VU VRIJE UNIVERSITEIT AMSTERDAM

Bayesian Networks

Knowledge Representation

- Andreas Sauter
- Dec. 2023
- (Content adapted from Erman Acar)

Overview

1. Foundations
   Degrees of Belief, Belief Dynamics, Independence, Bayes Theorem, Marginalization

2. Bayesian Networks
   Graphs and their Independencies, Bayesian Networks, d-Separation

3. Tools for Inference
   Factors, Variable Elimination, Elimination Order, Interaction Graphs, Graph pruning

4. Exact Inference in Bayesian Networks
   Posterior Marginal, Maximum – A-posteriori, Most Probable Explanation

Lecture 2: Bayesian Networks

Lecture Overview

Directed Acyclic Graphs

Nodes, Edges, Ancestry, Special Paths

Bayesian Networks

Motivation, Formal Definition

Independence through DAGs

Markov Property, Symmetry, Decomposition, Weak Union, Contraction

d-Separation

d-Blocked Paths, d-Separation, Pruning for d-Separation

Directed Acyclic Graphs

Motivation

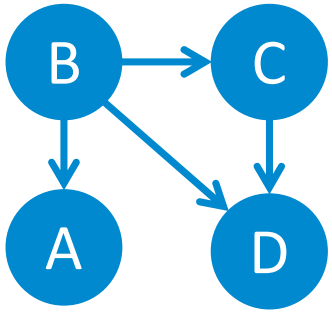
As we will see later, graphs are an important component of Bayesian networks.

This section serves as a refresher of basic and relevant concepts of graphs.

We will consider only directed acyclic graphs (DAGs) in the context of BNs.

These are graphs that have:

- Nodes
- Directed edges
- No cycles



SA(0

Basic Definitions

The nodes V represent propositional variables.

E.g: Alarm, Burglary, Earthquake

The directed edges E represent direct influences.

We define a DAG $G = (V, E)$ as a set of directed edges E over nodes V that do not induce a cycle.

A path π from $A \in V$ to $B \in V$ is a node-edge sequence in which B can be reached from A.

A directed path is a path in which all edges are directed towards the end node.

Slide 7

SA(0 Add definition of leaf node

Sauter, A. (Andreas), 2022-12-07T12:32:02.564

Ancestry

The parents $Pa(A)$ of a node A is the set of variables which have a directed edge to A.

The descendants $Desc(A)$ of a node A is the set of variables which can be reached by a directed path from A.

The non-descendants $non\_Desc(A)$ of a node A is the set of all variables which are neither descendants nor parents of A.

A leaf node is a node without descendants.

Special Paths

Three special types of paths play an important role in BNs.

For $A, B, W \in V$ we call a path π a:

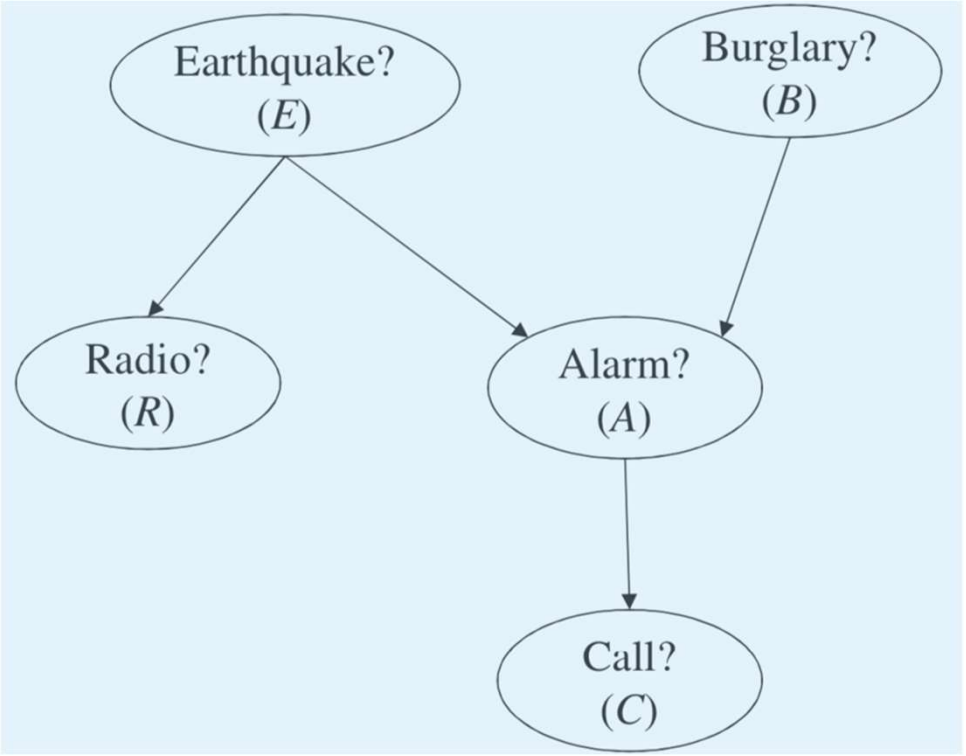
- Sequence if $\pi = A \rightarrow W \rightarrow B$
- Fork if $\pi = A \leftarrow W \rightarrow B$
- Collider if $\pi = A \rightarrow W \leftarrow B$

