Bayesian networks – Representation

Pekka Parviainen

University of Helsinki

Prob. Models, Spring 2012

A two-variable case

- Assume two binary (Bernoulli distributed) variables A and B.
- Two examples of the joint distribution $P(A, B)$:

	$B = 0$	$B = 1$	$P(A)$
$A = 0$	0.08	0.02	0.10
$A = 1$	0.72	0.18	0.90
$P(B)$	0.80	0.20	

$P(A, B) = P(A)P(B)$

We only need the marginals $P(A)$ and $P(B)$

	$B = 0$	$B = 1$	$P(A)$
$A = 0$	0.08	0.02	0.10
$A = 1$	0.18	0.72	0.90
$P(B)$	0.26	0.74	

$P(A, B) \neq P(A)P(B)$

We need the full table (or $P(A, B) = P(A)P(B \mid A)$).

Independence

- Recall that if $P(A, B) = P(A)P(B)$, then A and B are independent (and thus $P(B \mid A) = P(B)$).
- Independence can be used to separate from all joint distributions $P(A, B)$ the subset where the independence holds.
- Independence simplifies (constrains) things:
  - Model family $'A \perp B'$ is a subset of distributions
  - Model family $'\text{not } A \perp B'$ is the set of all distributions.

Two model families

- Model family M_1: $A \perp B$
  - Parameters: $\theta_1 = P(A = 1)$, $\theta_2 = P(B = 1)$
- Model family M_2:
  - Parameters: $\theta_1 = P(A = 1 \mid B = 1)$, $\theta_2 = P(A = 1 \mid B = 0)$, $\theta_3 = P(B = 1)$.
  - OR: $\theta_1 = P(B = 1 \mid A = 1)$, $\theta_2 = P(B = 1 \mid A = 0)$, $\theta_3 = P(A = 1)$.
  - OR: $\theta_1 = P(B = 1, A = 1)$, $\theta_2 = P(B = 1, A = 0)$, $\theta_3 = P(A = 0, B = 1)$.
- Hence, the model family M determines the necessary parameters, and fixing the values of the parameters θ produces a model instantiation (a joint distribution).

On learning and inference with independence models

- Assume n (binary) random variables $X_1, X_2, \ldots, X_n$.
- Inference/reasoning:
  - Working with an instantiated model $P(X_1, X_2, \ldots, X_n \mid M, \theta)$, compute the conditional probability distribution for the things you want to know, given all that you know, marginalizing out all that you don't know and don't want to know.
  - In principle exponential, requires $O(2^n)$ operations.
  - Can be simplified if the joint distribution factorizes by independence.

On learning and inference with independence models

- Learning/model selection:
  - Learn the model structure M: what is (conditionally) independent of what? What is the most probable model M maximizing $P(M \mid D)$?
  - Learn the parameters θ determining the "local" conditional distributions.
- Model averaging over model structures:

$$P(X \mid D) = \sum_M P(X \mid D, M) P(M \mid D).$$

Types of independence

Let X, Y, and Z be random variables.

- X and Y are *conditionally independent* given Z, written as $X \perp Y \mid Z$, if $\forall x, y, z$ with $P(Z = z) > 0$,

  $$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z).$$

- If $Z = \varnothing$, then X and Y are (marginally) independent.
- If two random variables are not (conditionally) independent, they are (conditionally) dependent.

Examples

- Lung cancer $\perp$ Yellow teeth | Smoking
  - But: Lung cancer $\not\perp$ Yellow teeth
- Child's genes $\perp$ Grandparent's genes | Parents'genes
  - But: Child's genes $\not\perp$ Grandparent's genes
- Ability of Team A $\perp$ Ability of Team B
  - But: Ability of Team A $\not\perp$ Ability of Team B | Outcome of A vs. B game

Independence saves space

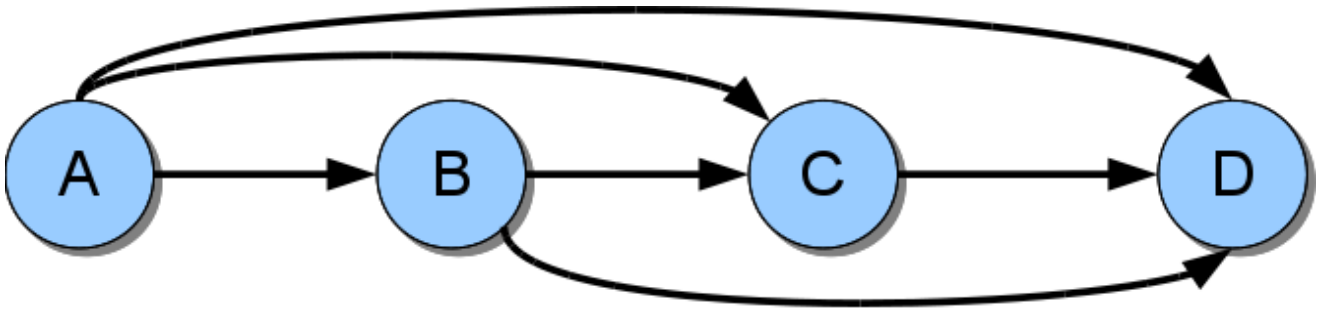
- If A and B are independent given C:

$$\begin{aligned} P(A,B,C) &= P(C)P(A \mid C)P(B \mid A,C) \\ &= P(C)P(A \mid C)P(B \mid C) \end{aligned}$$

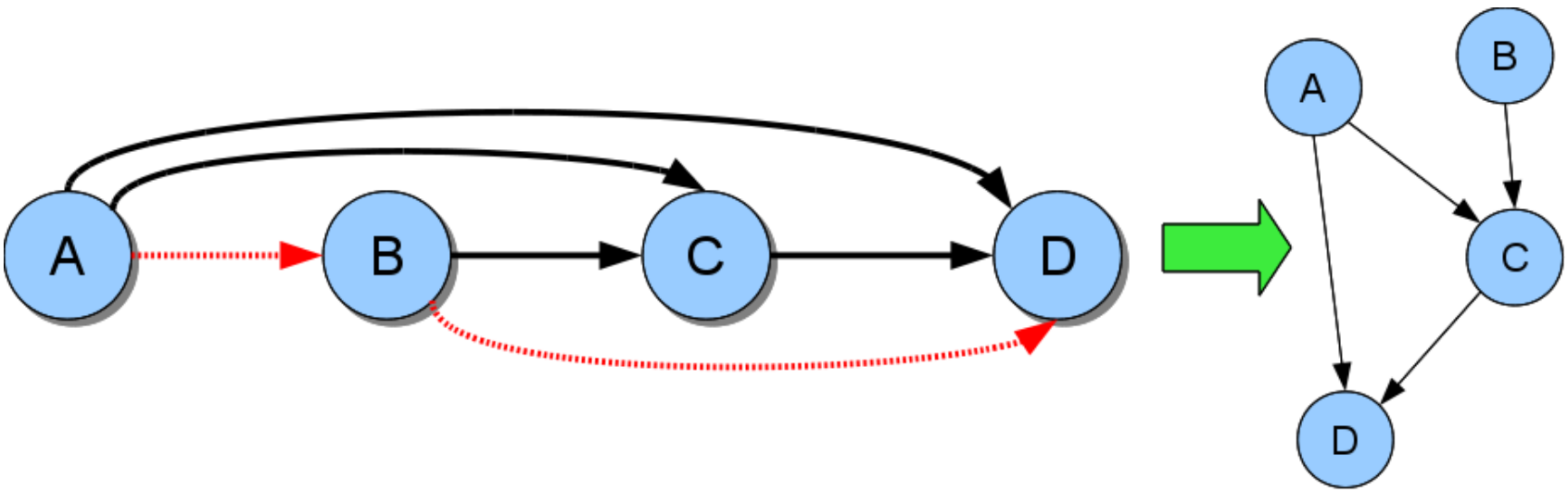
- Instead of having a full joint probability table for $P(A,B,C)$, we can have a table for $P(C)$ and tables $P(A \mid C = c)$ and $P(B \mid C = c)$ for each c.
  - Even for binary variables this saves space: $2^3 - 1 = 7$ vs. $1 + 2 + 2 = 5$.
  - With many variables and many independencies you save a lot.

