Artificial Intelligence

An Advanced Course About Foundations

UNIVERSITÀ DI PAVIA

Probabilistic Reasoning: Graphical Models

Marco Piastra

Factorizations & Graphs

Chain Factorization

- **Univariate factorization of a Joint Probability Distribution**

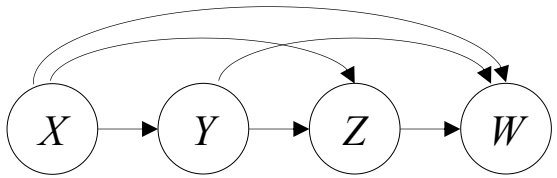
From the definition of conditional probability (*by applying the chain rule*)

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

Any joint probability distribution can be factorized in a way such that
each factor is *univariate* (i.e., one random variable as independent) conditional distribution.

- Each factorization depends on an arbitrary *sequence* of the *random variables*
- Hence factorizations are not *unique*: any sequence of random variables produces a legitimate factorization of the same kind

Graphical equivalent of a *univariate factorization*



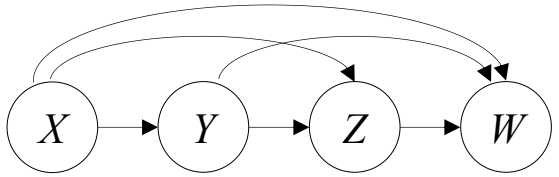
An oriented graph:

- each node represents a random variable (and the corresponding *univariate* factor)
- each arc represents a conditioning of a random variable over another one (i.e. *dependence*)

Chain Factorization

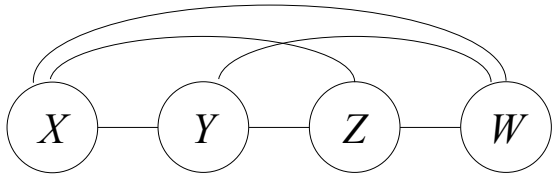
- **Graphical model**

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



This graph:

- is *acyclic*: if you follow the arrows, you will never return to the same node
- is *completely connected*: if you ignore arc orientations, every node is connected to any other node



In graph theory
a fully-connected, undirected graph
is called a CLIQUE

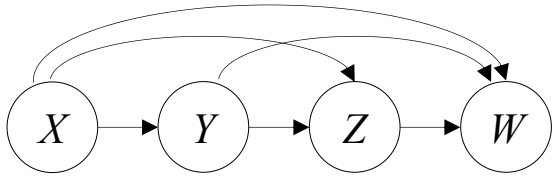
Any *univariate factorization* can be represented by a *graphical model*

Every *completely connected*, *acyclic* and *oriented graph* represents a legitimate *univariate factorization*

Chain Factorization and Independence Assumptions

- **Graphical model**

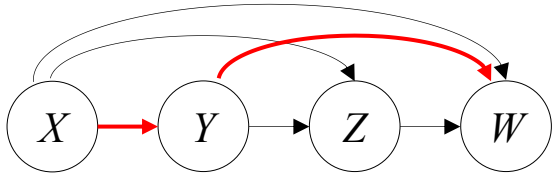
$$P(X,Y,Z,W) = P(X)P(Y|X)P(Z|X,Y)P(W|Y,X,Z)$$



This univariate factorization is always possible
The corresponding graph, ignoring arrows, is a CLIQUE

- **Independence**

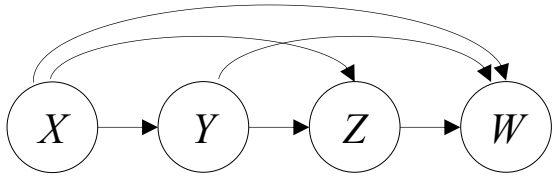
Let's remove a few arcs from the graph and rewrite the factorization accordingly