Examples

$Pa(A) = \{E, B\}$
$Desc(B) = \{A, C\}$
$Non\_Desc(B) = \{E, R\}$

$R \leftarrow E \rightarrow A \rightarrow C$ is a path
$B \rightarrow A \rightarrow C$ is a directed path

$R \leftarrow E \rightarrow A$ is a fork
$E \rightarrow A \leftarrow B$ is a collider
$E \rightarrow A \rightarrow C$ is a sequence



Bayesian Networks

Motivation

We know that joint probability tables are useful for reasoning about beliefs.

Unfortunately, representing this table needs 2^N rows even in the simplest case.

Bayesian networks address this issue by factorizing the joint probability distribution by means of the independence structure of the variables.

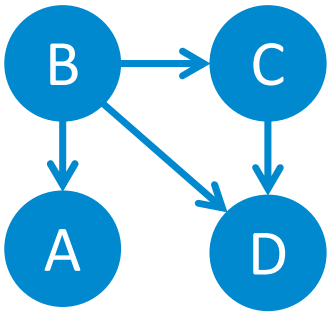
BNs acknowledge the fact that independence forms a significant aspect of beliefs and that it can be elicited relatively easily using the language of graphs.

Furthermore, BNs enable more efficient inference algorithms on probabilistic knowledge.

Example

A	B	C	D	P(A,B,C,D)
True	True	True	True	0.020
True	True	True	False	0.046
True	True	False	True	0.191
True	True	False	False	0.002
True	False	True	True	0.022
True	False	True	False	0.009
True	False	False	True	0.004
True	False	False	False	0.031
False	True	True	True	0.016
False	True	True	False	0.043
False	True	False	True	0.156
False	True	False	False	0.002
False	False	True	True	0.083
False	False	True	False	0.083
False	False	False	True	0.034
False	False	False	False	0.276

B	P(B)
True	0.47
False	0.53



C	B	P(C|B)
True	True	0.255
True	False	0.35
False	True	0.745
False	False	0.65

A	B	P(A|B)
True	True	0.55
True	False	0.1
False	True	0.45
False	False	0.9

D	B	C	P(D|B,C)
True	True	True	0.3
True	True	False	0.99
True	False	True	0.5
True	False	False	0.11
False	True	True	0.7
False	True	False	0.01
False	False	True	0.5
False	False	False	0.89

Definition

Intuitively, a Bayesian network consists of a structure and a parametrization where the overall joint distribution can be determined with the chain rule.

Formally: Given a DAG $G = (V, E)$ and a set of conditional probability tables $P = \{P_{X_i}(X_i|Pa_G(X_i)) : \forall X_i \in V\}$, a tuple $N = (G, P)$ is a Bayesian network if

$$\mathrm{P(V)} = \Pi_{X_i \in V} P_{X_i}(X_i|Pa_G(X_i))$$

Where $P(V)$ is the joint distribution over all variables and $Pa_G(X_i)$ denotes the parents of X_i w.r.t. the graph G.

Notations

For a less cluttered formulation, we will use the following notations.

For a variable X, we will write x if $X = true$ and $\neg x$ if $X = false$.

We will denote the conditional probability table $P_{X_i}(X_i|Pa_G(X_i))$ as $\Theta_{X_i|U_i}$ where $U_i = Pa_G(X_i)$

For an assignment $X_i = x_i \; and \; Pa_G(X_i) = u_i$ we will denote $\Theta_{x_i|u_i}$

Instantiations

An assignment of all network variables will be called a network instantiation.

A conditional probability table is compatible with an assignment z, denoted $\Theta_{x|u} \sim z$ iff the instantiations x, u and z agree on their assignments.

e.g. $\Theta_{a|\neg b} \sim a, \neg b, c$ and $\Theta_c \sim a, \neg b, c$

For an instantiation z we can re-write the chain rule to compute its probability as

$$\Pr(z) = \Pi_{\Theta_{x|u} \sim z} \Theta_{x|u}$$

E.g. Assume a sequence $A \rightarrow B \rightarrow C$ then $Pr(a, b, c) = \Theta_{c|b}\Theta_{b|a}\Theta_a$

