Bayesian networks

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Significant parts of this material come from
the lectures on Bayesian networks which are part of
Artificial Intelligence course by Pieter Abbeel and Dan Klein.
The original lectures can be found at `http://ai.berkeley.edu`

Introduction

Uncertainty

Probabilistic reasoning is one of the frameworks that allow us to maintain our beliefs and knowledge in uncertain environments.

Uncertainty

Probabilistic reasoning is one of the frameworks that allow us to maintain our beliefs and knowledge in uncertain environments.

Usual scenario:

- **Observed variables X (evidence):** known things related to the state of the world; often imprecise, noisy (info from sensors, symptoms of a patient, etc.).
- **Unobserved, hidden variables K:** unknown, but important aspects of the world; we need to reason about them (what the position of an object is, whether a disease is present, etc.)
- **Model:** describes the relations among hidden and observed variables; allows us to reason.

Uncertainty

Probabilistic reasoning is one of the frameworks that allow us to maintain our beliefs and knowledge in uncertain environments.

Usual scenario:

- **Observed variables X (evidence):** known things related to the state of the world; often imprecise, noisy (info from sensors, symptoms of a patient, etc.).
- **Unobserved, hidden variables K:** unknown, but important aspects of the world; we need to reason about them (what the position of an object is, whether a disease is present, etc.)
- **Model:** describes the relations among hidden and observed variables; allows us to reason.

Models (including probabilistic)

- describe how (a part of) the world works.
- are always approximations or simplifications:
  - They cannot acount for everything (they would be as complex as the world itself).
  - They represent only a chosen subset of variables and interactions between them.
  - "All models are wrong; some are useful." — George E. P. Box

Uncertainty

Probabilistic reasoning is one of the frameworks that allow us to maintain our beliefs and knowledge in uncertain environments.

Usual scenario:

- **Observed variables X (evidence):** known things related to the state of the world; often imprecise, noisy (info from sensors, symptoms of a patient, etc.).
- **Unobserved, hidden variables K:** unknown, but important aspects of the world; we need to reason about them (what the position of an object is, whether a disease is present, etc.)
- **Model:** describes the relations among hidden and observed variables; allows us to reason.

Models (including probabilistic)

- describe how (a part of) the world works.
- are always approximations or simplifications:
  - They cannot acount for everything (they would be as complex as the world itself).
  - They represent only a chosen subset of variables and interactions between them.
  - "All models are wrong; some are useful." — George E. P. Box

A **probabilistic model** is a joint distribution over a set of random variables.

Notation

Random variables (start with capital letters):

$$K, X, Weather, \dots$$

Values of random variables (start with lower-case letters):

$$x_1, e_i, rainy, \dots$$

Probability distribution of a random variable:

$P(X)$ or P_X

Probability of a random event:

$P(X = x_1)$ or $P_X(x_1)$

Shorthand for a probability of a random event (if there is no chance of confusion):

$P(+r)$ meaning $P(Rainy = true)$ or
$P(r)$ meaning $P(Weather = rainy)$

Probability cheatsheet

Conditional probability:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Product rule:

$$P(X,Y) = P(X|Y)P(Y)$$

Bayes rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_i P(y|x_i)P(x_i)}$$

Chain rule:

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdot \ldots = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

$X \perp\!\!\!\perp Y$ (X and Y are **independent**) iff

$$\forall x, y : P(x,y) = P(x)P(y)$$

$X \perp\!\!\!\perp Y|Z$ (X and Y are **conditinally independent** given Z) iff

$$\forall x, y, z : P(x,y|z) = P(x|z)P(y|z)$$

Joint probability distribution

Joint distribution over a set of variables $X_1, \ldots, X_n$ (here descrete) assigns a probability to each combination of values:

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(x_1, \ldots, x_n)$$

For a proper probability distribution:

$$\forall x_1, \ldots, x_n : P(x_1, \ldots, x_n) \geq 0 \qquad \text{and} \qquad \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) = 1$$

Joint probability distribution

Joint distribution over a set of variables $X_1, \ldots, X_n$ (here descrete) assigns a probability to each combination of values:

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(x_1, \ldots, x_n)$$

For a proper probability distribution:

$$\forall x_1, \ldots, x_n : P(x_1, \ldots, x_n) \geq 0 \qquad \text{and} \qquad \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) = 1$$

Probabilistic inference

- Compute a desired probability from other known probabilities (e.g. marginal or conditional from joint).
- **Conditional probabilities** turn out to be the most interesting ones:
  - They represent our or agent's beliefs given the evidence (measured values of observable variables).
  - $P(\text{bus on time}|\text{rush our}) = 0.8$
- Probabilities change with new evidence:
  - $P(\text{bus on time}) = 0.95$
  - $P(\text{bus on time}|\text{rush our}) = 0.8$
  - $P(\text{bus on time}|\text{rush our, dry roads}) = 0.85$

Contents

- What is a Bayesian network?
- How it encodes the joint probability distributions?
- What independence assumptions does it encode?
- How to perform reasoning using BN?

Bayesian networks

What's wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$):

What's wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$): $n_{\text{params}} = 2^n - 1$
- In general:

What’s wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$): $n_{\text{params}} = 2^n - 1$
- In general: $n_{\text{params}} \geq d^n - 1$

What's wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$): $n_{\text{params}} = 2^n - 1$
- In general: $n_{\text{params}} \geq d^n - 1$

Two issues with full joint probability distribution:

- It is usually *too large* to represent explicitly!
- It is very hard to learn (estimate from data, or elicit from domain experts) the vast number of parameters!

What's wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$): $n_{\text{params}} = 2^n - 1$
- In general: $n_{\text{params}} \geq d^n - 1$

Two issues with full joint probability distribution:

- It is usually *too large* to represent explicitly!
- It is very hard to learn (estimate from data, or elicit from domain experts) the vast number of parameters!

Bayesian networks (BN) can represent (or approximate) complex joint distributions (models) using simple, local distributions (conditional probabilities), if we are willing to impose some conditional independence assumptions on the domain.

- We describe how *variables locally interact*.
- Local *interactions chain together* to give global, indirect interactions.
- BN requires *less parameters* than full joint distribution.
- The network structure and the local probability tables can be easilly elicited from domain experts, or learned from less data.

What's wrong with the joint distribution?

How many free parameters n_{params} has a probability distribution over n variables, each having at least d possible values?

- For all variables binary ($d = 2$): $n_{\text{params}} = 2^n - 1$
- In general: $n_{\text{params}} \geq d^n - 1$

Two issues with full joint probability distribution:

- It is usually *too large* to represent explicitly!
- It is very hard to learn (estimate from data, or elicit from domain experts) the vast number of parameters!

Bayesian networks (BN) can represent (or approximate) complex joint distributions (models) using simple, local distributions (conditional probabilities), if we are willing to impose some conditional independence assumptions on the domain.

- We describe how *variables locally interact*.
- Local *interactions chain together* to give global, indirect interactions.
- BN requires *less parameters* than full joint distribution.
- The network structure and the local probability tables can be easilly elicited from domain experts, or learned from less data.

Other names for BN:

- belief network, probabilistic network, causual network, knowledge map
- directed probabilistic graphical model

What is a Bayesian network?

A full joint probability distribution can *always* be *factorized into a product of conditional distributions*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_n | X_1, \ldots, X_{i-1}),$$

which can be *simplified using (conditional) independence assumptions*. In the extreme case, when all the variables are independent, the above simplifies to

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

What is a Bayesian network?

A full joint probability distribution can *always* be *factorized into a product of conditional distributions*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_n | X_1, \ldots, X_{i-1}),$$

which can be *simplified using (conditional) independence assumptions*. In the extreme case, when all the variables are independent, the above simplifies to

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

Bayesian network is a probabilistic graphical model that encodes such a factorization. It is defined by a *directed acyclic graph (DAG)* with

- a set of *nodes* representing the random variables,
- oriented *edges* representing the direct influences among variables, and
- *(un)conditional probability distributions* describing the probability distribution of each random variable given all its parents (i.e., not given all the preceding variables).

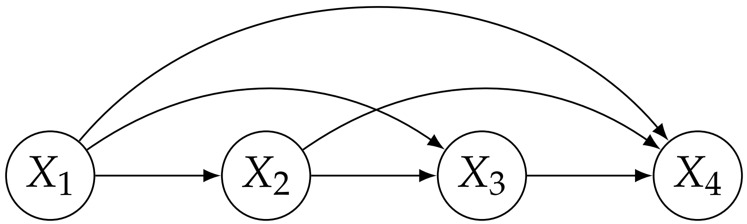
BN represents the following factorization of the joint probability:

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

What is a Bayesian network?

A full joint probability distribution can *always* be *factorized into a product of conditional distributions*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_n | X_1, \ldots, X_{i-1}),$$



which can be *simplified using (conditional) independence assumptions*. In the extreme case, when all the variables are independent, the above simplifies to

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

Bayesian network is a probabilistic graphical model that encodes such a factorization. It is defined by a *directed acyclic graph (DAG)* with

- a set of *nodes* representing the random variables,
- oriented *edges* representing the direct influences among variables, and
- *(un)conditional probability distributions* describing the probability distribution of each random variable given all its parents (i.e., not given all the preceding variables).

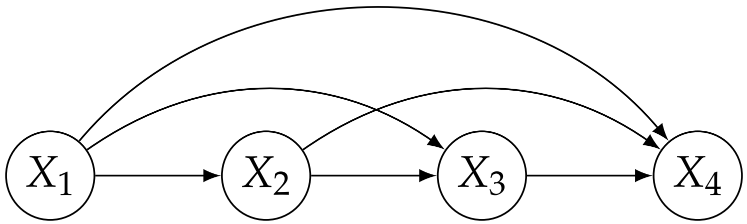
BN represents the following factorization of the joint probability:

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

What is a Bayesian network?

A full joint probability distribution can *always* be *factorized into a product of conditional distributions*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_n | X_1, \ldots, X_{i-1}),$$



which can be *simplified using (conditional) independence assumptions*. In the extreme case, when all the variables are independent, the above simplifies to

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

Bayesian network is a probabilistic graphical model that encodes such a factorization. It is defined by a *directed acyclic graph (DAG)* with

- a set of *nodes* representing the random variables,
- oriented *edges* representing the direct influences among variables, and
- *(un)conditional probability distributions* describing the probability distribution of each random variable given all its parents (i.e., not given all the preceding variables).

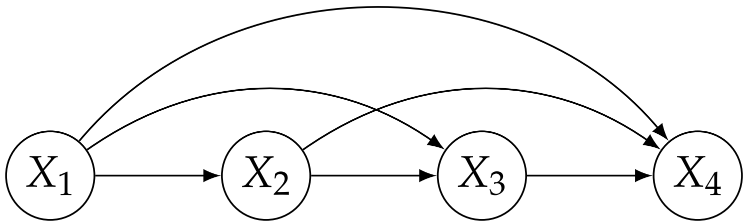
BN represents the following factorization of the joint probability:

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

What is a Bayesian network?