Chain Factorization and Independence Assumptions

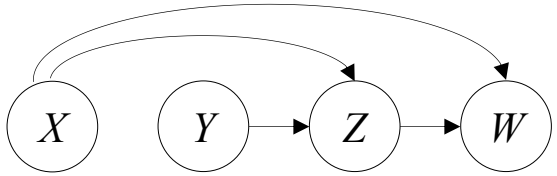
- **Graphical model**

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



- **Independence**

Let's remove a few arcs from the graph and rewrite the factorization accordingly



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

which is true only if

$$P(Y|X) = P(Y) \qquad \langle X \perp Y \rangle$$

Independence

$$P(W|X,Y,Z) = P(W|X,Z) \qquad \langle Y \perp W \,|X,Z \rangle$$

Conditional Independence

Determining Independence from Graphs

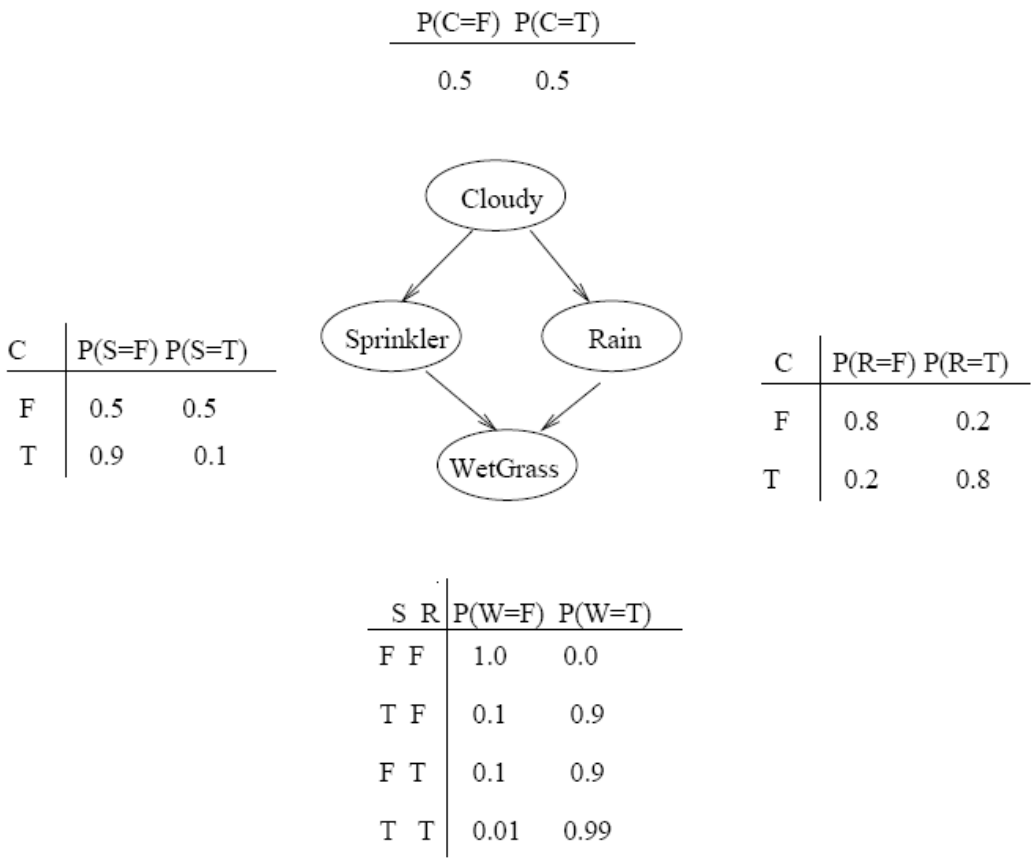
Graphical models (a.k.a. *Bayesian Networks*)