$$= \Pr(c|b)\Pr(b|a)\Pr(a)$$

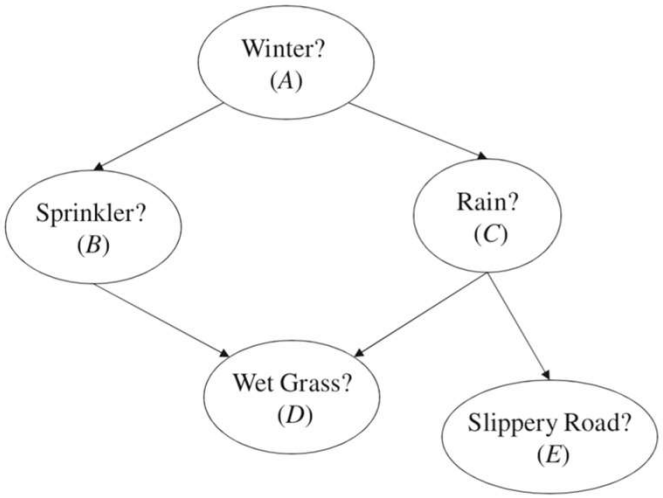
VU

Instantiations – Examples

$$\Pr(a, b, \neg c, d, \neg e)$$
$$= \Theta_a \Theta_{b|a} \Theta_{\neg c|a} \Theta_{d|b,\neg c} \Theta_{\neg e|\neg c}$$
$$= 0.6 * 0.2 * 0.2 * 0.9 * 1$$
$$= 0.0216$$

$$\Pr(\neg a, \neg b, \neg c, \neg d, \neg e)$$
$$= \Theta_{\neg a} \Theta_{\neg b|\neg a} \Theta_{\neg c|\neg a} \Theta_{\neg d|\neg b,\neg c} \Theta_{\neg e|\neg c}$$
$$= 0.4 * 0.25 * 0.9 * 1 * 1$$
$$= 0.09$$

From this we can recover $\Pr(A, B, C, D, E)$



A	Θ_A
true	.6
false	.4

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

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

Independence through DAGs

Introduction

If defined accordingly, DAGs can be a great tool to determine independence.

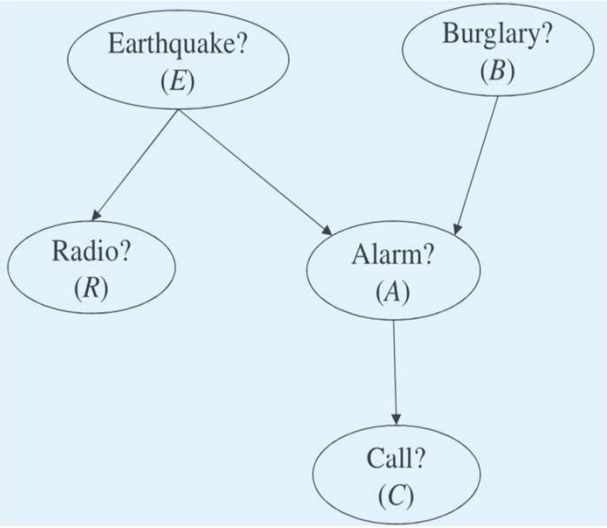
We don’t even need to know what the actual distributions are!

Some intuition:

An alarm directly triggers a call from the neighbour.

If we get a radio report that an earthquake took place, then the belief in A changes, which in turn changes our belief in C.

Yet, if we know already that no alarm was triggered, then the belief in the call stays unchanged. Hence, $C \perp\!\!\!\perp R \mid A$



Markov Property

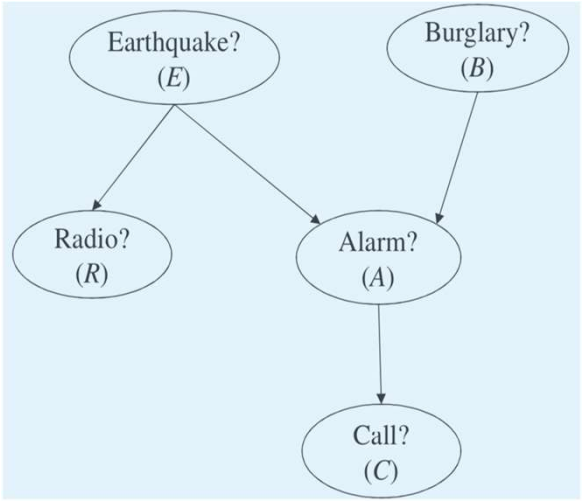
One information a DAG can give us about independence is the Markov Property.

The Markov property tells us that every variable is conditionally independent of its non-descendants given its parents, or formally for $\mathrm{N} \in \mathrm{V}$:

$$\{\mathrm{N}\} \perp\!\!\!\perp non\_Desc(N) \mid Pa(N)$$

Example:

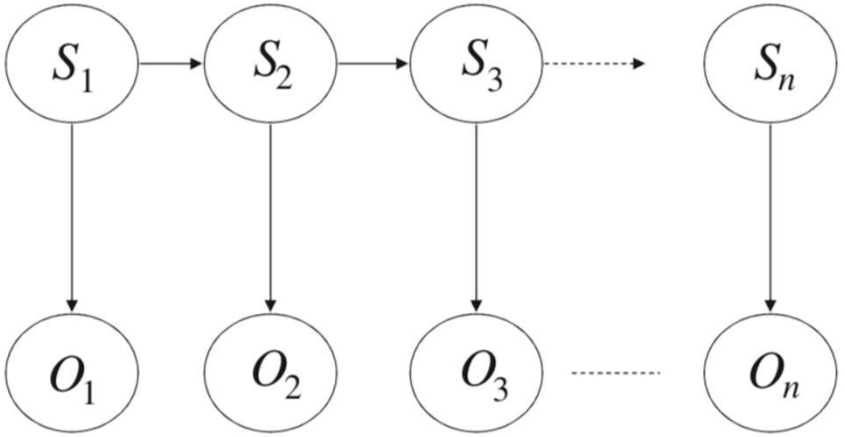
$$\{C\} \perp\!\!\!\perp \{B, E, R\} \mid \{A\}$$
$$\{R\} \perp\!\!\!\perp \{A, B, C\} \mid \{E\}$$
$$\{A\} \perp\!\!\!\perp \{R\} \mid \{B, E\}$$
$$\{B\} \perp\!\!\!\perp \{E, R\} \mid \emptyset$$
$$\{E\} \perp\!\!\!\perp \{B\} \mid \emptyset$$



Markov Property – Further Example

This property is also often used in the Hidden Markov Model (HMM) which has many applications in reinforcement learning, NLP, ect.

It consists of states at discrete time steps $S_1, S_2, \dots, S_T$ and observations resulting from the state $O_1, O_2, \dots, O_T$ (e.g. measurements), with the following structure:



The Markov property tells us, that
$\forall S_i: \{S_i\} \perp\!\!\!\perp \{S_1, \dots, S_{i-2}, O_1, \dots O_{i-2}\} \mid \{S_{i-1}\}$

Informally: “The current state can be fully determined by the previous state”

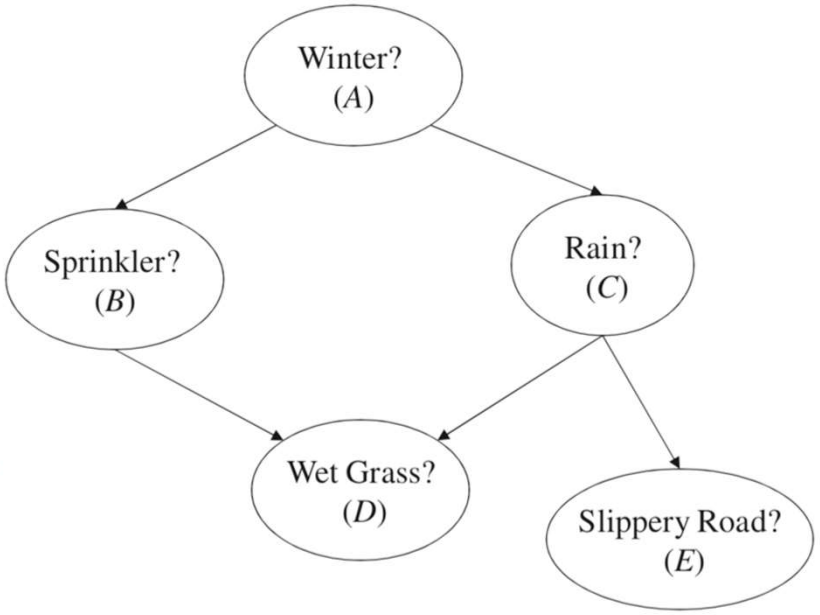
VU

Properties of Independence

Although all independencies of the Markov property are encoded by the DAG, the DAG implies even more independencies.

E.g. the graph on the right also implies $\{D\} \perp\!\!\!\perp \{E\} \mid \{A, C\}$, which does not follow from the Markov property.

These additional independencies can be derived by a set of independence properties.



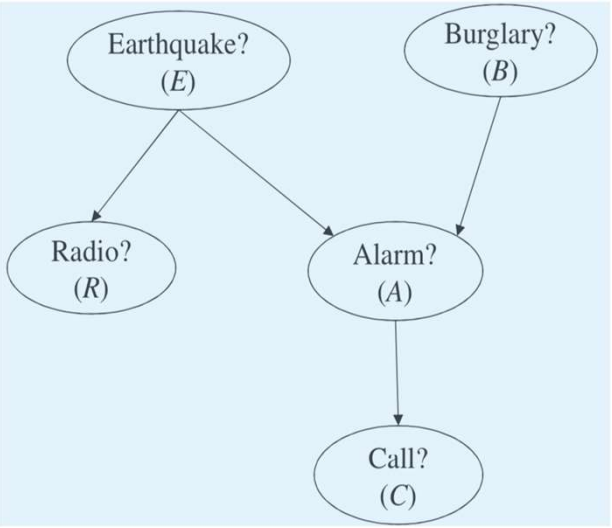
Symmetry (recap)

Intuitively, if observing Y does not influence our belief in X, then learning X does not influence our belief in Y either.