Chain rule – Independence – BN

Chain rule: $P(A, B, C, D) = P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C)$

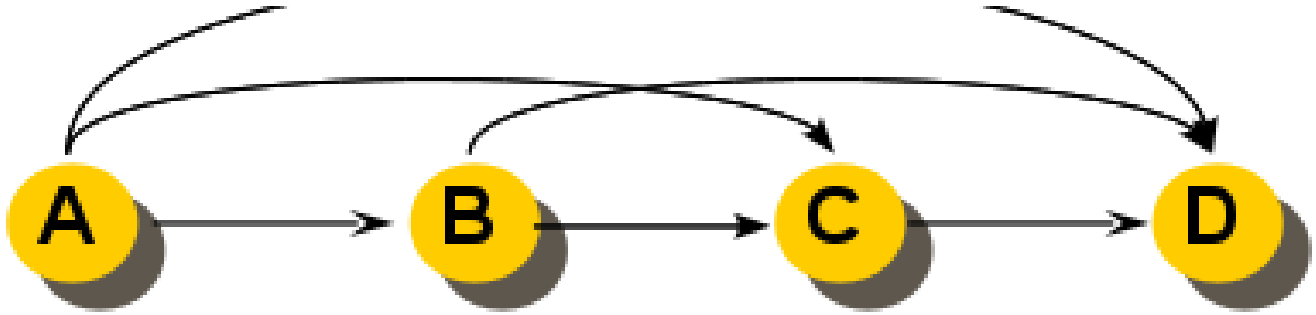


Independence: $P(A, B, C, D) = P(A)P(B)P(C \mid A, B)P(D \mid A, C)$

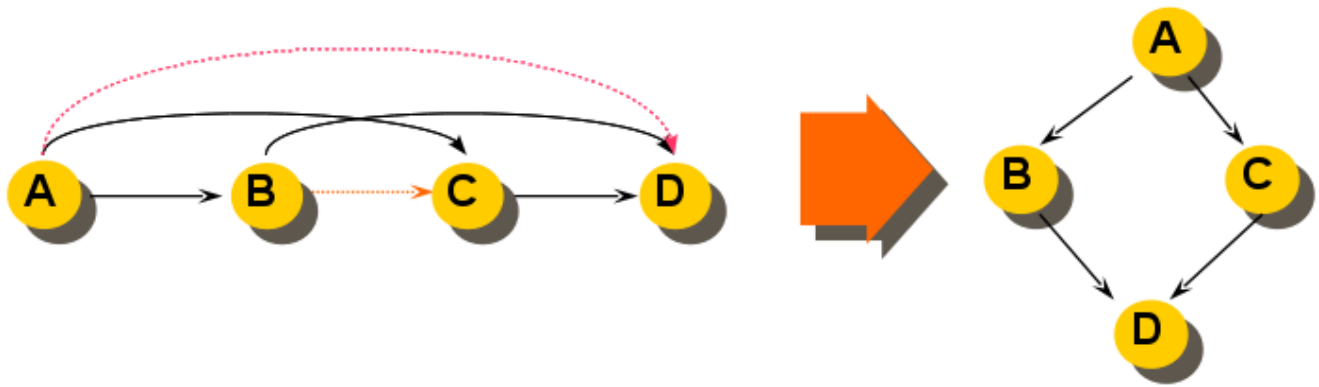


Another example

- $P(A, B, C, D) = P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C)$



- Assume $P(C \mid A, B) = P(C \mid A)$ and $P(D \mid A, B, C) = P(D \mid B, C)$

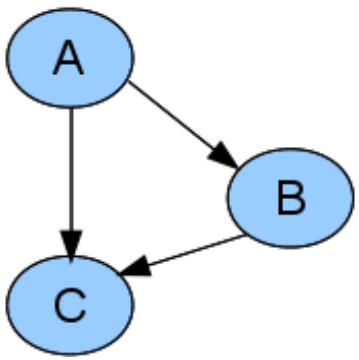


- Missing arcs encode conditional independence.

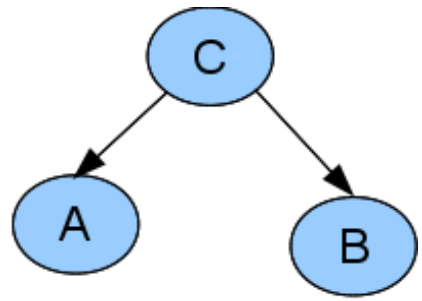
But order can matter

- $P(A, B, C) = P(C, A, B)$
- $P(A)P(B \mid A)P(C \mid A, B) = P(C)P(A \mid C)P(B \mid A, C)$
- And if A and B are conditionally independent given C:
  1. $P(A, B, C) = P(A)P(B \mid A)P(C \mid A, B)$
  2. $P(C, A, B) = P(C)P(A \mid C)P(B \mid C)$

1.



2.



And the point is?

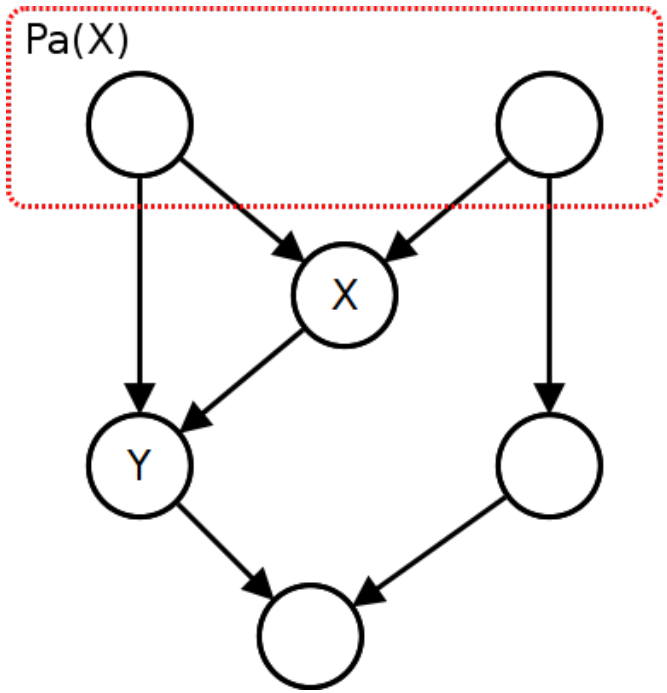
- Simple conditional probabilities are easier to determine than the full joint probabilities.
- In many domains, the underlying structure corresponds to relatively sparse networks so only a small number of conditional probabilities are needed

Bayesian networks: basics

- A Bayesian network is a model of probabilistic dependencies between the domain variables.
- The model can be described as a list of (in)dependencies, but is is usually more convenient to express them in a graphical form as a directed acyclic network.
- The nodes in the network correspond to the domain variables, and the arcs reveal the underlying dependencies, i.e., the hidden structure of the domain of your data.
- The "quantitative strengths" of the dependencies are modeled as conditional probability distributions (not shown in the graph).

Directed acyclic graph (DAG)

- A directed graph with no (directed) cycles.



- If there is an arc from X to Y, then X is called a parent of Y and Y is a child of X.
- The parents of of node X are denoted by $Pa_G(X)$.