A full joint probability distribution can *always* be *factorized into a product of conditional distributions*

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_n | X_1, \dots, X_{i-1}),$$



which can be *simplified using (conditional) independence assumptions*. In the extreme case, when all the variables are independent, the above simplifies to

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i).$$

Bayesian network is a probabilistic graphical model that encodes such a factorization. It is defined by a *directed acyclic graph (DAG)* with

- a set of *nodes* representing the random variables,
- oriented *edges* representing the direct influences among variables, and
- *(un)conditional probability distributions* describing the probability distribution of each random variable given all its parents (i.e., not given all the preceding variables).

BN represents the following factorization of the joint probability:

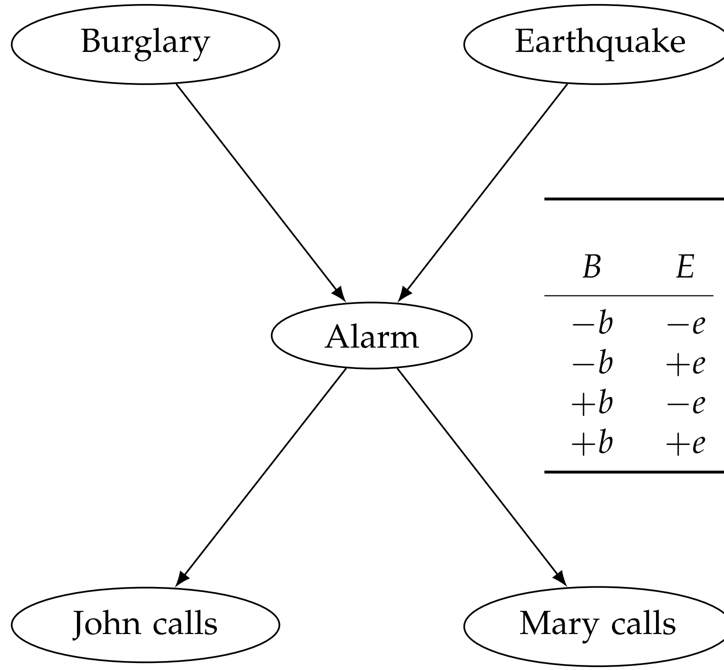
$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

A particular BN (usually) cannot represent any joint distribution!

BN example

$P(B)$	
$+b$	$-b$
0.001	0.999

$P(E)$	
$+e$	$-e$
0.002	0.998



		$P(A\|B,E)$	
B	E	$+a$	$-a$
$-b$	$-e$	0.001	0.999
$-b$	$+e$	0.29	0.71
$+b$	$-e$	0.94	0.06
$+b$	$+e$	0.95	0.05

	$P(J\|A)$	
A	$+j$	$-j$
$-a$	0.05	0.95
$+a$	0.9	0.1

	$P(M\|A)$	
A	$+m$	$-m$
$-a$	0.01	0.99
$+a$	0.7	0.3

Introduction

Bayesian networks
- Issues
- BN
- BN example
- Independence
- Independence?
- Conditional independence
- Causality
- Assumptions in BN
- Independence in BN
- Causal chain
- Common cause
- Common effect
- D-separation
- D-sep examples

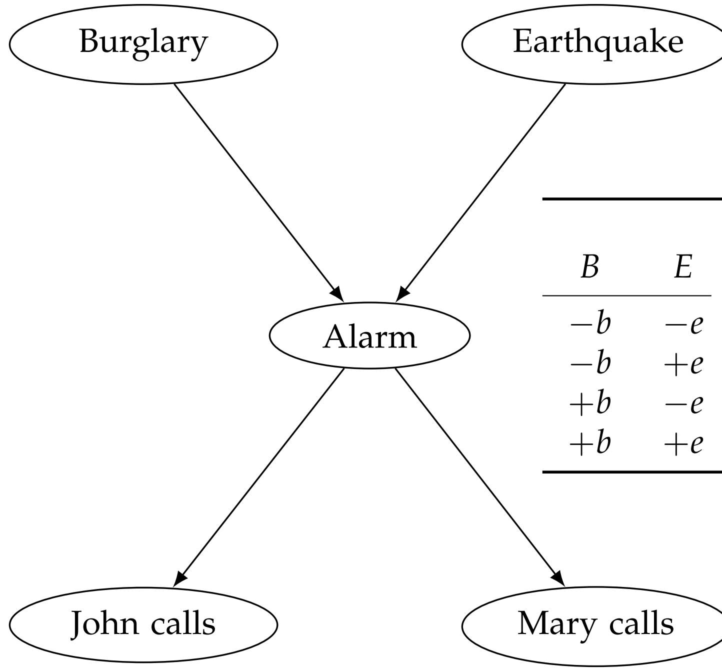
Inference

Summary

BN example

$P(B)$	
$+b$	$-b$
0.001	0.999

$P(E)$	
$+e$	$-e$
0.002	0.998



		$P(A\|B,E)$	
B	E	$+a$	$-a$
$-b$	$-e$	0.001	0.999
$-b$	$+e$	0.29	0.71
$+b$	$-e$	0.94	0.06
$+b$	$+e$	0.95	0.05

	$P(J\|A)$	
A	$+j$	$-j$
$-a$	0.05	0.95
$+a$	0.9	0.1

	$P(M\|A)$	
A	$+m$	$-m$
$-a$	0.01	0.99
$+a$	0.7	0.3

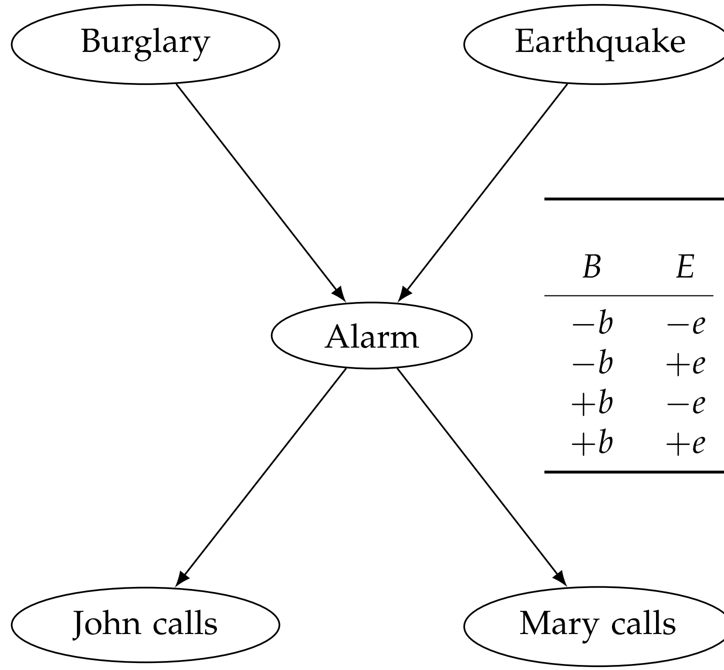
The joint probability is factorized by this BN as

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

BN example

$P(B)$	
$+b$	$-b$
0.001	0.999

$P(E)$	
$+e$	$-e$
0.002	0.998



		$P(A\|B,E)$	
B	E	$+a$	$-a$
$-b$	$-e$	0.001	0.999
$-b$	$+e$	0.29	0.71
$+b$	$-e$	0.94	0.06
$+b$	$+e$	0.95	0.05

	$P(J\|A)$	
A	$+j$	$-j$
$-a$	0.05	0.95
$+a$	0.9	0.1

	$P(M\|A)$	
A	$+m$	$-m$
$-a$	0.01	0.99
$+a$	0.7	0.3

The joint probability is factorized by this BN as

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

What is the probability of $+b, -e, -a, +j, -m$?

$$P(+b,-e,-a,+j,-m) = P(+b)P(-e)P(-a|+b,-e)P(+j|-a)P(-m|-a) =$$
$$= 0.001 \cdot 0.998 \cdot 0.06 \cdot 0.05 \cdot 0.99 \doteq 3 \cdot 10^{-6}$$

Independence

Two variables X and Y are independent ($X \perp\!\!\!\perp Y$) iff

$$\forall x, y : P(x, y) = P(x)P(y),$$

which implies that

$$\forall x, y : P(x|y) = P(x) \qquad \text{and} \qquad \forall x, y : P(y|x) = P(y)$$

Independence as a modeling assumption:

- Empirical distributions are at best "close to independence"; assuming independence may thus be too strong.
- Nevertheless, sometimes a reasonable assumption; what can we assume about variables $Weather, Umbrella, Cavity, Toothache$?
- Example: Having n unfair, but independent coin flips:
  - A general joint $P(X_1, \ldots, X_n)$ with no assumptions has $2^n - 1$ free parameters.
  - $P(X_1, \ldots, X_n)$ factorized using independence assumptions to $P(X_1) \cdot \ldots \cdot P(X_n)$ has just n free parameters.

How to check independence?

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$P_2(T, W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

1. Compute marginal distributions of individual variables ($P(T), P(W)$) from the joint distribution (P_1).
2. Create a new joint distribution (P_2) from the marginals assuming independence of the variables.
3. Is the new joint the same as the original one? Then the variables are indeed independent.

Conditional independence

Two variables X and Y are conditionally independent given another variable Z ($X \perp\!\!\!\perp Y | Z$) iff

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z),$$

which implies that

$$\forall x, y, z : P(x|y, z) = P(x|z) \qquad \text{and} \qquad \forall x, y, z : P(y|x, z) = P(y|z)$$

Conditional independence

Two variables X and Y are conditionally independent given another variable Z ($X \perp\!\!\!\perp Y | Z$) iff

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z),$$

which implies that

$$\forall x, y, z : P(x|y, z) = P(x|z) \qquad \text{and} \qquad \forall x, y, z : P(y|x, z) = P(y|z)$$

Conditional independence as a modeling assumption:

- It is our most basic and robust form of knowledge about uncertain environments.
- In practice, measuring certain variable often breaks mutual influence of 2 other variables (or vice versa, it introduces influence amonge variables that were originally independent).
- Conditional independence assumptions are very suitable to model real world!

Causality

Suppose we want to model 2 variables:

- R: Does it rain?
- T: Is there high traffic?

Which of the 2 models is correct?

$$R \longrightarrow T$$

$$P(R,T) = P(R)P(T|R)$$

$$R \longleftarrow T$$

$$P(R,T) = P(T)P(R|T)$$

Causality

Suppose we want to model 2 variables:

- R: Does it rain?
- T: Is there high traffic?

Which of the 2 models is correct?

$$R \longrightarrow T$$

$$P(R,T) = P(R)P(T|R)$$

$$R \longleftarrow T$$

$$P(R,T) = P(T)P(R|T)$$

- In this case for 2 variables, both models can represent any joint distribution over R and T.
- We prefer the *causal* orientation (rain influences/causes traffic, not vice versa) because
  - the structure is then more intuitive and describes how things work in the world;
  - the resulting BN is often simpler (nodes have fewer parents);
  - the conditional probabilities are easier to obtain.
- In practice, *BN needn't be causal*, especially when variables are missing.
  - Imagine variables $YellowFingers$ and $Cancer$. They are correlated, but neither causes the other. Both are caused by smoking (which is a missing variable).
  - Arrows can reflect *correlation, not causation*.
- What do the arrows really mean?
  - They define BN topology which may happen to encode causal structure.
  - BN topology defines the factorization of the joint distribution, i.e. the conditional independence assumptions.