Formally, symmetry is expressed as the following:

$$\{X\} \perp\!\!\!\perp \{Y\} | \{Z\} \Leftrightarrow \{Y\} \perp\!\!\!\perp \{X\} | \{Z\}$$

Example: If the graph encodes $\{A\} \perp\!\!\!\perp \{R\} \mid \{B, E\}$ because of the Markov property, then it also encodes $\{R\} \perp\!\!\!\perp \{A\} \mid \{B, E\}$ because of symmetry.



Decomposition

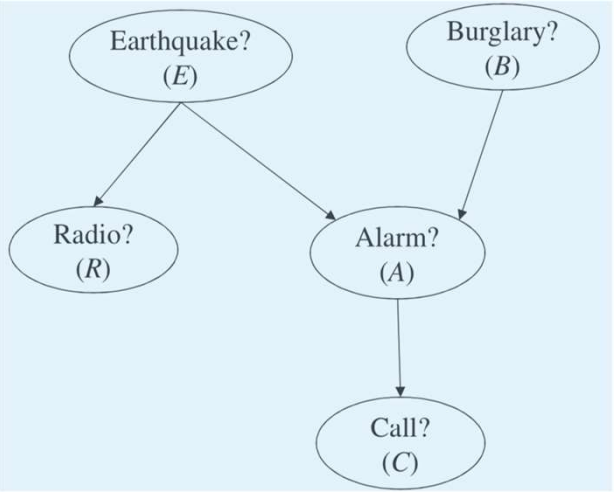
Intuitively, if observing $\{Y, W\}$ does not influence our belief in X, then learning Y alone or W alone, will not influence our belief in X.

Formally, decomposition can be expressed as the following:

$$\{X\} \perp\!\!\!\perp \{Y, W\} \mid \{Z\} \Rightarrow \{X\} \perp\!\!\!\perp \{W\} \mid \{Z\} \wedge \{X\} \perp\!\!\!\perp \{Y\} \mid \{Z\}$$

Note: The opposite direction does not hold in general.

Example: If we know $\{R\} \perp\!\!\!\perp \{A, C\} \mid \{E\}$, then also $\{R\} \perp\!\!\!\perp \{A\} \mid \{E\}$ and $\{R\} \perp\!\!\!\perp \{C\} \mid \{E\}$.



Decomposition - Application

More generally, with the help of decomposition we can state that

$$\forall W \subseteq non\_Desc(X): X \perp\!\!\!\perp W \mid Pa(X)$$

This is especially useful when calculating a joint probability distribution with the chain rule.

Example: $\Pr(R, C, A, E, B) = \Pr(R|\, C, A, E, B)\Pr(C|A, E, B)\, \Pr(A|E, B)\, \Pr(E|B)\, \Pr(B)$

Can be simplified to: $\Pr(R, C, A, E, B) = \Pr(R|E)\, \Pr(C|A)\, \Pr(A|E, B)\, \Pr(E)\, \Pr(B)$

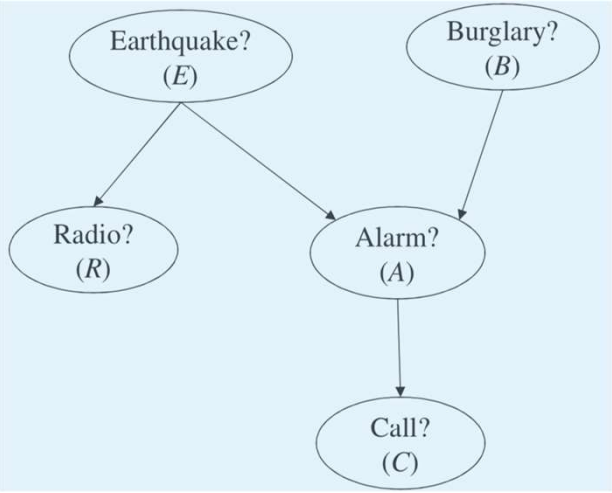
Weak Union

Intuitively, this expresses that, if the information $\{Y, W\}$ is not relevant to our belief in X, then the partial information Y will not make W relevant.

Formally, weak union can be expressed as the following:

$$\{X\} \perp\!\!\!\perp \{Y, W\} \mid \{Z\} \Rightarrow \{X\} \perp\!\!\!\perp \{W\} \mid \{Z, Y\}$$

Example: If we know $\{C\} \perp\!\!\!\perp \{B, R\} \mid \{A\}$, then we can conclude by weak union also $\{C\} \perp\!\!\!\perp \{R\} \mid \{A, B\}$.