Independencies and factorizations

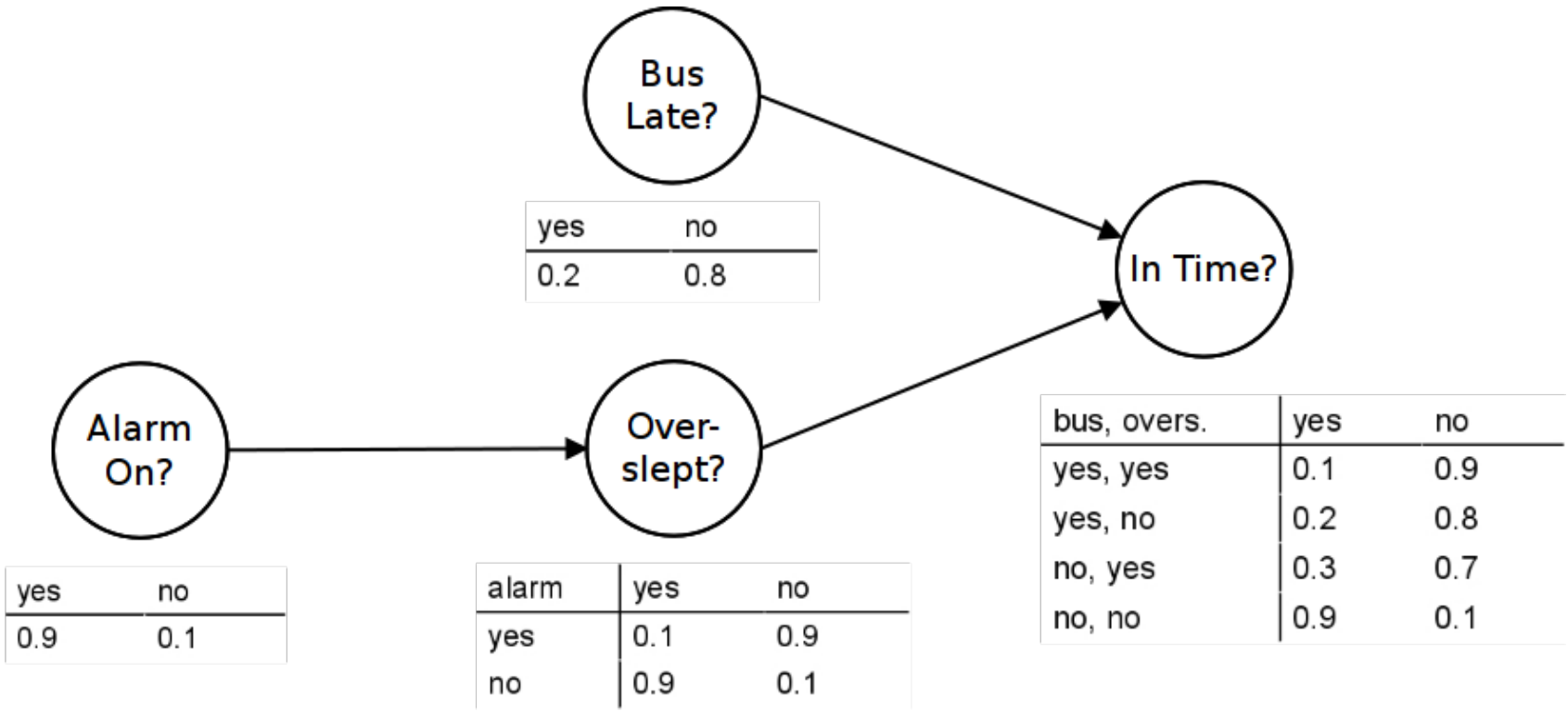
- Let P be a distribution. We define $\mathcal{I}(P)$ to be the set of independence assertions of the form $(X \perp Y \mid \mathbf{Z})$ that hold in P.
- Let G be a DAG. We say that a distribution P factorizes according to G if P can be expressed as a product

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_G(X_i)).$$

Bayesian networks: the textbook definition

A Bayesian (belief) network representation for a probability distribution P on a domain $(X_1, ..., X_n)$ is a pair (G, θ), where G is such a directed acyclic graph whose nodes correspond to the variables $X_1, ..., X_n$ that P factorizes according to G. The second component θ is a set consisting of all the conditional probabilities of the form $P(X|Pa_G(X))$.

A Bayesian network



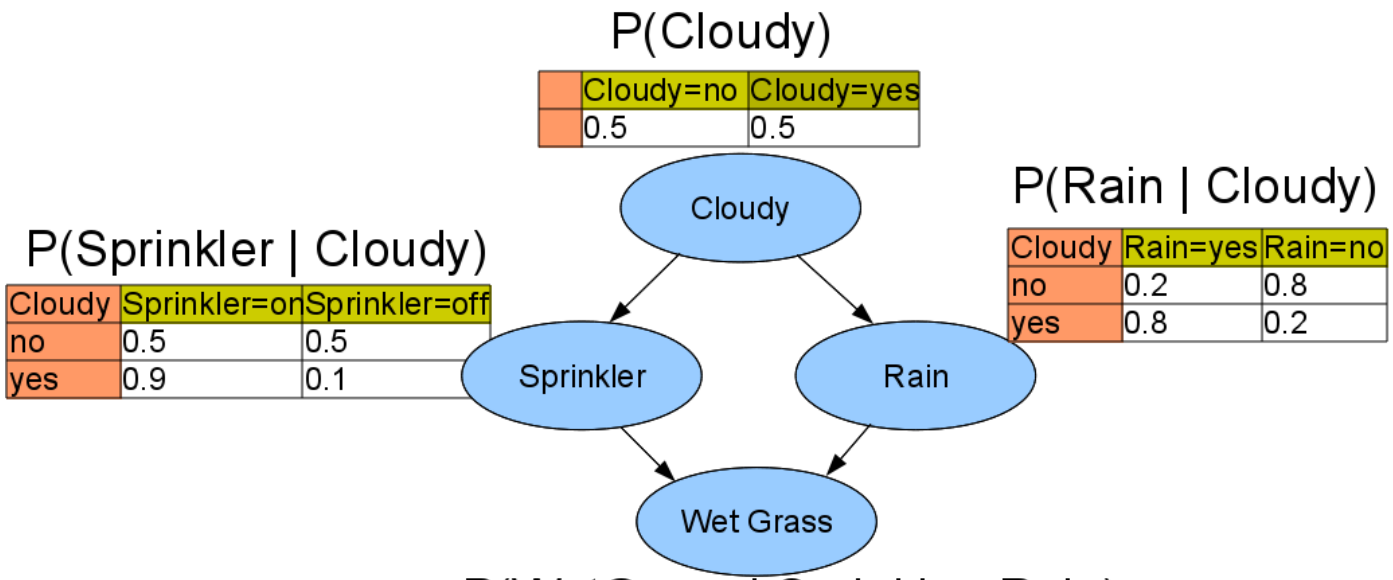
yes	no
0.2	0.8

bus, overs.	yes	no
yes, yes	0.1	0.9
yes, no	0.2	0.8
no, yes	0.3	0.7
no, no	0.9	0.1

yes	no
0.9	0.1

alarm	yes	no
yes	0.1	0.9
no	0.9	0.1

Another Bayesian network



Bayesian networks as factorizations

- If we have a DAG G, we denote the parents of the node (variable) X_i with $Pa_G(X_i)$ and a value configuration of $Pa_G(x_i)$ with $pa_G(x_i)$. Then

$$P(x_1, x_2, \ldots, x_n \mid G) = \prod_{i=1}^{n} P(x_i \mid pa_G(X_i)),$$

  where $P(x_i \mid pa_G(X_i))$ are called local probabilities.
- Local probabilities, that is, the conditional probability tables (CPTs) are determined by the parameters θ.

Compactness of a Bayesian network

- Assume we have n binary variables.
- To determine a CPT, one needs $2^n - 1$, that is, $O(2^n)$ parameters
- If the distribution factorizes according to a DAG where each node has at most k parents, one needs at most $O(n2^k)$ parameters.
- For example, if $n = 30$ and $k = 10$, a Bayesian network would have about 30000 parameters while a CPT would require over a billion parameters.

What do Bayesian networks have to offer?

- Encoding of the covariation between "input" variables – BN can handle incomplete data sets.
- Allows one to learn about causal relationships (predictions in the presence of interventions).
- Natural way of combining domain knowledge and data as a single model.
- Computationally efficient inference algorithms for multi-dimensional domains.

Bayesian networks?

- Very poor name, nothing "Bayesian" per se.
- A parametric probabilistic model that
  - Can be used for Bayesian inference (or not).
  - Can be learned via Bayesian methods (or not).
  - Is conveniently represented as a graph (a probabilistic graphical model).
  - Has a clear semantic foundation based on independencies.