Assumptions in BN

- Each BN defines a factorization of the joint distribution.
- The factorization is possible due to (conditional) independence assumptions we are willing to make:

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

- Beyond the above "chain rule $\rightarrow$ BN" explicit conditional independence assumptions, often additional implicit assumptions exist. (They can be read off the graph.)
- For modeling, it is important to understand all the assumptions made when the BN graph is chosen.

Assumptions in BN

- Each BN defines a factorization of the joint distribution.
- The factorization is possible due to (conditional) independence assumptions we are willing to make:

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

- Beyond the above "chain rule $\rightarrow$ BN" explicit conditional independence assumptions, often additional implicit assumptions exist. (They can be read off the graph.)
- For modeling, it is important to understand all the assumptions made when the BN graph is chosen.

Example:

$$X \longrightarrow Y \longrightarrow Z \longrightarrow W$$

- This BN enforces the following simplification of the chain rule:

$$P(X)P(Y|X)P(Z|X,Y)P(W|X,Y,Z) = P(X)P(Y|X)P(Z|Y)P(W|Z)$$

- Explicit assumptions from these simplifications:

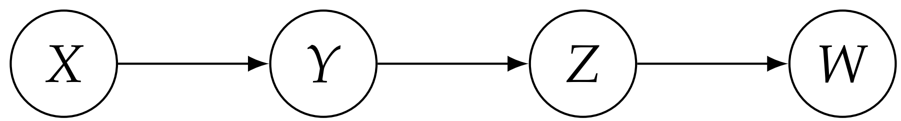
Assumptions in BN

- Each BN defines a factorization of the joint distribution.
- The factorization is possible due to (conditional) independence assumptions we are willing to make:

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

- Beyond the above "chain rule $\rightarrow$ BN" explicit conditional independence assumptions, often additional implicit assumptions exist. (They can be read off the graph.)
- For modeling, it is important to understand all the assumptions made when the BN graph is chosen.

Example:



- This BN enforces the following simplification of the chain rule:

$$P(X)P(Y|X)P(Z|X,Y)P(W|X,Y,Z) = P(X)P(Y|X)P(Z|Y)P(W|Z)$$

- Explicit assumptions from these simplifications:

$$\begin{aligned} P(Z|X,Y) = P(Z|Y) \quad &\Longrightarrow \quad Z \perp\!\!\!\perp X|Y \\ P(W|X,Y,Z) = P(W|Z) \quad &\Longrightarrow \quad W \perp\!\!\!\perp X,Y|Z \end{aligned}$$

- Additional implicit assumption:

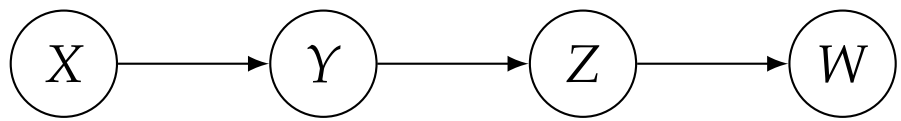
Assumptions in BN

- Each BN defines a factorization of the joint distribution.
- The factorization is possible due to (conditional) independence assumptions we are willing to make:

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

- Beyond the above "chain rule → BN" explicit conditional independence assumptions, often additional implicit assumptions exist. (They can be read off the graph.)
- For modeling, it is important to understand all the assumptions made when the BN graph is chosen.

Example:



- This BN enforces the following simplification of the chain rule:

$$P(X)P(Y|X)P(Z|X,Y)P(W|X,Y,Z) = P(X)P(Y|X)P(Z|Y)P(W|Z)$$

- Explicit assumptions from these simplifications:

$$\begin{aligned} P(Z|X,Y) = P(Z|Y) \quad &\Longrightarrow \quad Z \perp\!\!\!\perp X|Y \\ P(W|X,Y,Z) = P(W|Z) \quad &\Longrightarrow \quad W \perp\!\!\!\perp X,Y|Z \end{aligned}$$

- Additional implicit assumption:

$$W \perp\!\!\!\perp X|Y$$

Independence in BN

Question about a BN:

- **Are certain 2 variables independent given certain evidence?**
- Can we answer this by studying local structures in BN?

Independence in BN

Question about a BN:

- **Are certain 2 variables independent given certain evidence?**
- Can we answer this by studying local structures in BN?

Why is this question important?

- Assume we want to answer query about X and we have evidence on Y.
- If we can analyze the BN structure and find a set of variables Z which are independent of X given Y, we can greatly simplify the inference (because Z has no effect on X)!

Independence in BN

Question about a BN:

- **Are certain 2 variables independent given certain evidence?**
- Can we answer this by studying local structures in BN?

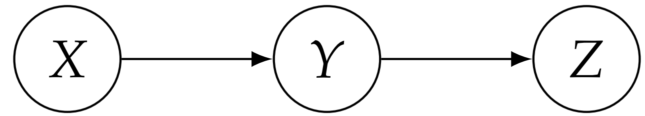
Why is this question important?

- Assume we want to answer query about X and we have evidence on Y.
- If we can analyze the BN structure and find a set of variables Z which are independent of X given Y, we can greatly simplify the inference (because Z has no effect on X)!

D-separation

- A condition/algorithm for answering such queries.
- Study independence properties for triplets of variables.
- Analyze complex cases in terms of the included triplets.
- Triplets can have only 3 possible configurations which cover all cases:
  - "Causal chain" (linear structure)
  - "Common cause" (diverging structure)
  - "Common effect" (converging structure)

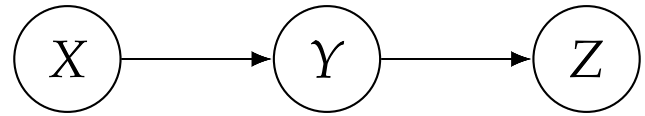
Causal chain



$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Example: low atmospheric pressure (X) causes rain (Y) which causes high traffic (Z).

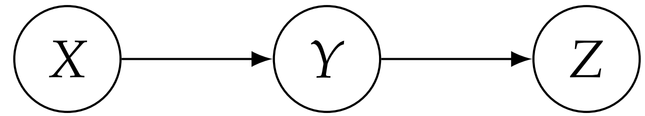
Causal chain



$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Example: low atmospheric pressure (X) causes rain (Y) which causes high traffic (Z).
- **Are X and Z guaranteed to be independent?**

Causal chain



$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Example: low atmospheric pressure (X) causes rain (Y) which causes high traffic (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.
  - But despite that, in some particular cases they can be independent. How?

Causal chain

$$X \rightarrow Y \rightarrow Z$$

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Example: low atmospheric pressure (X) causes rain (Y) which causes high traffic (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.
  - But despite that, in some particular cases they can be independent. How?
- **Are X and Z guaranteed to be independent given Y?**

Causal chain

$$X \rightarrow Y \rightarrow Z$$

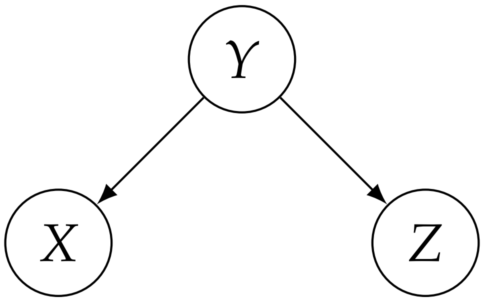
$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Example: low atmospheric pressure (X) causes rain (Y) which causes high traffic (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.
  - But despite that, in some particular cases they can be independent. How?
- **Are X and Z guaranteed to be independent given Y?**
  - **YES!**

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z|y)$$

  - Evidence along the chain **blocks** the mutual influence between the two outer variables.

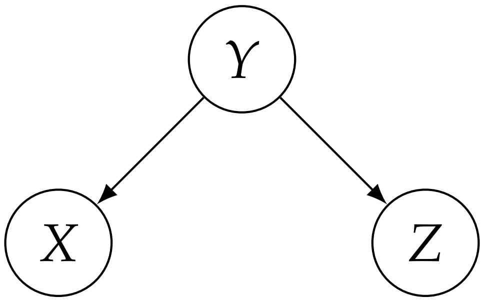
Common cause



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Example: upcoming project deadline (Y) causes both high traffic on student fora (X) and fill computer labs (Z).

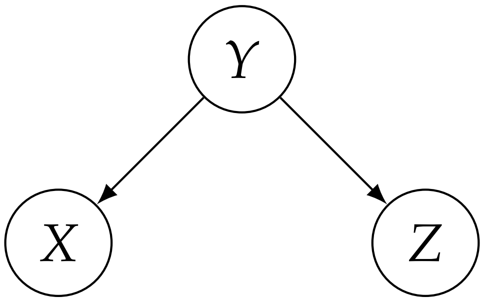
Common cause



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Example: upcoming project deadline (Y) causes both high traffic on student fora (X) and fill computer labs (Z).
- **Are X and Z guaranteed to be independent?**

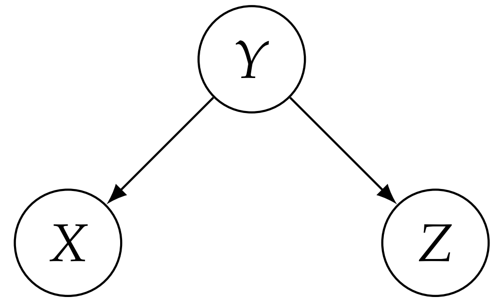
Common cause



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Example: upcoming project deadline (Y) causes both high traffic on student fora (X) and fill computer labs (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.

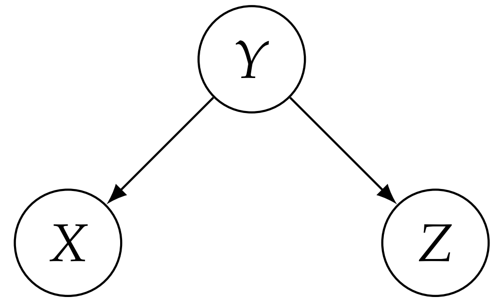
Common cause



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Example: upcoming project deadline (Y) causes both high traffic on student fora (X) and fill computer labs (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.
- **Are X and Z guaranteed to be independent given Y?**

Common cause



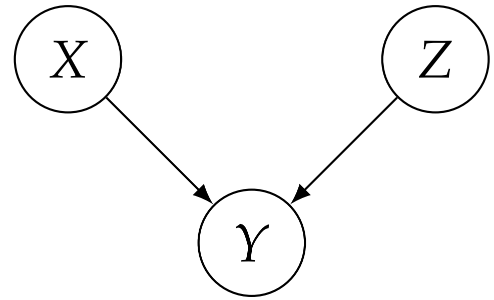
$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Example: upcoming project deadline (Y) causes both high traffic on student fora (X) and fill computer labs (Z).
- **Are X and Z guaranteed to be independent?**
  - **No.**
  - You can easilly find a counterexample, i.e. CPTs for which X and Z are not independent, i.e. they are not guaranteed to be independent.
- **Are X and Z guaranteed to be independent given Y?**
  - **YES!**

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} = P(z|y)$$

  - Evidence on the cause **blocks** the mutual influence between all effects.