Structure and numbers, instead of just numbers

- A structured, pre-numerical representation of a joint probability

  Each graphical model is an *oriented* graph

  - nodes are *random variables*
  - arcs represent *dependence*



From graphical models to joint probability

- Joint probability factorization

A *chain factorization* like the following is always allowed

$$P(C, S, R, W) = P(C)P(S|C)P(R|C, S)P(W|C, S, R)$$

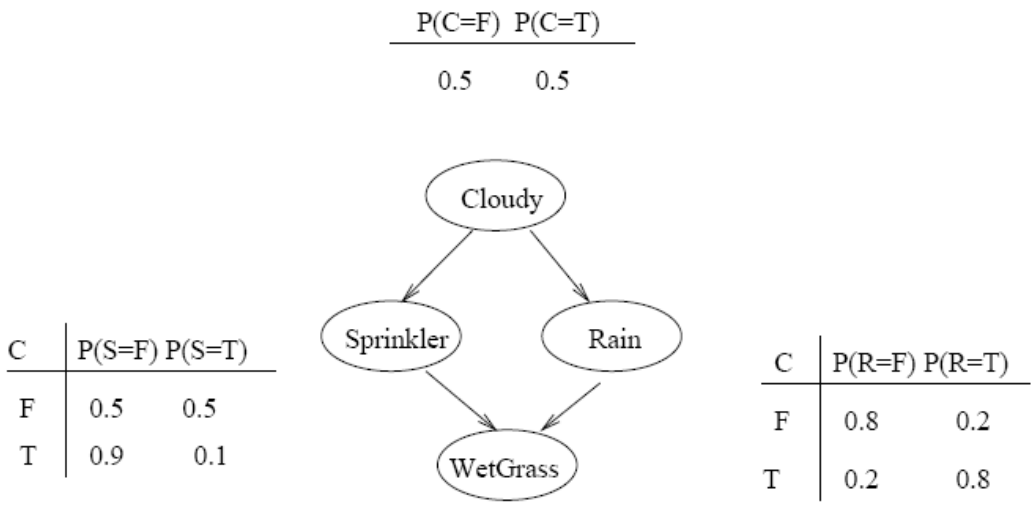
Hint: apply the definition of conditional probability repeatedly (such factorization is not unique)

A complete specification of a joint probability would require $2^4 = 16$ *values*
The values in figure are just 9

Factorization for a graphical model

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

where $parents(X_i)$ are the nodes from which there is an entry arc to X_i



For this example, the above rule produces:

$$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$$

Note the difference from above

Independence assumptions: $<R \perp S \mid C>, <W \perp C \mid R, S>$

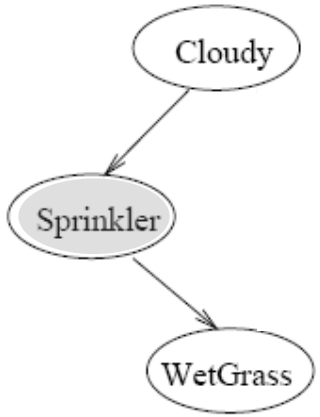
Patterns in Graphical Models

- ***Sequence* or *Chain***

Consider the graph on the right

$$P(C,S,W) = P(C)P(S|C)P(W|S)$$

sequence



Now suppose we observe S

$$\begin{aligned} P(C,W|S) &= \frac{P(C,S,W)}{P(S)} \\ &= \frac{P(C)P(S|C)P(W|S)}{P(S)} \\ &= \frac{P(C,S)}{P(S)}P(W|S) \\ &= P(C|S)P(W|S) \end{aligned}$$