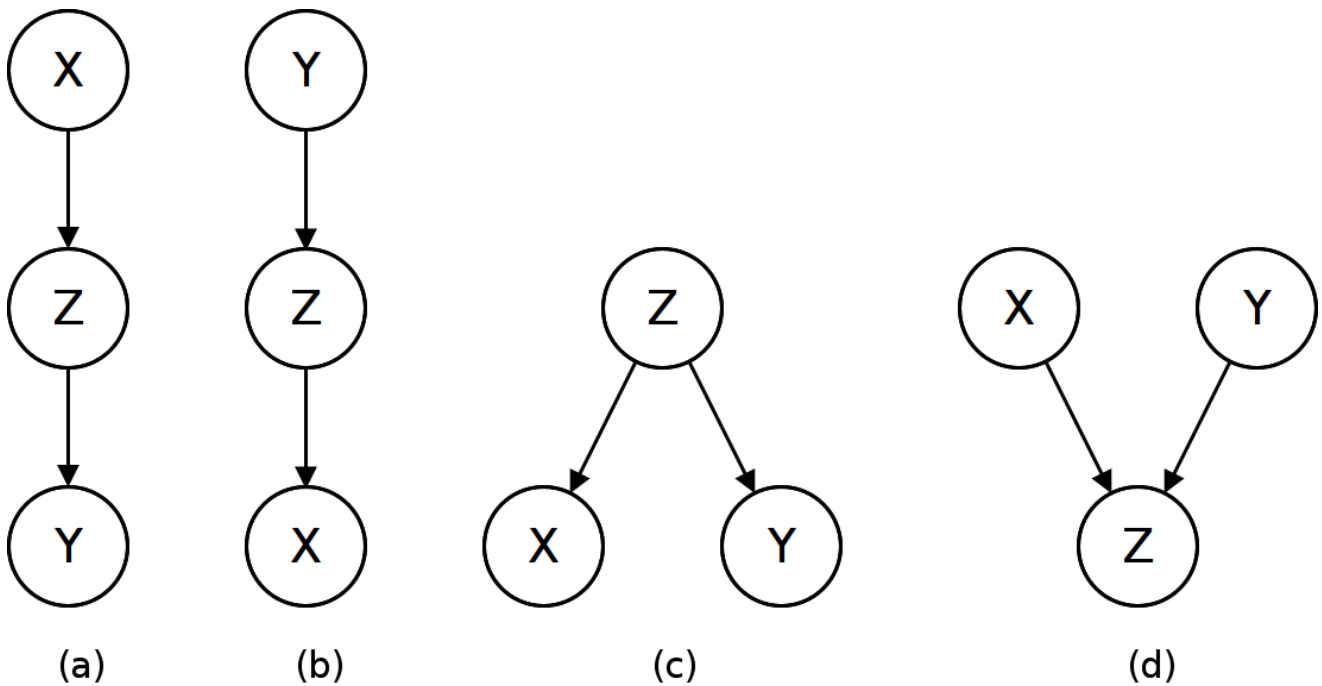
Independencies in a Bayesian network

- When can we guarantee that independence $X \perp Y \mid \mathbf{Z}$ holds (here $\mathbf{Z}$ is a set of variables) given that the distribution factorizes according to the DAG?
- Direct connections
  - If there is an arc $X \to Y$, then it is always possible to construct a distribution such that X and Y are correlated regardless of any information about other variables.
  - For example, X is uniformly distributed (regardless of its parents) and $Y = X$.

Indirect connections

How can information about X "flow" to Y?

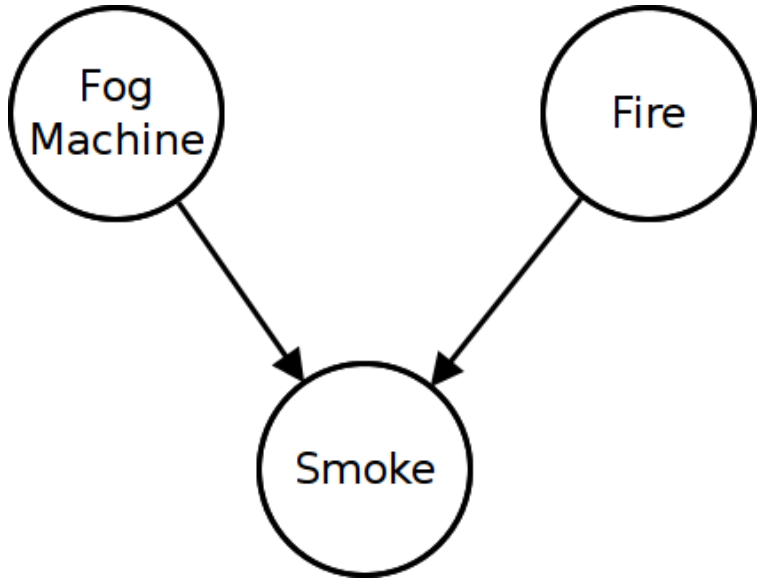
- a) Indirect causal effect
- b) Indirect evidential effect
- c) Common cause
- d) Common effect



Indirect connections

- Indirect causal effect, indirect evidential effect and common cause
  - Information can leak from X to Y if Z is not known. However, if Z is known, no information leaks through.
- Common effect
  - If Z is not known, no information can leak from X to Y. However, if Z is known, information can leak through.

Explaining away – an example



- Assume we have the following probabilities: the probability that there is a fire $P(F = 1) = 0.2$, the probability that the fog machine is on $P(M = 1) = 0.1$ and the probability of observing smoke given F and M as follows $P(S = 1 \mid F = 0, M = 0) = 0.1$, $P(S = 1 \mid F = 1, M = 0) = 0.8$, $P(S = 1 \mid F = 0, M = 1) = 0.8$, and $P(S = 1 \mid F = 1, M = 1) = 0.9$.

Explaining away – an example

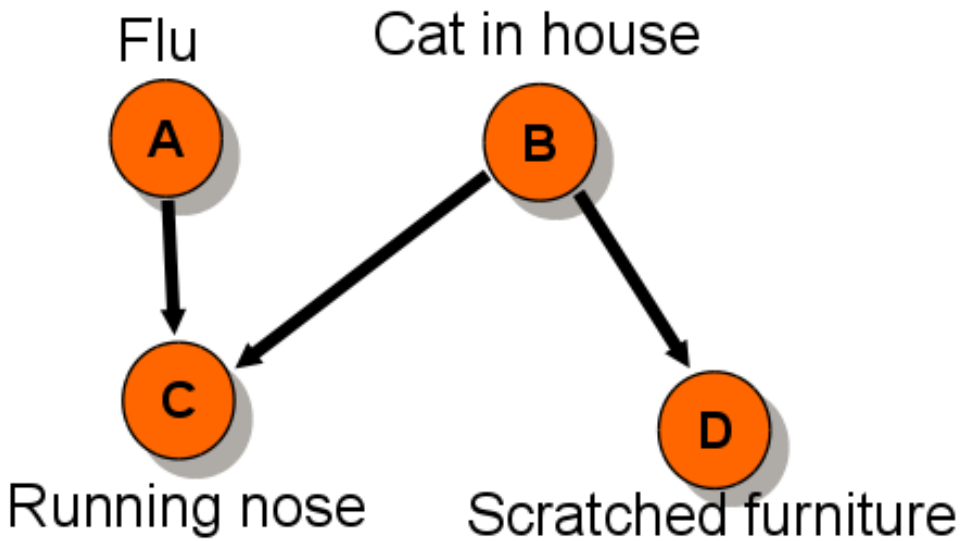
- Suppose we observe that there is smoke, that is, $S = 1$.
- Now, the probability of that there is a fire increases to $P(F = 1 \mid S = 1) = 0.54$.
- Subsequently we observe that a fog machine is on. Now $P(F = 1 \mid S = 1, M = 1) = 0.22$.
- F and M are not conditionally independent given S.

Explaining away

- Berkson's paradox, selection bias
- When the probability of one explanation increases, the alternative explanations become less probable (they are explained away).

Explaining away – another example

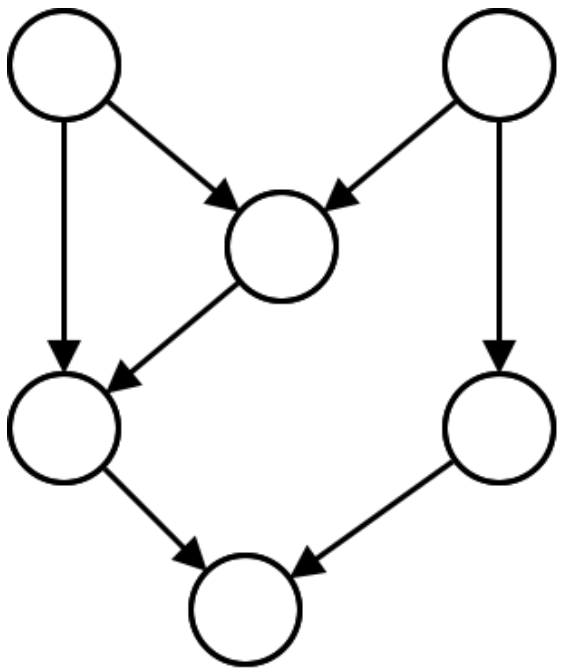
- Given $C = 1$, the probability of $A = 1$ is about 51%, and the probability of $B = 1$ is also about 51%.
- Given $C = 1$ and $D = 1$, the probability of $A = 1$ goes down to 13% while the probability of $B = 1$ goes up to 91%.