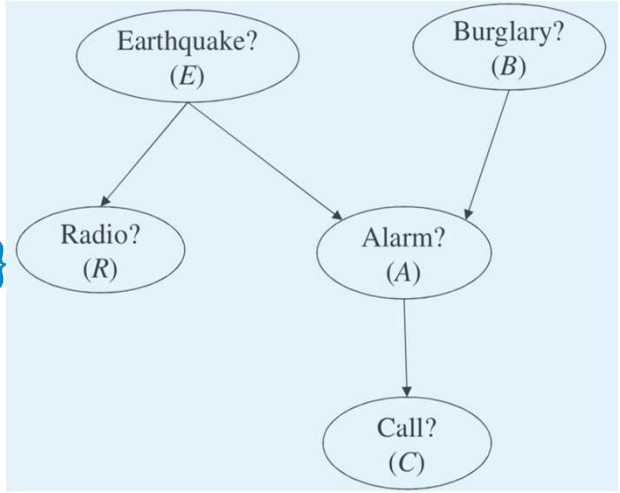
Contraction

Intuitively, given Z, if observing an irrelevant information Y makes W irrelevant, then W, Y must have been irrelevant from the start.

Formally, contraction can be expressed as the following:

$$\{X\} \perp\!\!\!\perp \{Y\} \mid \{Z\} \wedge \{X\} \perp\!\!\!\perp \{W\} \mid \{Y, Z\} \Rightarrow \{X\} \perp\!\!\!\perp \{Y, W\} \mid \{Z\}$$

Example: If we know $\{C\} \perp\!\!\!\perp \{B\} \mid \{A\} \wedge \{C\} \perp\!\!\!\perp \{E\} \mid \{A, B\}$ then $\{C\} \perp\!\!\!\perp \{E, B\} \mid \{A\}$



d-Separation

Introduction

We have seen that deriving new independencies from the Markov property can be cumbersome.

Luckily, there is a graphical test called d-separation which captures the same independencies as the rules described before.

If X is d-separated from Y by Z, we denote this as $X \perp^d Y \mid Z$.

Each d-separation implies an independence in a Bayesian network:

$$\forall X, Y, Z \subseteq V\colon X \perp^d Y \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$$

Important: Not the other way round!

d-Blocked Paths

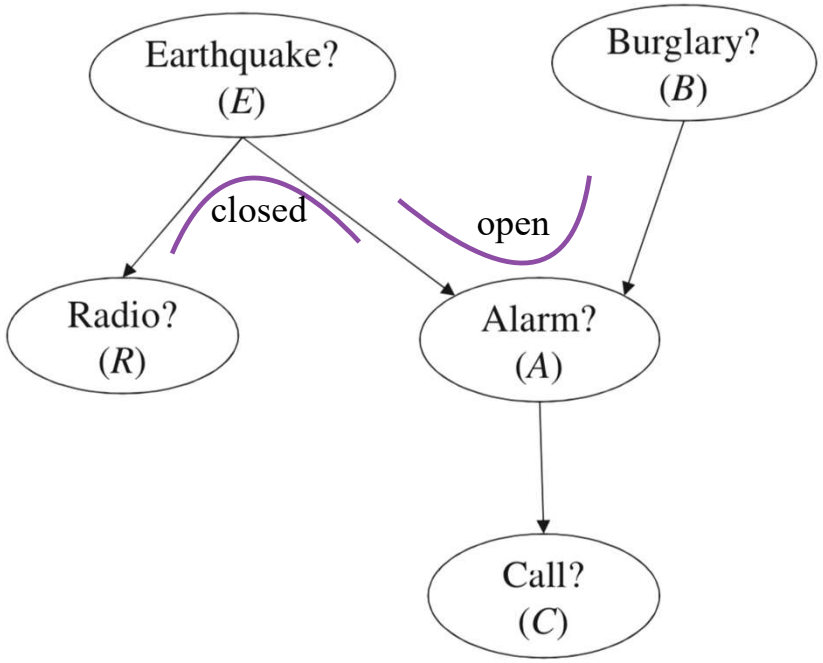
Recall the three special paths described in before: sequence, fork, collider

These can be, what we call, d-blocked in the following cases:

- A sequence $A \rightarrow W \rightarrow B$ is d-blocked by Z, iff W ∈ Z.
- A fork $A \leftarrow W \rightarrow B$ is d-blocked by Z, iff W ∈ Z.
- A collider $A \rightarrow W \leftarrow B$ is d-blocked by Z, iff **neither** W **nor** any desc(W) ∈ Z.

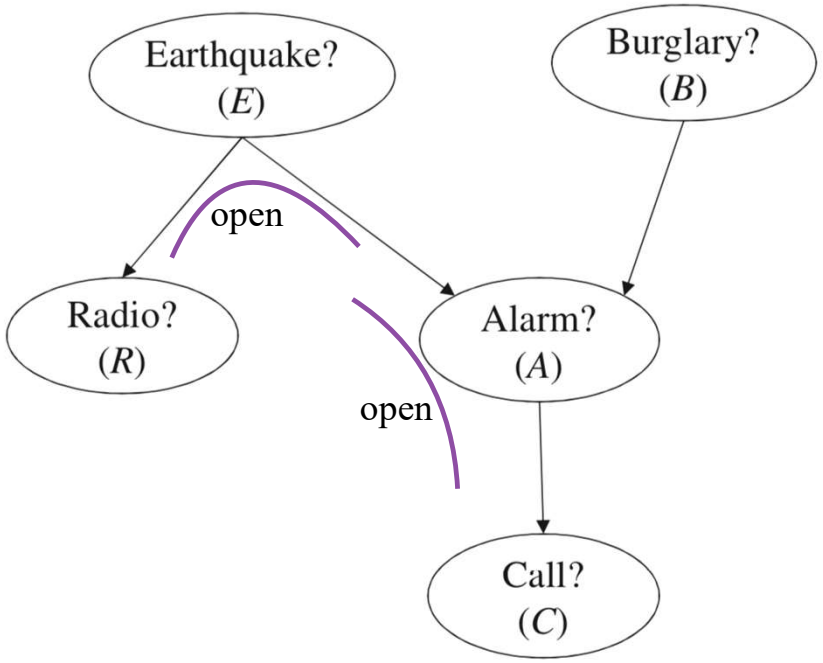
In general, a path is d-blocked iff it contains at least one d-blocked sequence, fork or collider.