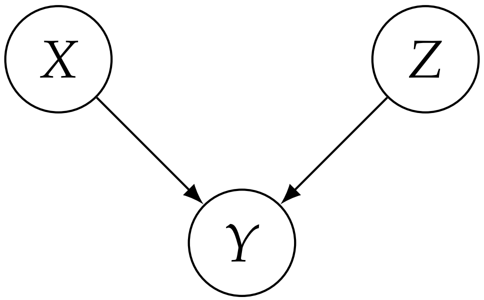
Common effect



$$P(x, y, z) = P(x)P(y)P(z|x, y)$$

- Example: Rain (X) and a football match at nearby stadium (Z) both cause increased traffic (Y).

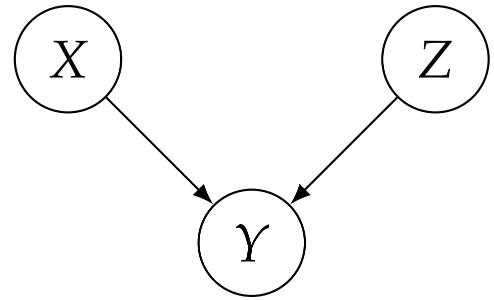
Common effect



$$P(x,y,z) = P(x)P(y)P(z|x,y)$$

- Example: Rain (X) and a football match at nearby stadium (Z) both cause increased traffic (Y).
- **Are X and Z guaranteed to be independent?**

Common effect

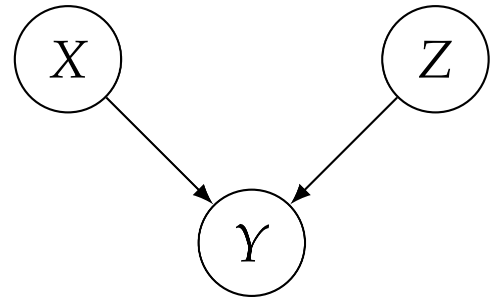


$$P(x,y,z) = P(x)P(y)P(z|x,y)$$

- Example: Rain (X) and a football match at nearby stadium (Z) both cause increased traffic (Y).
- **Are X and Z guaranteed to be independent?**
  - **Yes.**

$$P(x,y) = \sum_z P(x,y,z) = \sum_z P(x)P(y)P(z|x,y) = P(x)P(y)$$

Common effect



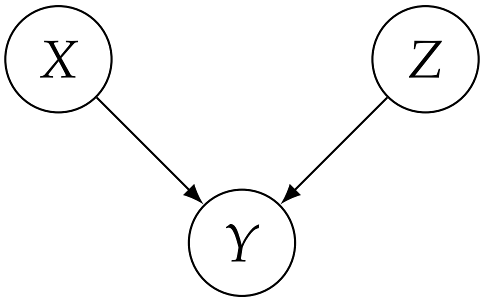
$$P(x, y, z) = P(x)P(y)P(z|x, y)$$

- Example: Rain (X) and a football match at nearby stadium (Z) both cause increased traffic (Y).
- **Are X and Z guaranteed to be independent?**
  - **Yes.**

$$P(x, y) = \sum_z P(x, y, z) = \sum_z P(x)P(y)P(z|x, y) = P(x)P(y)$$

- **Are X and Z guaranteed to be independent given Y?**

Common effect



$$P(x, y, z) = P(x)P(y)P(z|x, y)$$

- Example: Rain (X) and a football match at nearby stadium (Z) both cause increased traffic (Y).
- **Are X and Z guaranteed to be independent?**
  - **Yes.**

$$P(x, y) = \sum_z P(x, y, z) = \sum_z P(x)P(y)P(z|x, y) = P(x)P(y)$$

- **Are X and Z guaranteed to be independent given Y?**
  - **NO!**
  - Seeing traffic (y) puts the rain (X) and the football game (Z) in competition as explanation.
  - The opposite of the previous 2 cases: observing an effect **activates** influence between possible causes.
  - The influence is activated also when we *observe any descendant of Y!*

D-separation

Question:

- Are variables X and Y independent given evidence on $Z_1, \ldots, Z_k$, i.e. can we write $X \perp\!\!\!\perp Y | \{Z_1, \ldots, Z_k\}$?

D-separation

Question:

- Are variables X and Y independent given evidence on $Z_1, \ldots, Z_k$, i.e. can we write $X \perp\!\!\!\perp Y | \{Z_1, \ldots, Z_k\}$?

Answer:

- Check all (undirected!) paths between X and Y.
- If *all paths are inactive/blocked*, we say that X and Y are *d-separated by $Z_1, \ldots, Z_k$*. Then independence is guaranteed, i.e.

$$X \perp\!\!\!\perp Y | \{Z_1, \ldots, Z_k\}$$

- Otherwise, if *at least one path is active*, we say that X and Y are *d-connected*. Independence is not guaranteed.

Introduction

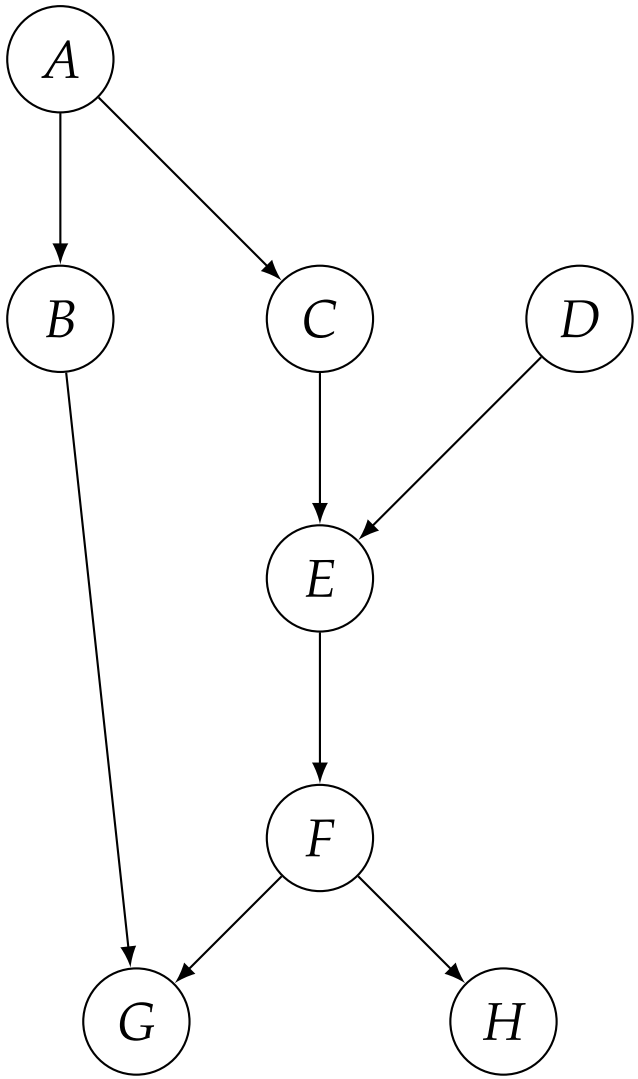
Bayesian networks

- Issues
- BN
- BN example
- Independence
- Independence?
- Conditional independence
- Causality
- Assumptions in BN
- Independence in BN
- Causal chain
- Common cause
- Common effect
- D-separation
- D-sep examples

Inference

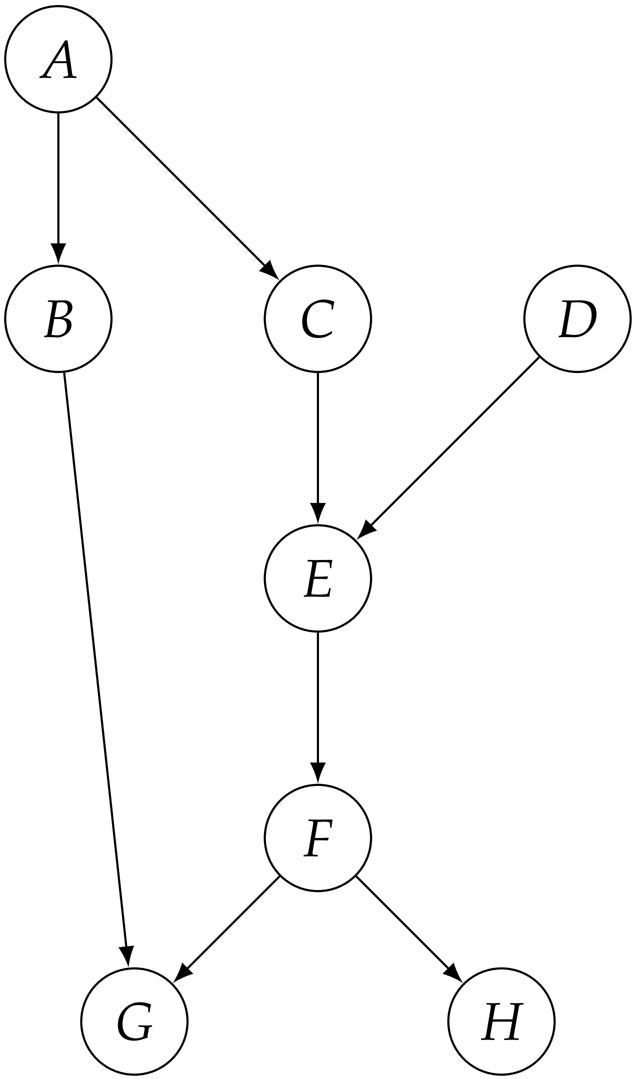
Summary

D-sep examples



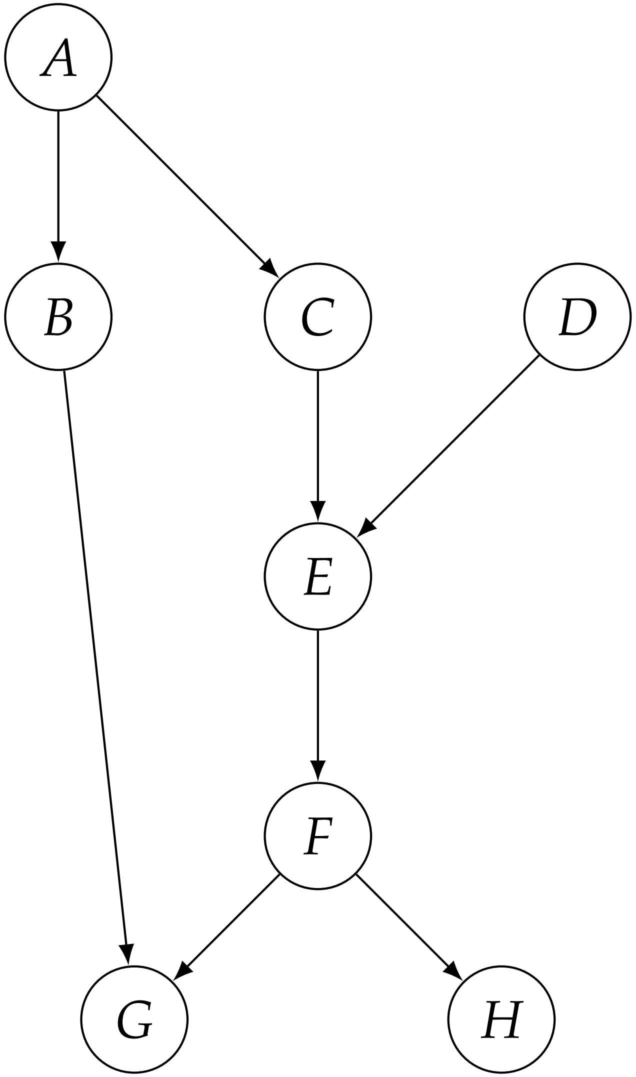
$B \perp\!\!\!\perp C | A$?