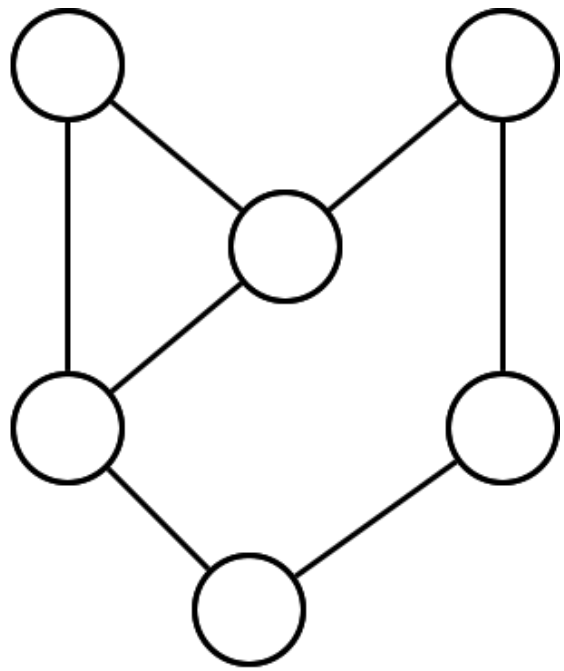
P(A=1)=0.05
P(B=1)=0.05
P(C=1|A=0,B=0)=0.001
P(C=1|A=1,B=0)=0.95
P(C=1|A=0,B=1)=0.95
P(C=1|A=1,B=1)=0.99
P(D=1|B=1)=0.99
P(D=1|B=0)=0.1

Skeleton

- Skeleton of a DAG is a undirected graph which is obtained by removing the arrowheads from the DAG.



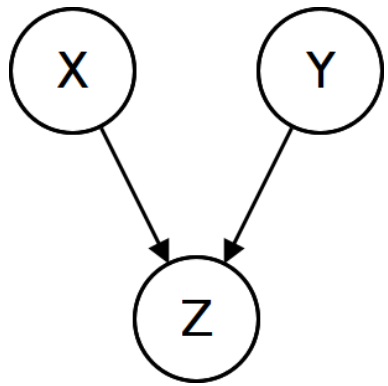
DAG



Skeleton

Trails and v-structures

- A trail in a BN is a cycle-free sequence (path) of edges in the corresponding skeleton.
- A node Z is a head-to-head node along a trail if there are two consecutive arc $X \rightarrow Z$ and $Z \leftarrow Y$ on that trail.

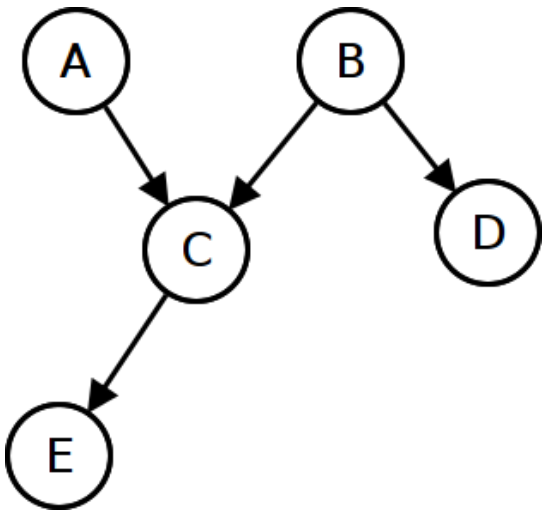


- If there are arcs $X \rightarrow Z$ and $Z \leftarrow Y$ and there is no arc between X and Y, then the triple (X, Z, Y) forms a *v-structure*.

d-separation

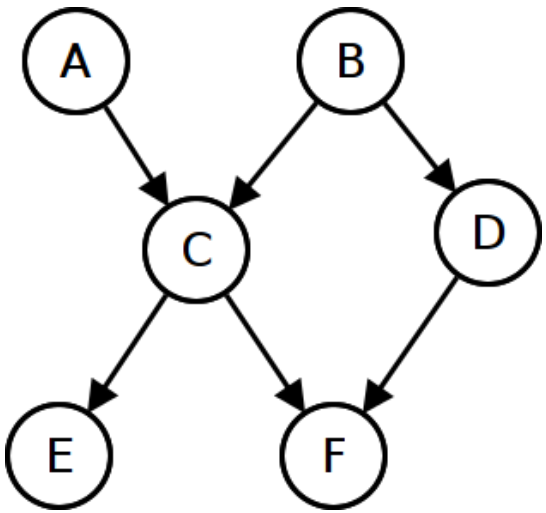
- Nodes X and Y are d-connected by nodes $\mathbf{Z}$ along a trail from X to Y if
  - Every head-to-head node along the trail is in $\mathbf{Z}$ or has a descendant in $\mathbf{Z}$.
  - None of the other nodes along the trail is in $\mathbf{Z}$.
- Nodes X and Y are d-separated by nodes $\mathbf{Z}$ if they are not d-connected by $\mathbf{Z}$ along any trail from X to Y.
- The set of independencies that correspond to d-separation is denoted by $\mathcal{I}(G)$.

Reading independencies



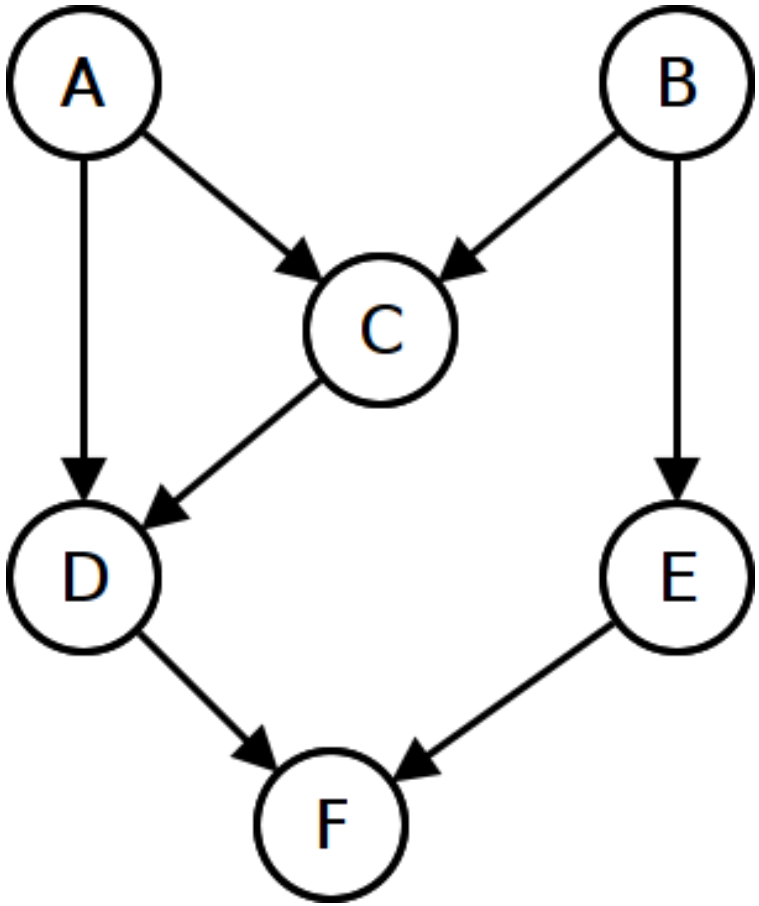
$$A \perp B$$
$$A \perp D$$
$$A \perp E \mid \{C\}$$
$$B \perp E \mid \{C\}$$
$$C \perp D \mid \{B\}$$
$$D \perp E \mid \{B\}$$

Reading independencies



$$A \perp B$$
$$A \perp D$$
$$A \perp E \mid \{C\}$$
$$A \perp F \mid \{C, B\}$$
$$B \perp E \mid \{C\}$$
$$B \perp F \mid \{C, D\}$$
$$C \perp D \mid \{B\}$$
$$D \perp E \mid \{B\}$$
$$E \perp F \mid \{C\}$$

Let's practice...