Examples



Is $R \leftarrow E \rightarrow A \leftarrow B$ d-blocked by $\{E, C\}$?

Yes!



Is $R \leftarrow E \rightarrow A \rightarrow C$ d-blocked by $\emptyset$?

No!

d-Separation

We now have all the tools to properly define d-separation.

Given disjoint sets $X, Y, Z \subseteq V$, we say X and Y are d-separated by Z, iff every path between a node in X to a node in Y is d-blocked by Z.

Intuitively this means that there is no way information can flow between X and Y when we condition on Z.

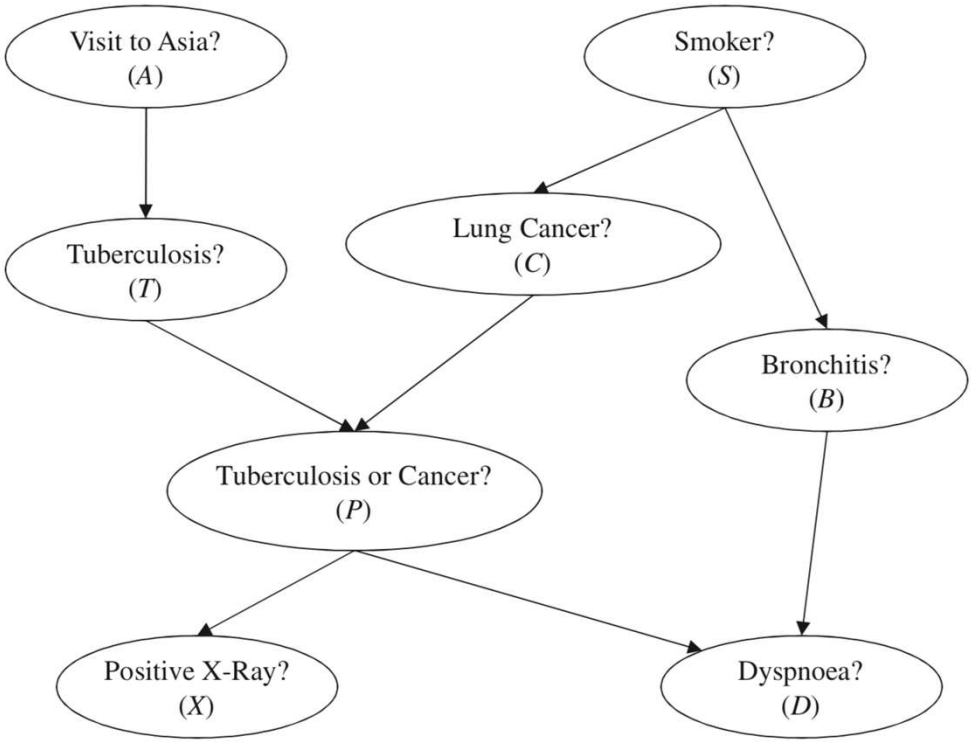
Examples

Does $\{B\} \perp^d \{C\} \mid \{S\}$ hold?
Yes!

Does $\{X\} \perp^d \{S\} \mid \{C, D\}$ hold?
No!

Does $\{X\} \perp^d \{S\} \mid \{C\}$ hold?
Yes!

Does $\{X, S\} \perp^d \{D\} \mid \{B, P\}$ hold?
Yes!



D-Separation through Pruning

Paths between sets of nodes can be exponentially many. The following method guarantees that d-separation can be decided in linear time in the graph size.

Given a DAG G and disjoint sets of nodes $X, Y, Z \subseteq V$ a pruned DAG G' w.r.t. Z is determined by:

- Deleting every leaf node $W \notin X \cup Y \cup Z$,
- Deleting all edges outgoing from nodes in Z,
- Performing both rules iteratively until they can't be applied anymore.

Then, $X \perp^d Y \mid Z$ in $G \Leftrightarrow X$ and Y are disconnected in G' w.r.t. Z.

