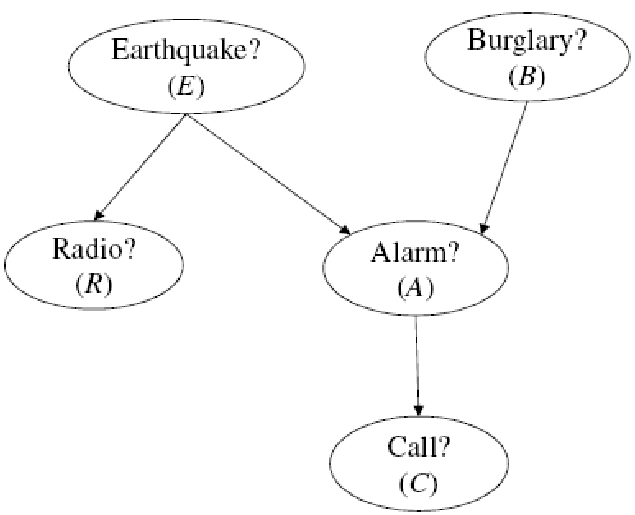
This leads to the chain rule of Bayesian networks:

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

Weak Union

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

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

Contraction

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

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

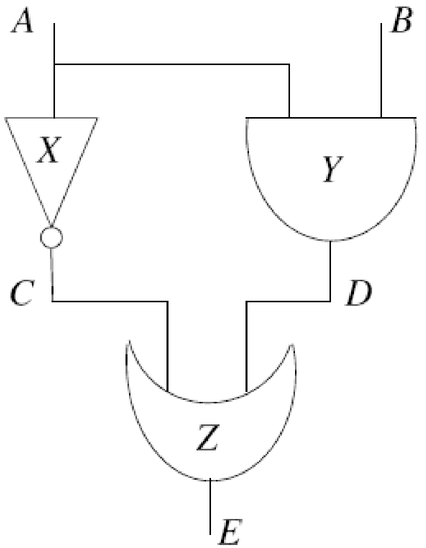
One can view Contraction as a weaker version of Composition. Recall that Composition does not hold for probability distributions.

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.

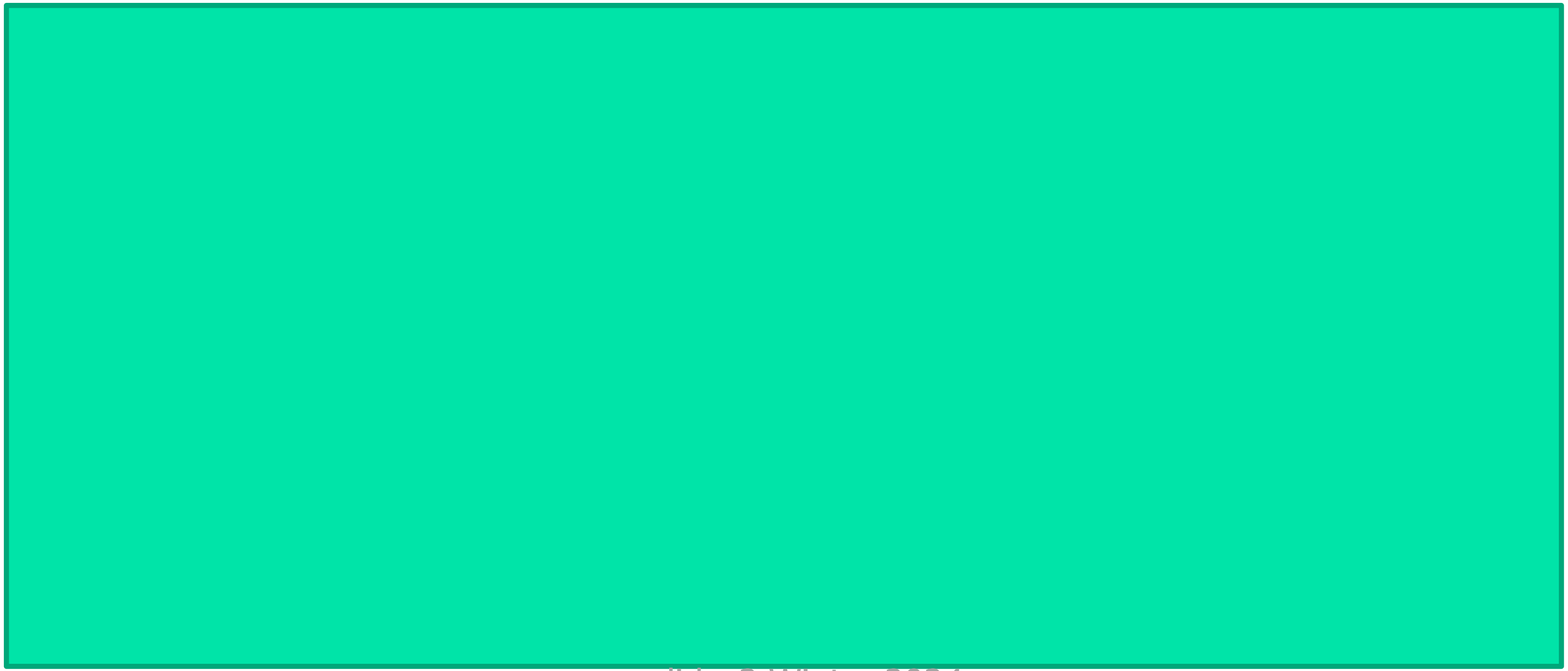
A strictly positive distribution cannot capture logical constraints.

Intersection

Holds only for strictly positive distributions

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

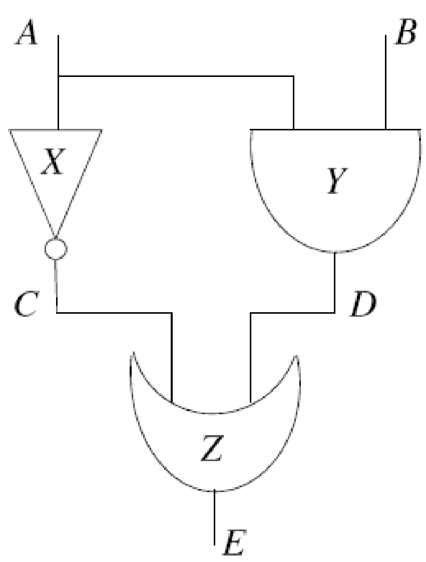


Intersection

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.



- 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

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

- 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

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

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

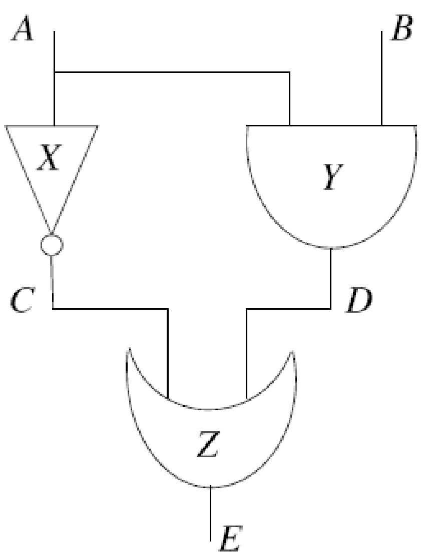
If information **w** is irrelevant given **y**, and **y** is irrelevant given **w**, then combined information **yw** is irrelevant to start with.

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

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