Soundness

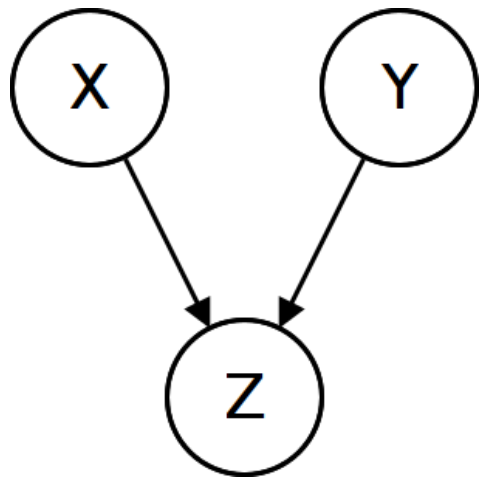
- If a distribution P factorizes according to G, then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$.
  - $\mathcal{I}(P)$ is the set of (conditional) independencies in P.

Faithfulness

- A distribution P a faithful to a graph G if whenever $(X \perp Y \mid \mathbf{Z}) \in \mathcal{I}(P)$ then $\mathrm{d}-\mathrm{sep}_G(X; Y \mid \mathbf{Z})$.
- In other words, parametrizations of the local probability distributions don't cause any additional independencies.
- For almost all distributions P that factorize over G, that is, for all distributions expect for a set of measure zero in the space of CPD parametrizations, we have $\mathcal{I}(G) = \mathcal{I}(P)$.

Example of unfaithfulness

- Consider three binary variables: X, Y, Z.
- $P(X = 0) = 0.5$, $P(Y = 0) = 0.5$, and $P(Z = 0 \mid X = 0, Y = 0) = 1$, $P(Z = 0 \mid X = 1, Y = 0) = 0$, $P(Z = 0 \mid X = 0, Y = 1) = 0$, and $P(Z = 0 \mid X = 1, Y = 1) = 1$ (XOR).



Expressiveness of Bayesian networks

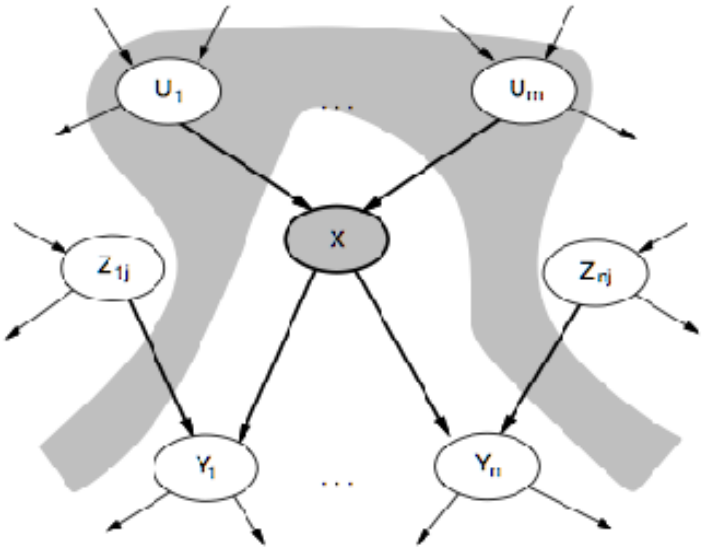
- Any distribution can be represented by a BN (complete graph).
- However, all sets of independence statements are not representable with DAGs.
  - Example, consider four variables A, B, C, and D. Assume that $A \perp D \mid \{B, C\}$ and $B \perp C \mid \{A, D\}$ and there are no other independencies.

Markov conditions

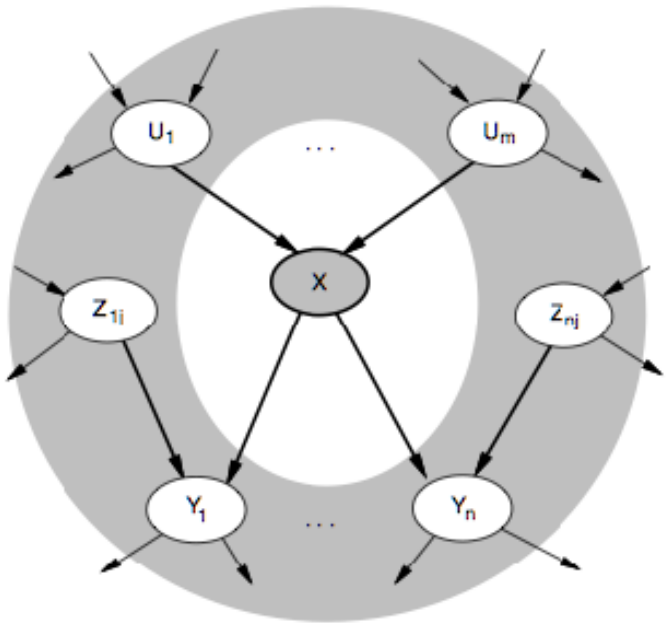
- The following conditions are equivalent in Bayesian networks:
  - Local (parental) Markov condition
    - X is independent of its non-descendants given its parents.
  - Global Markov condition
    - X and Y are independent given Z, iff they are d-separated by Z.
- Markov blanket
  - X is independent of any other set of variables given its parents, children, and the parents of its children (= Markov blanket).

Local Markov conditions visualized

From Russell & Norvig's book:

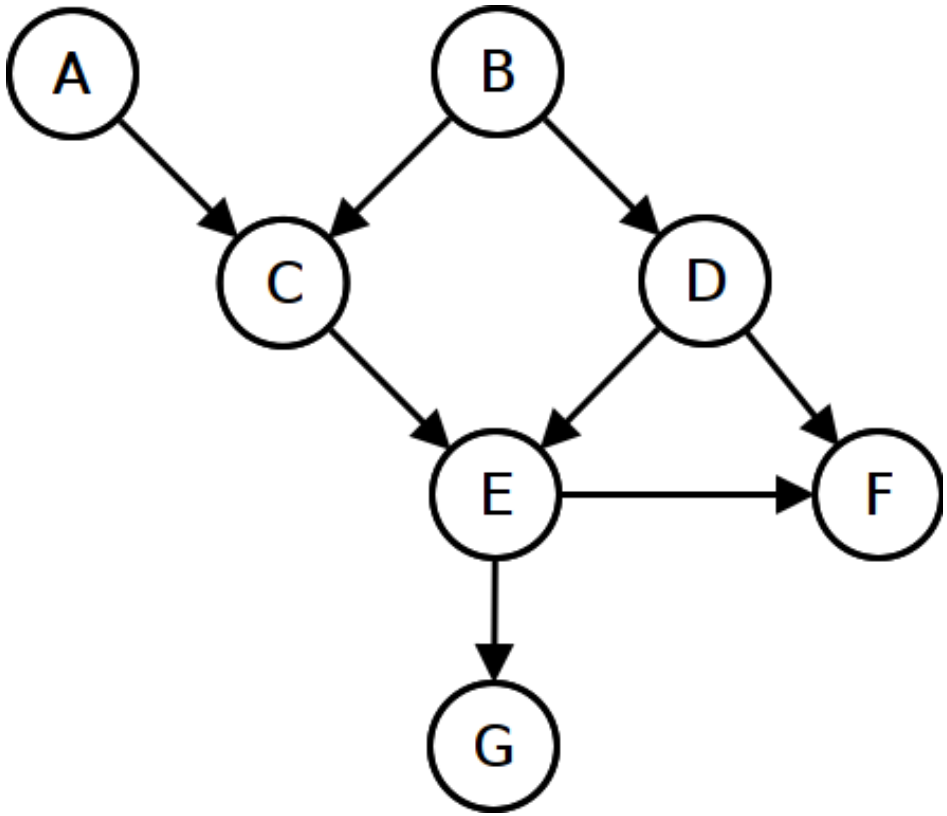


"X is conditionally independent of its non-descendants, given its parents"



"X is conditionally independent of all the other variables, given its Markov blanket"

Let's practice...