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

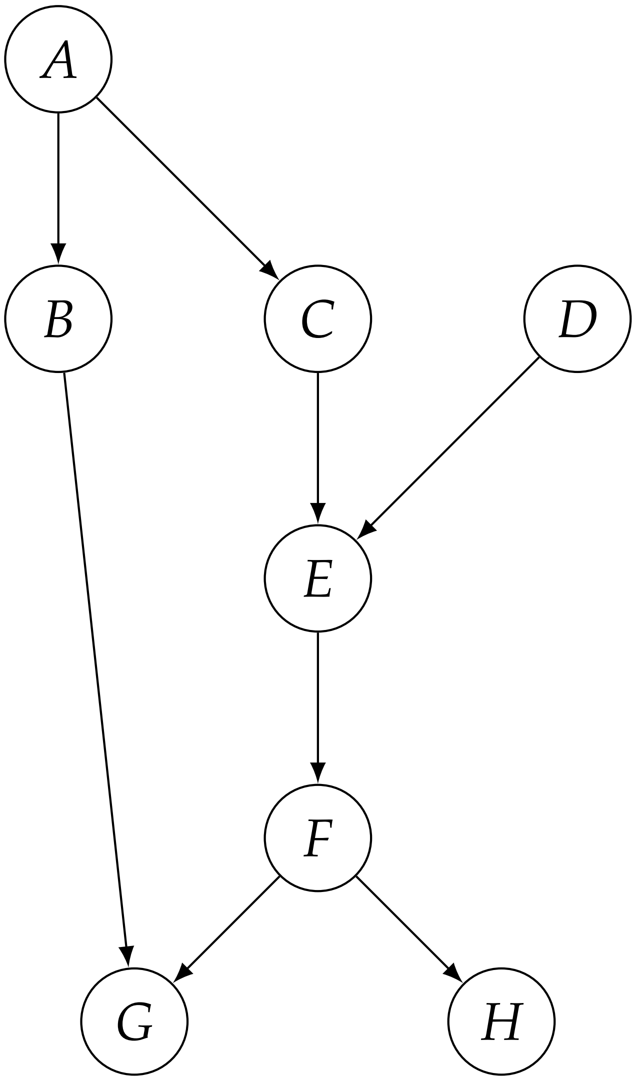
D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

D-sep examples

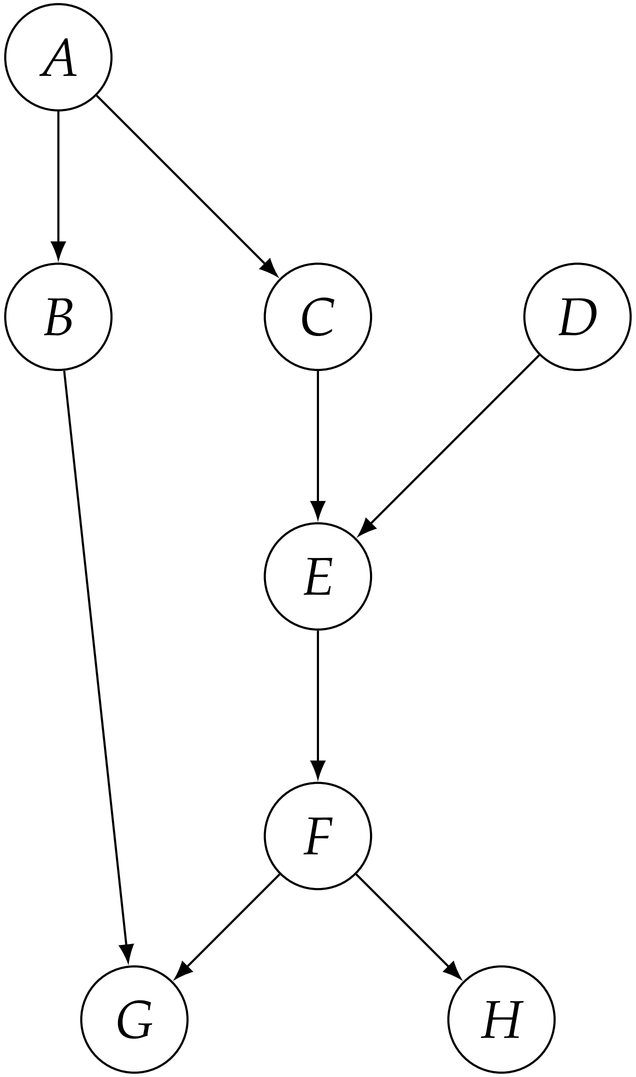


$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$?

D-sep examples

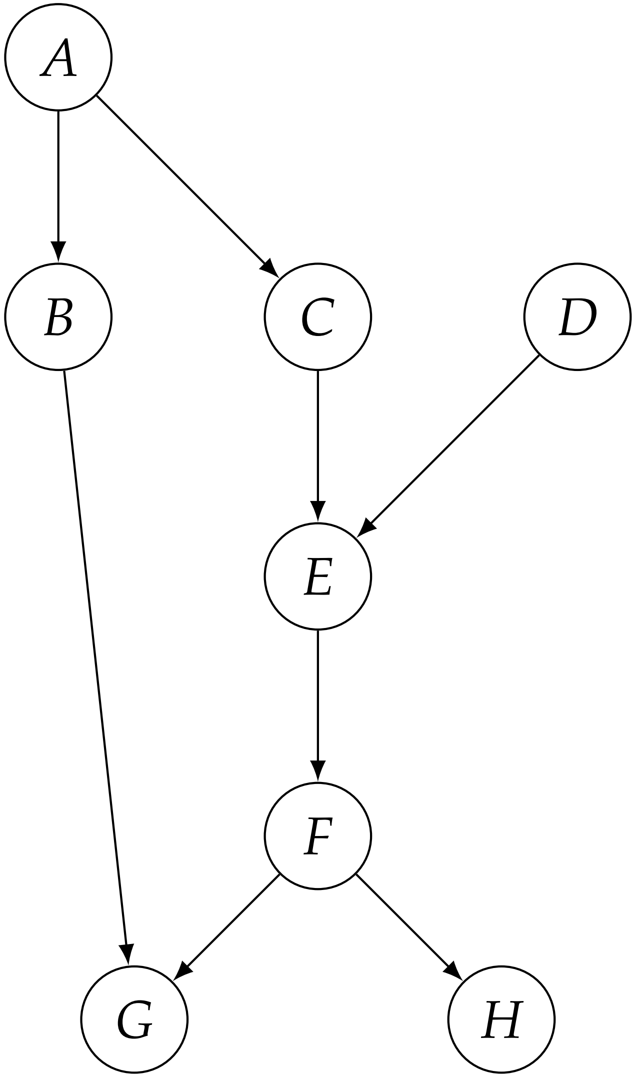


$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

D-sep examples



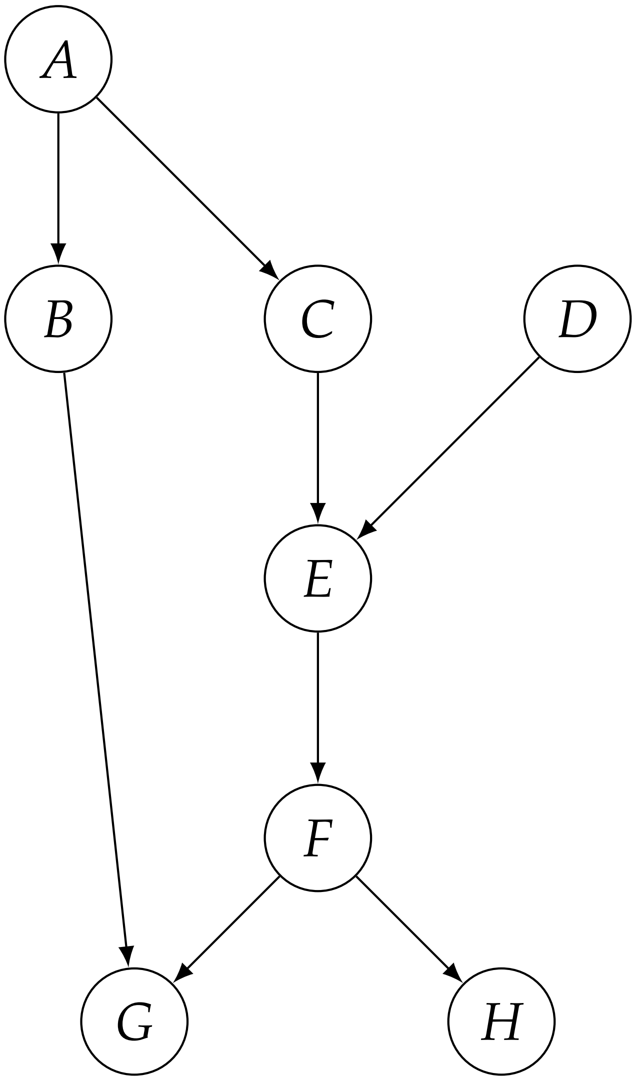
$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

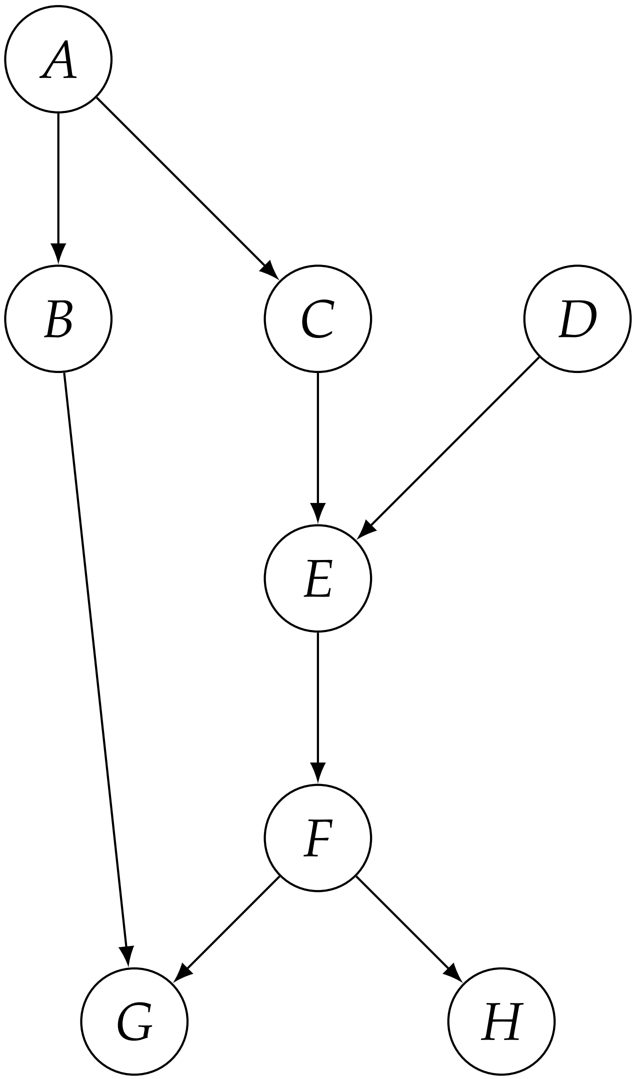
- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

$C \perp\!\!\!\perp D | F$?