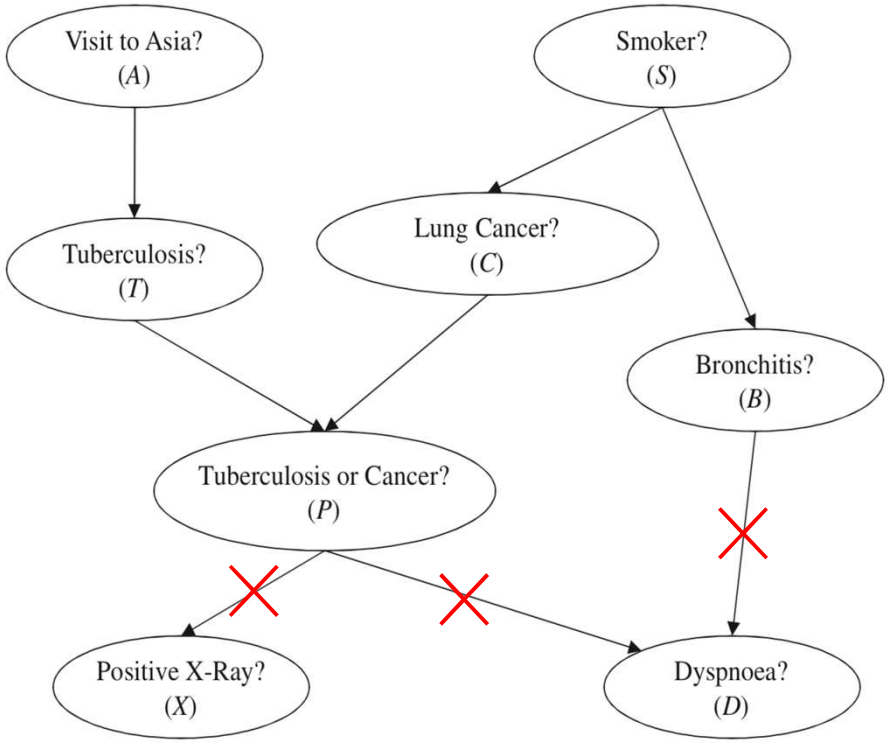
Examples

Recall rules:

- Deleting every leaf node $W \notin X \cup Y \cup Z$,
- Deleting all edges outgoing from nodes in Z,

Is $\{A, S\} \perp^d \{D, X\} \mid \{B, P\}$?

Yes!



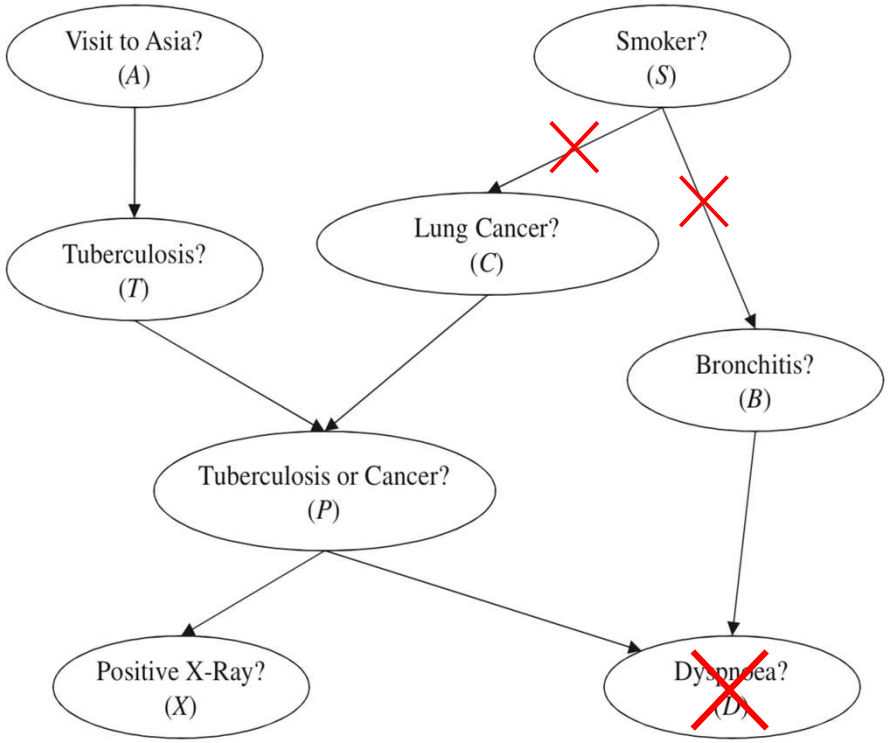
Examples

Recall rules:

- Deleting every leaf node $W \notin X \cup Y \cup Z$,
- Deleting all edges outgoing from nodes in Z,

Is $\{T, C\} \perp^d \{B\} \mid \{S, X\}$?

Yes!



Lecture 2: Summary

- We introduced DAGs and defined important graph-theoretic concepts.
- We investigated which independencies a DAG can encode.
- We introduced and formally defined Bayesian networks.
- We saw an easy way to derive the independencies implied by the structure of a Bayesian Network.