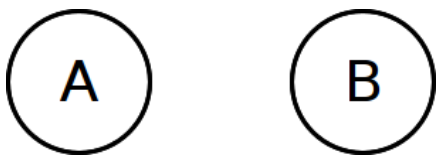
Equivalence classes

Back to the two-variable case...

Model M_1:

A and B are independent.

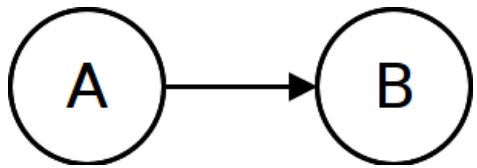
$P(A, B) = P(A)P(B)$



Model M_2:

A and B are dependent.

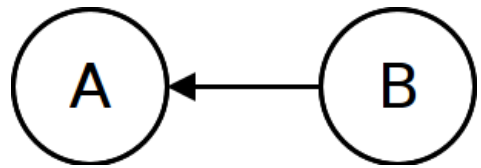
$P(A, B) = P(A)P(B \mid A)$



Model M_3:

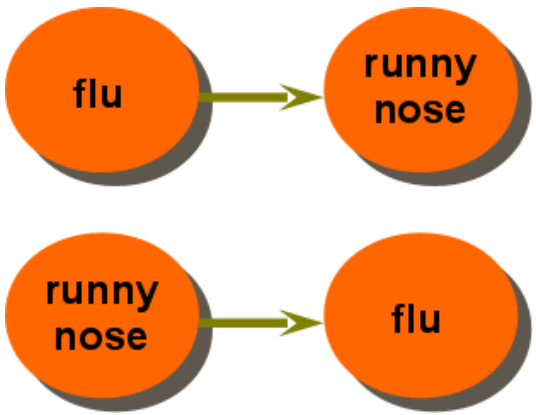
A and B are dependent.

$P(A, B) = P(B)P(A \mid B)$



Equivalence classes

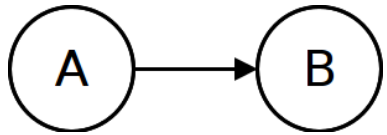
- An equivalence class is a set of BN structures which can used for representing exactly the same set of probability distributions.
- The "causally natural" version makes it easier to determine the conditional probabilities.

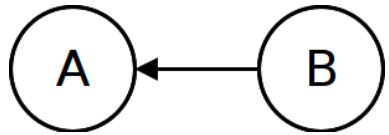


P(flu, ns) = P(flu)P(rn | flu)

P(flu, rn) = P(rn)P(flu| rn)

Bayes' theorem visualized

- $P_2(A, B) = P_2(A)P_2(B \mid A)$ 

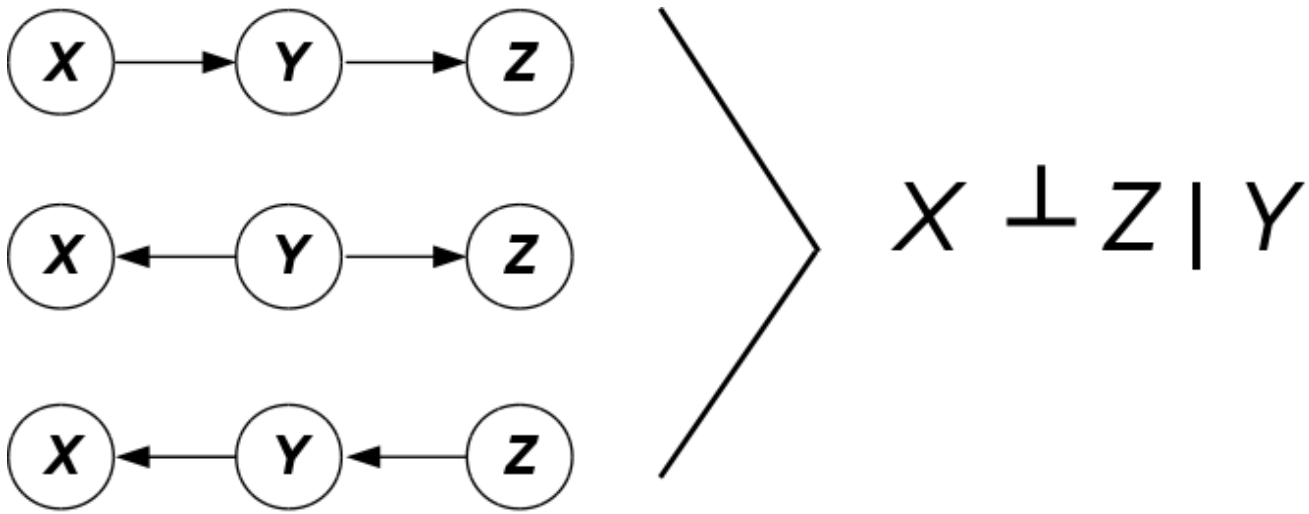
- $P_3(A, B) = P_3(B)P_3(A \mid B)$ 

- Assume $P_2(A)$ and $P_2(B \mid A)$ fixed.
- If $P_2(A, B) = P_3(A, B)$, then

$$P_3(A \mid B) = \frac{P_2(A)P_2(B \mid A)}{P_3(B)}$$

Equivalent network structures

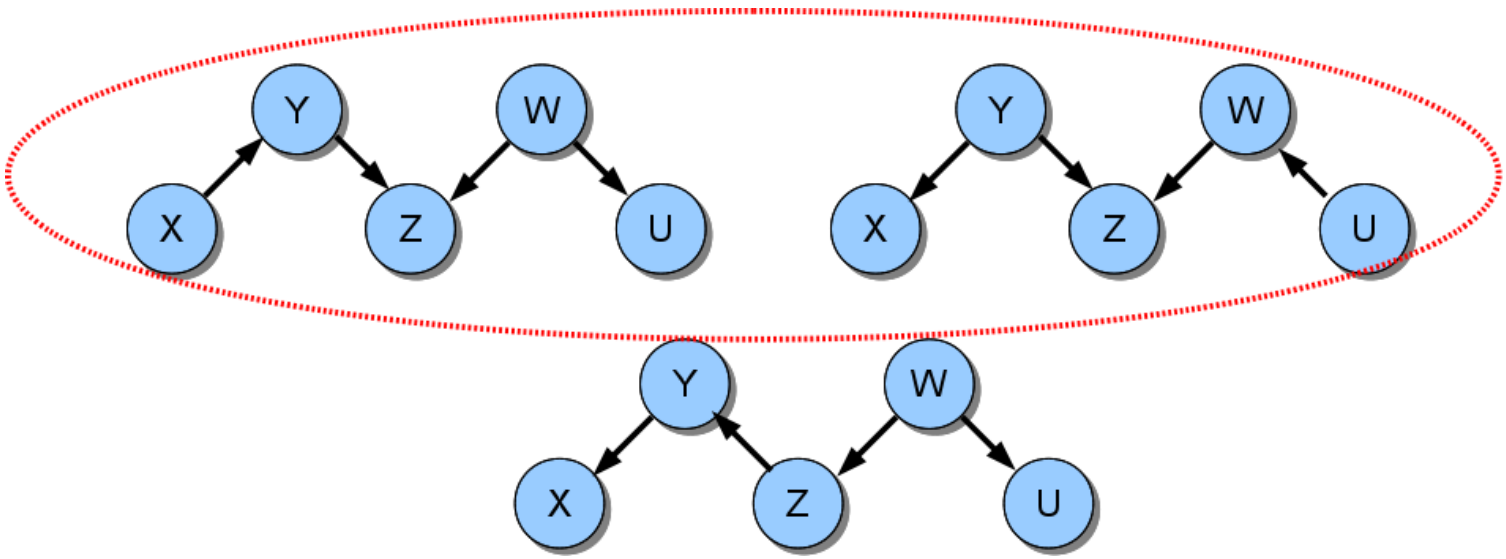
- Two network structures for the domain X are Markov equivalent (or independence equivalent) if they encode the same set of conditional independence statements.
- Example:



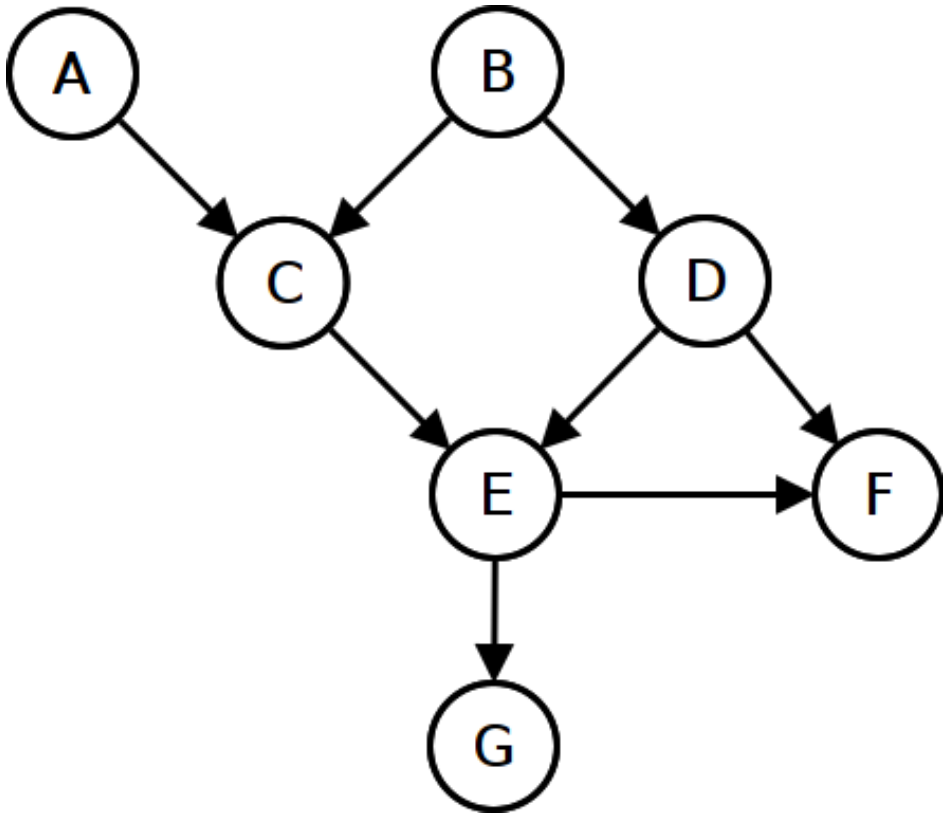
- Markov equivalence class: the set of (all) structures that encode the same set of conditional independence statements.

Equivalent network structures

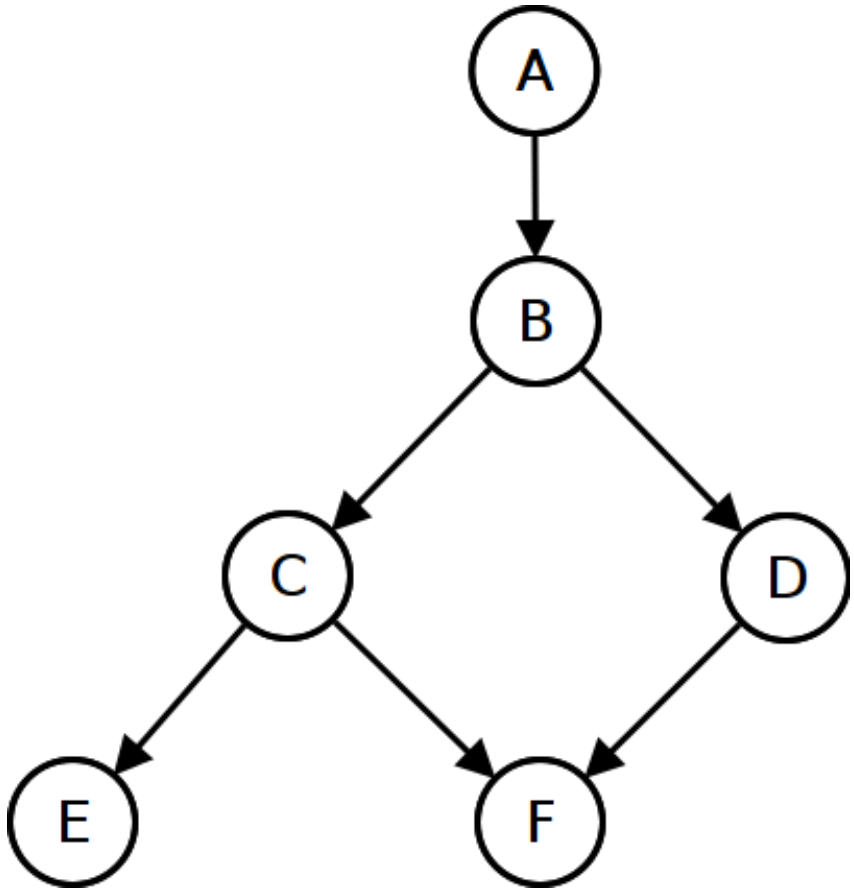
- Verma (1990): Two network structures are Markov equivalent if and only if:
  - They have the same skeleton.
  - They have the same v-structures.



Let's practice...



Let's practice...



Causal Bayesian networks

- A Bayesian network is causal of the arcs correspond to direct cause–effect relationships.
- It is possible to predict the consequences of interventions.
- Markov equivalent networks are not equivalent anymore.

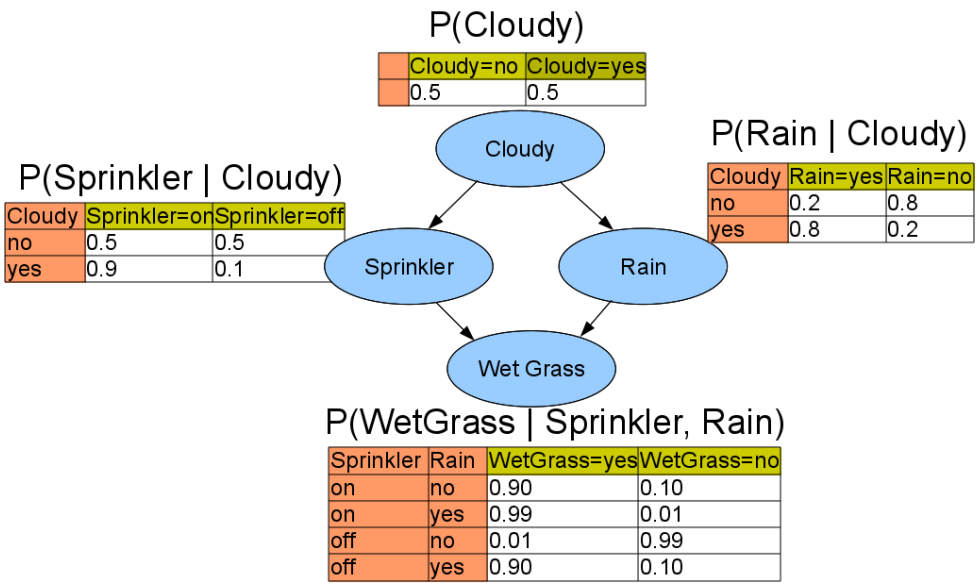
Causal order

- Causes first, then effects.
- Since causes render direct consequences independent yielding smaller CPTs.
- Causal CPTs are easier to assess by human experts.
- Smaller CPTs are easier to estimate reliably from a finite set of observations (data).

Bayesian networks as generative models

- How to generate random vectors from a Bayesian network?
  1. Order the nodes in a topological order.
  2. Generated values one by one in the topological order given all the previously generated values.

Example



- Sample parent first
  - $P(C)$: $(0.5, 0.5) \rightarrow yes$
  - $P(S \mid C = yes)$: $(0.9, 0.1) \rightarrow on$
  - $P(R \mid C = yes)$: $(0.8, 0.2) \rightarrow no$
  - $P(W \mid S = on, R = no)$: $(0.9, 0.1) \rightarrow yes$
- $P(C, S, R, W) = P(yes, on, no, yes) = 0.5 \times 0.9 \times 0.2 \times 0.9 = 0.081.$

Further readings

- Barber (2011), Chapter 3.
- Neapolitan (2004), Chapters 1.3 and 2.