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

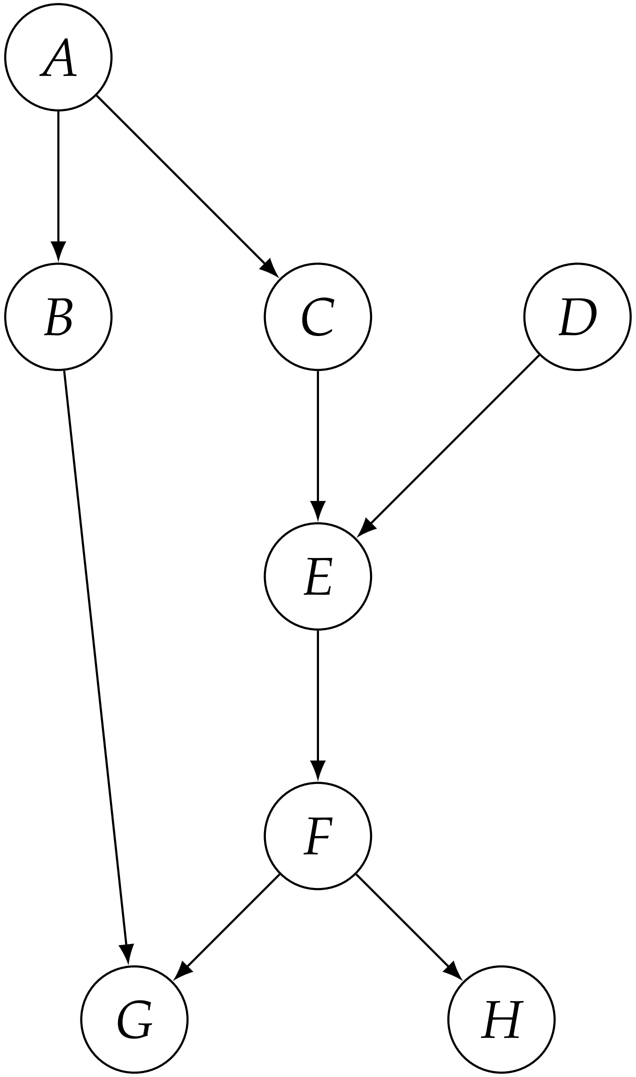
- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

$C \perp\!\!\!\perp D | F$? NO! Why?

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

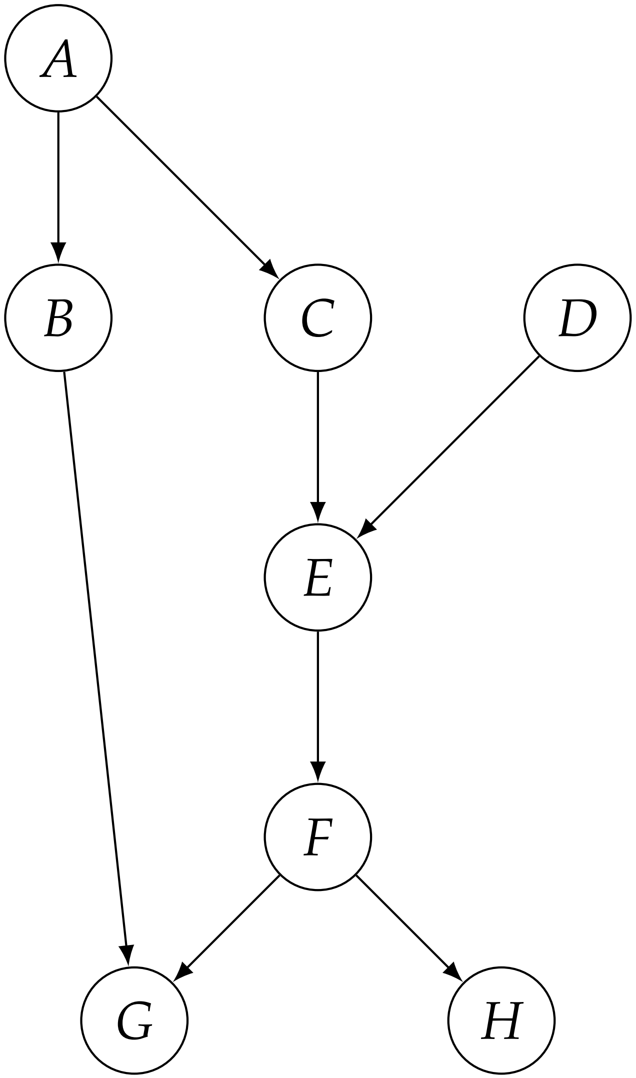
$A \perp\!\!\!\perp F | E$? YES! Why?

- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

$C \perp\!\!\!\perp D | F$? NO! Why?

- C, A, B, G, F, E, D is blocked by evidence on F and by missing evidence on G
- C, E, D is activated by the evidence on F which is a descendant of E.

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

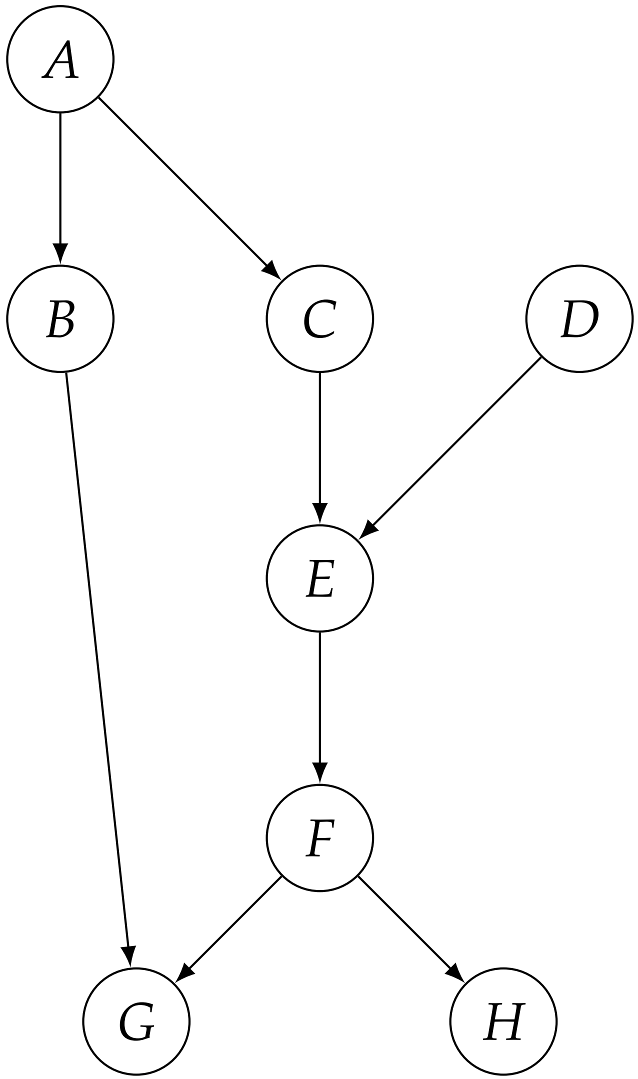
- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

$C \perp\!\!\!\perp D | F$? NO! Why?

- C, A, B, G, F, E, D is blocked by evidence on F and by missing evidence on G
- C, E, D is activated by the evidence on F which is a descendant of E.

$A \perp\!\!\!\perp G | \{B, F\}$? YES! Why?

D-sep examples



$B \perp\!\!\!\perp C | A$? YES! Why?

- B, A, C blocked by evidence on A
- B, G, F, E, C not active — missing evidence on G

$A \perp\!\!\!\perp F | E$? YES! Why?

- A, C, E, F blocked by evidence on E
- A, B, G, F not active — missing evidence on G

$C \perp\!\!\!\perp D | F$? NO! Why?

- C, A, B, G, F, E, D is blocked by evidence on F and by missing evidence on G
- C, E, D is activated by the evidence on F which is a descendant of E.

$A \perp\!\!\!\perp G | \{B, F\}$? YES! Why?

- A, B, G blocked by evidence on B
- A, C, E, F, G blocked by evidence on F

Inference

What is inference?

Inference

- Calculation of some useful quantity from a joint probability distribution.
- Examples:
  - Posterior probability:

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

  - Most likely explanation:

$$\arg\max_q P(Q = q|E_1 = e_1, \dots, E_k = e_k)$$

What is inference?

Inference

- Calculation of some useful quantity from a joint probability distribution.
- Examples:
  - Posterior probability:

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

  - Most likely explanation:

$$\arg\max_q P(Q = q|E_1 = e_1, \dots, E_k = e_k)$$

General case: The set of all variables $X_1, \dots, X_n$ is formally divided into

- evidence variables $E_1, \dots, E_k = e_1, \dots, e_k$,
- query variable(s) Q,
- hidden variables $H_1, \dots, H_r$,
- and assuming we know the joint $P(X_1, \dots, X_n)$ we want to compute (e.g.)

$$P(Q|E_1 = e_1, \dots, E_k = e_k).$$

- How to do it?

Inference by enumeration

Given the joint distribution $P(X_1, \dots, X_n) = P(Q, H_1, \dots, H_r, E_1, \dots, E_k)$:

$$P(Q|e_1, \dots, e_k) = \frac{P(Q, e_1, \dots, e_k)}{P(e_1, \dots, e_k)}$$

$$P(Q, e_1, \dots, e_k) = \sum_{h_1, \dots, h_r} P(Q, h_1, \dots, h_r, e_1, \dots, e_k)$$

$$P(e_1, \dots, e_k) = \sum_{q, h_1, \dots, h_r} P(q, h_1, \dots, h_r, e_1, \dots, e_k)$$

Inference by enumeration

Given the joint distribution $P(X_1, \ldots, X_n) = P(Q, H_1, \ldots, H_r, E_1, \ldots, E_k)$:

$$P(Q|e_1, \ldots, e_k) = \frac{P(Q, e_1, \ldots, e_k)}{P(e_1, \ldots, e_k)}$$

$$P(Q, e_1, \ldots, e_k) = \sum_{h_1, \ldots, h_r} P(Q, h_1, \ldots, h_r, e_1, \ldots, e_k)$$

$$P(e_1, \ldots, e_k) = \sum_{q, h_1, \ldots, h_r} P(q, h_1, \ldots, h_r, e_1, \ldots, e_k)$$

This is computationally equivalent to:

1. From $P(Q, H_1, \ldots, H_r, E_1, \ldots, E_k)$, select all the entries consistent with $e_1, \ldots, e_k$.
2. Sum out all H to get "joint" of Query and evidence:

$$P(Q, e_1, \ldots, e_k) = \sum_{h_1, \ldots, h_r} P(Q, h_1, \ldots, h_r, e_1, \ldots, e_k)$$

3. Normalize the distribution:

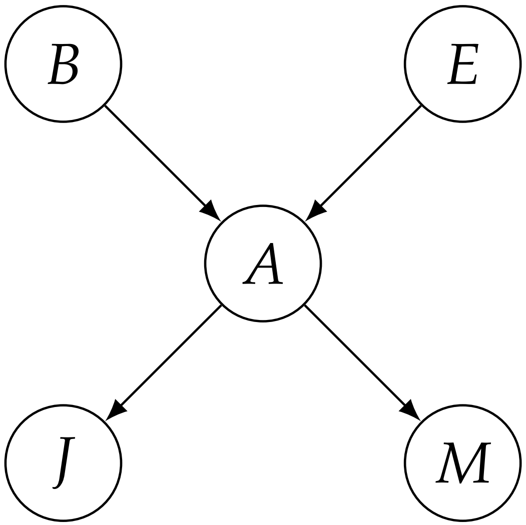
$$P(Q|e_1, \ldots, e_k) = \frac{1}{Z} P(Q, e_1, \ldots, e_k), \qquad \text{where} \qquad Z = \sum_q P(q, e_1, \ldots, e_k).$$

This is often written as $P(Q|e_1, \ldots, e_k). \propto_Q P(Q, e_1, \ldots, e_k)$.

Enumeration in BN

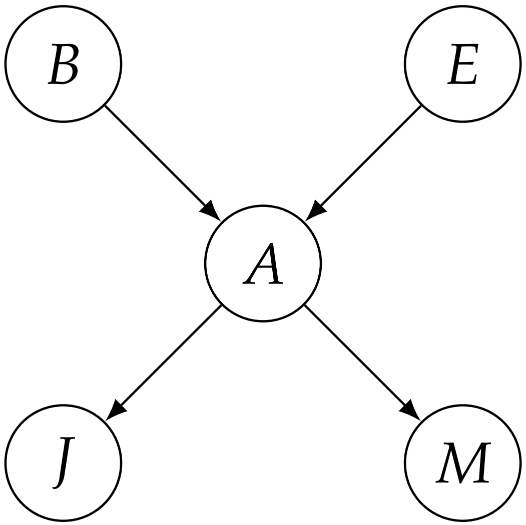
- Given unlimited time, inference in BN is easy.
- Example:

$$P(B|+j,+m) \propto_B P(B,+j,+m) = \sum_{e,a} P(B,e,a,+j,+m) =$$
$$= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a) =$$
$$= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)+$$
$$+ P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)$$



Enumeration in BN

- Given unlimited time, inference in BN is easy.
- Example:

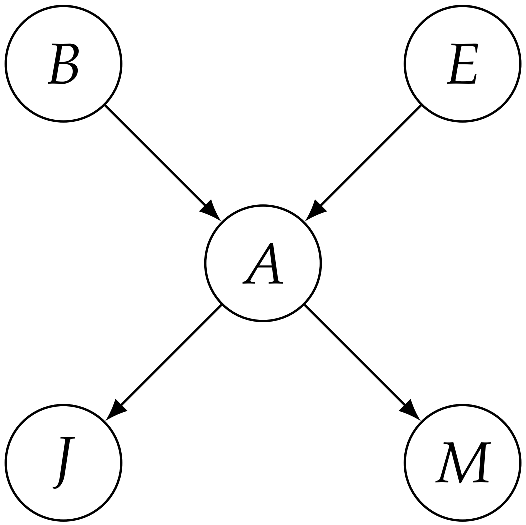


$$
\begin{aligned}
P(B|+j,+m) \propto_B P(B,+j,+m) &= \sum_{e,a} P(B,e,a,+j,+m) = \\
&= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a) = \\
&= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a)+ \\
&+ P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a)+ \\
&+ P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)+ \\
&+ P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)
\end{aligned}
$$

What if the BN would be much larger?

Enumeration in BN

- Given unlimited time, inference in BN is easy.
- Example:



$$P(B|+j,+m) \propto_B P(B,+j,+m) = \sum_{e,a} P(B,e,a,+j,+m) =$$

$$= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a) =$$

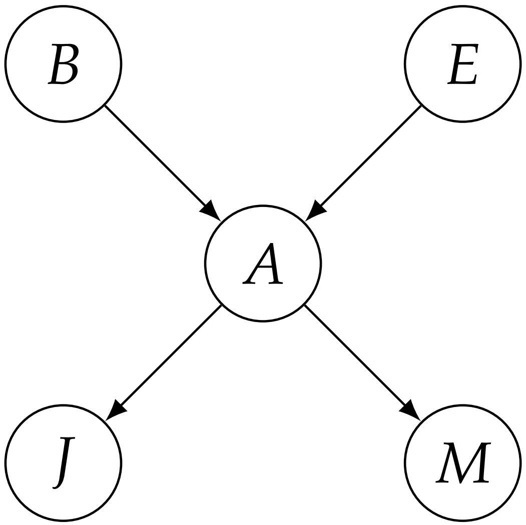
$$= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)+$$
$$+ P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)$$

What if the BN would be much larger? Inference by enumeration would be

- very slow, because
- it first creates the whole joint distribution before it can sum out the hidden variables! Inference by *enumeration has exponential complexity*!

Enumeration in BN

- Given unlimited time, inference in BN is easy.
- Example:



$$P(B|+j,+m) \propto_B P(B,+j,+m) = \sum_{e,a} P(B,e,a,+j,+m) =$$
$$= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a) =$$
$$= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a)+$$
$$+ P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)+$$
$$+ P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)$$

What if the BN would be much larger? Inference by enumeration would be

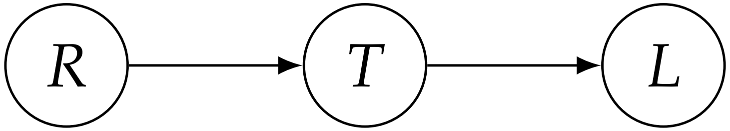
- very slow, because
- it first creates the whole joint distribution before it can sum out the hidden variables! Inference by *enumeration has exponential complexity*!

What about joining only such part of the distribution that would allow us to sum out a hidden variable as soon as possible?

- **Variable elimination**: Interleave joining and marginalization!
- Still *worst-case exponential complexity*, but in practice much faster than inference by enumeration!

Enumeration vs Variable elimination (VE)

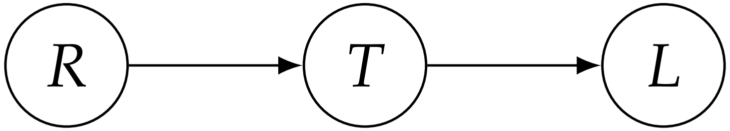
- R: Rain
- T: Traffic
- L: Late for school
- $P(L) = ?$



$$P(R, T, L) = P(R)P(T|R)P(L|T)$$

Enumeration vs Variable elimination (VE)

- R: Rain
- T: Traffic
- L: Late for school
- $P(L) = ?$



$$P(R, T, L) = P(R)P(T|R)P(L|T)$$

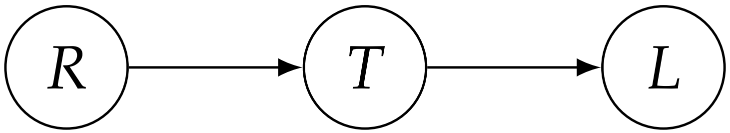
Inference by enumeration:

- Build the full joint first.
- Then sum out hidden variables.

$$P(L) = \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\ \underbrace{P(r)P(t|r)}_{\text{Join on } r\text{: } P(r,t)}}_{\text{Join on } t\text{: } P(r,t,L)}}_{\text{Eliminate } r\text{: } P(t,L)}}_{\text{Eliminate } t\text{: } P(L)}$$

Enumeration vs Variable elimination (VE)

- R: Rain
- T: Traffic
- L: Late for school
- $P(L) = ?$



$$P(R, T, L) = P(R)P(T|R)P(L|T)$$

Inference by enumeration:

- Build the full joint first.
- Then sum out hidden variables.

$$P(L) = \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\ \underbrace{P(r)P(t|r)}_{\text{Join on } r:\ P(r,t)}}_{\text{Join on } t:\ P(r,t,L)}}_{\text{Eliminate } r:\ P(t,L)}}_{\text{Eliminate } t:\ P(L)}$$

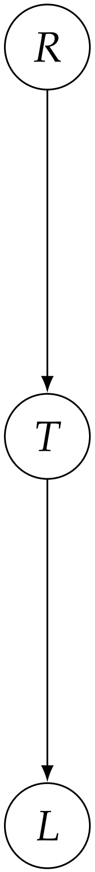
Inference by variable elimination:

- Perform a “small” join.
- Marginalize as soon as you can.

$$P(L) = \underbrace{\sum_t \underbrace{P(L|t) \underbrace{\sum_r\ \underbrace{P(r)P(t|r)}_{\text{Join on } r:\ ,P(r,t)}}_{\text{Eliminate } r:\ P(t)}}_{\text{Join on } t:\ P(t,L)}}_{\text{Eliminate } t:\ P(L)}$$

VE example

Initial factors:



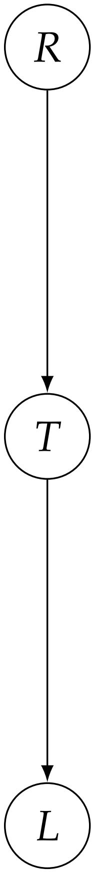
$P(R)$	
$+r$	0.1
$-r$	0.9

$P(T\|R)$		
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

VE example

Initial factors:



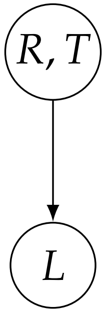
$P(R)$	
$+r$	0.1
$-r$	0.9

$P(T\|R)$		
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After join on R:

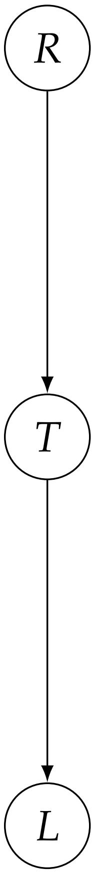
$P(R,T)$		
$+r$	$+t$	0.08
$+r$	$-t$	0.02
$-r$	$+t$	0.09
$-r$	$-t$	0.81



$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

VE example

Initial factors:



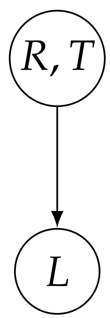
$P(R)$	
$+r$	0.1
$-r$	0.9

$P(T\|R)$		
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After join on R:

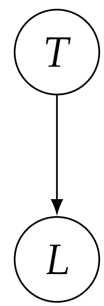
$P(R,T)$		
$+r$	$+t$	0.08
$+r$	$-t$	0.02
$-r$	$+t$	0.09
$-r$	$-t$	0.81



$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After eliminating R:

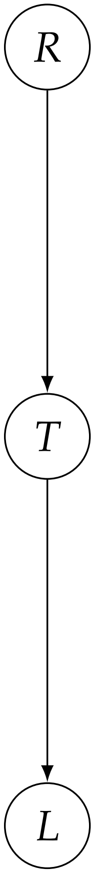
$P(T)$	
$+t$	0.17
$-t$	0.83



$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

VE example

Initial factors:



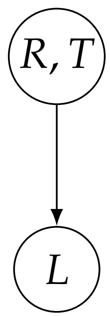
$P(R)$	
$+r$	0.1
$-r$	0.9

$P(T\|R)$		
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After join on R:

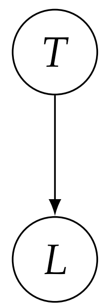
$P(R,T)$		
$+r$	$+t$	0.08
$+r$	$-t$	0.02
$-r$	$+t$	0.09
$-r$	$-t$	0.81



$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

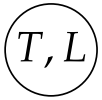
After eliminating R:

$P(T)$	
$+t$	0.17
$-t$	0.83



$P(L\|T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

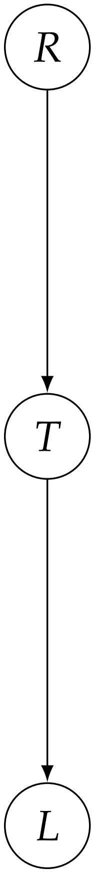
After join on T:



$P(T,L)$		
$+t$	$+l$	0.051
$+t$	$-l$	0.119
$-t$	$+l$	0.083
$-t$	$-l$	0.747

VE example

Initial factors:



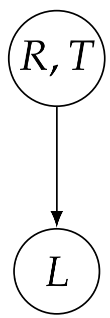
$P(R)$	
$+r$	0.1
$-r$	0.9

$P(T \mid R)$		
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

$P(L \mid T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After join on R:

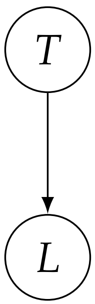
$P(R,T)$		
$+r$	$+t$	0.08
$+r$	$-t$	0.02
$-r$	$+t$	0.09
$-r$	$-t$	0.81



$P(L \mid T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

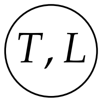
After eliminating R:

$P(T)$	
$+t$	0.17
$-t$	0.83



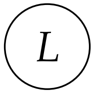
$P(L \mid T)$		
$+t$	$+l$	0.3
$+t$	$-l$	0.7
$-t$	$+l$	0.1
$-t$	$-l$	0.9

After join on T:



$P(T,L)$		
$+t$	$+l$	0.051
$+t$	$-l$	0.119
$-t$	$+l$	0.083
$-t$	$-l$	0.747

After eliminating T:



$P(L)$	
$+l$	0.134
$-l$	0.866

Evidence in VE

If there is some Evidence in VE, e.g. if $P(L|+r)$ is required:

- Use only factors which correspond to the evidence, i.e. for the above example,
  - instead of $P(R)$, use $P(+r)$,
  - instead of $P(T|R)$, use $P(T|+r)$,
  - use $P(L|T)$ as before (evidence does not affect it).
- Eliminate all variables except query Q and evidence e.
- Result of VE will be a (partial) joint distribution of Q and e, i.e. for the above example, we would get

$$P(+r, L).$$

- To get $P(L|+r)$, just normalize $P(+r, L)$, i.e.

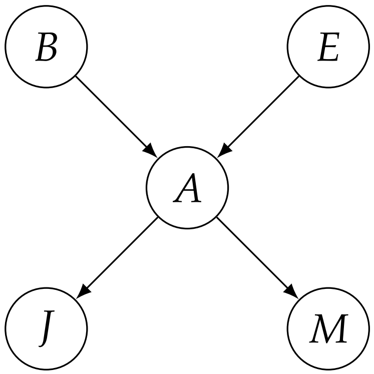
$$P(L|+r) \propto_L P(+r, L).$$

General variable elimination

Query: $P(Q|E_1 = e_1, \ldots, E_k = e_k)$

1. Start with the initial CPTs, instantiated with the evidence $e_1, \ldots, e_k$.
2. While there are any hidden variables:
   - Choose a hidden variable H.
   - Join all factors containing H.
   - Eliminate (sum out) H.
3. Join all remaining factors and normalize.

VE Example 2



Query: $P(B|+j,+m) = ?$

1. Start with the given CPTs corresponding to evidence $+j,+m$:

$$\boxed{P(B) \quad P(E) \quad P(A|B,E) \quad P(+j|A) \quad P(+m|A)}$$

2. Choose hidden variable A and all factors containing it:

$$\left.\begin{array}{l} P(+j|A) \\ P(+m|A) \\ P(A|B,E) \end{array}\right\} \overset{\text{Join on } A}{\Longrightarrow} P(+j,+m,A|B,E) \overset{\text{Sum out } A}{\Longrightarrow} P(+j,+m|B,E)$$

$$\boxed{P(B) \quad P(E) \quad P(+j,+m|B,E)}$$

3. Choose hidden variable E and all factors containing it:

$$\left.\begin{array}{l} P(E) \\ P(+j,+m|B,E) \end{array}\right\} \overset{\text{Join on } E}{\Longrightarrow} P(+j,+m,E|B) \overset{\text{Sum out } E}{\Longrightarrow} P(+j,+m|B)$$

$$\boxed{P(B) \quad P(+j,+m|B)}$$

4. No hidden variables left. Finish with B

$$\left.\begin{array}{l} P(B) \\ P(+j,+m|B) \end{array}\right\} \overset{\text{Join on } B}{\Longrightarrow} P(+j,+m,B) \overset{\text{Normalize}}{\Longrightarrow} P(B|+j,+m),$$

which is the result we were looking for.

VE Comments

- Conceptually, VE just replaces the computation of

$$uwy + uwz + uxy + uxz + vwy + vwz + vxy + vxz$$

  with the equivalent computation of

$$(u + v)(w + x)(y + z)$$

  to improve computational efficiency!
- The computational and space complexity of VE is determined by the largest factor (probability table) generated during the process.
- The elimination ordering can greatly affect the size of the largest factor.
- Does there always exist an ordering that only results in small factors? **NO!**
- Inference in BN can be reduced to SAT problem, i.e. **inference in BN is NP-hard**. No known efficient exact probabilistic inference in general.
- For **polytrees**, we can always find an efficient ordering!
  - Polytree is a directed graph with no undirected cycles.

Sampling

Due to the exponential (worst-case) complexity of enumeration and variable elimination, *exact inference may be intractable for large BNs.* $\Longrightarrow$ *Approximate inference using sampling.*

Sampling

Due to the exponential (worst-case) complexity of enumeration and variable elimination, *exact inference may be intractable for large BNs.* $\Longrightarrow$ *Approximate inference using sampling.*

Sampling

- Draw N samples from a sampling distribution S.
- Compute an approximate posterior probability.
- Show that this converges to the true probability P with increasing N.

Sampling

Due to the exponential (worst-case) complexity of enumeration and variable elimination, *exact inference may be intractable for large BNs.* $\Longrightarrow$ *Approximate inference using sampling.*

Sampling

- Draw N samples from a sampling distribution S.
- Compute an approximate posterior probability.
- Show that this converges to the true probability P with increasing N.

Why sampling?

- Learning: get samples from a distribution you do not know.
- Inference: getting a sample is faster than computing the right answer (e.g. with VE).

Sampling

Due to the exponential (worst-case) complexity of enumeration and variable elimination, *exact inference may be intractable for large BNs.* $\Longrightarrow$ *Approximate inference using sampling.*

Sampling

- Draw N samples from a sampling distribution S.
- Compute an approximate posterior probability.
- Show that this converges to the true probability P with increasing N.

Why sampling?

- Learning: get samples from a distribution you do not know.
- Inference: getting a sample is faster than computing the right answer (e.g. with VE).

Sampling in BNs:

- **Prior sampling**: generates samples from joint $P(X_1, \ldots, X_n)$.
- **Rejection sampling**: generates samples from conditional $P(Q|e)$.
- **Likelihood weighting**: generates samples from conditional $P(Q|e)$. Better than rejection sampling if evidence is unlikely.
- **Gibbs sampling**: generates samples from conditional $P(Q|e)$.

Gibbs sampling

Procedure:

1. Start with an arbitrary instantiation (realization) $x_1, \ldots, x_n$ of all variables consistent with the evidence.
2. Choose one of the non-evidence variables (sequentially, or systematically uniformly), say x_i, and resample its value from $P(X_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, i.e. keeping all the other variables and the evidence fixed.
3. Repeat step 2 for a long time.

Gibbs sampling

Procedure:

1. Start with an arbitrary instantiation (realization) $x_1, \ldots, x_n$ of all variables consistent with the evidence.
2. Choose one of the non-evidence variables (sequentially, or systematically uniformly), say x_i, and resample its value from $P(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, i.e. keeping all the other variables and the evidence fixed.
3. Repeat step 2 for a long time.

Properties:

- The sample resulting from the above procedure converges to the right distribution.
- Why is this better than sampling from the joint distribution?
  - In BN, sampling a variable given all the other variables is usually much easier than sampling from the full joint distribution.
  - Only a join on the variable to be sampled is needed: this factor depends only on the variable's parents, its children and its children's parents (*Markov blanket*).
- Gibbs sampling is a special case of Metropolis-Hastings algorithm which belongs to more general methods called **Markov chain Monte Carlo (MCMC) methods**.
  - Methods for sampling from a distribution.
  - The samples are not independent; instead, the neighbors in their stream are very similar to each other.
  - Yet, their distribution converges to the right one, and e.g. sample averages are still consistent estimators.

Summary

Competencies

After this lecture, a student shall be able to . . .

- explain why the joint probability distribution is an awkward model of domains with many random variables;
- define what a Bayesian network is, and describe how it solves the issues with joint probability;
- explain how BN factorize the joint distribution, and compare it with the factorization we get from chain rule;
- write down factorization of the joint probability given the BN graph, and vice versa, draw the BN graph given a factorization of the joint probability;
- explain the relation between the direction of edges in BN and the causality;
- given the structure of a BN, check whether 2 variables are guaranteed to be independent using the concept of D-separation;
- describe and prove the conditional (in)dependence relations among variable triplets (causual chain, common cause, common effect);
- describe inference by enumeration and explain why it is unwieldy for BN;
- explain the difference between inference by enumeration and by variable elimination (VE);
- explain what makes VE more suitable for BN than enumeration;
- describe the features (complexity) of exact inference by enumeration and VE in BN;
- explain how we can use sampling to make approximate inference in BN;
- describe Gibbs sampling.