Chain rule

This implies $\langle C \perp W \,|S\rangle$

Conditional independence

Patterns in Graphical Models

- ***Fork***

Consider the graph on the right

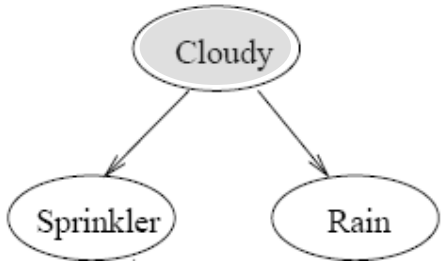
$$P(C, S, R) = P(C)P(S|C)P(R|C)$$

Now suppose we observe C

$$\begin{aligned} P(R, S|C) &= \frac{P(C, S, R)}{P(C)} \\ &= \frac{P(C)P(S|C)P(R|C)}{P(C)} \\ &= \frac{P(C, S)}{P(C)} P(R|C) \\ &= P(S|C)P(R|C) \end{aligned}$$

This implies $\langle R \perp S \,|C\rangle$ — *Conditional independence*

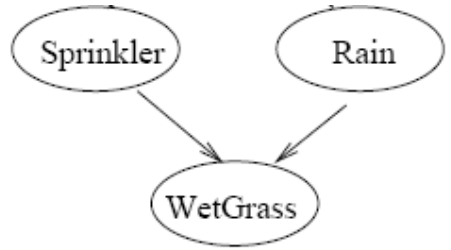
fork



Patterns in Graphical Models

- ***Join* or *Collider*** *CAUTION: this case is different from the previous two*

join or *collider*



Consider the graph on the right

$$P(R, S, W) = P(S)P(R)P(W|S, R)$$

Marginalizing with respect to W

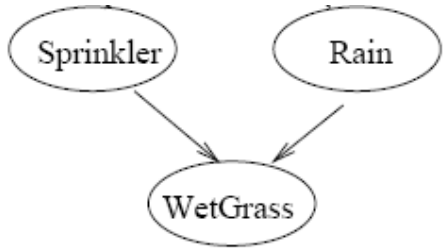
$$\begin{aligned} P(R, S) &= \sum_W P(S)P(R)P(W|S, R) \\ &= P(S)P(R)\sum_W P(W|S, R) \\ &= P(S)P(R) \end{aligned}$$

This implies $\langle R \perp S \rangle$ — *Marginal independence*

Patterns in Graphical Models

- ***Join* or *Collider*** *CAUTION: this case is different from the previous two*

join or *collider*



Consider the graph on the right

$$P(R,S,W) = P(S)P(R)P(W|S,R)$$

which is true only if $\langle S \perp R \rangle$ — *Independence (also 'Marginal Independence')*

Now suppose we observe W

$$P(R,S|W) = \frac{P(R,S,W)}{P(W)}$$

$$= \frac{P(S)P(R)P(W|S,R)}{P(W)}$$

$$\neq P(S|W)P(R|W)$$

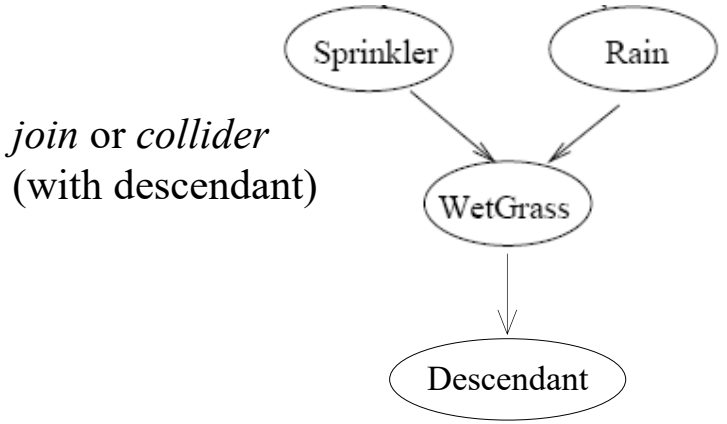
No further simplification possible

This implies $\langle S \not\perp R \,|W\rangle$

Patterns in Graphical Models

- ***Join* or *Collider***

The same loss of independence occurs if you observe any of the descendants...



join or *collider* (with descendant)

Consider the graph on the right

$$P(R, S, W, D) = P(S)P(R)P(W|S, R)P(D|W)$$

which is true only if $\langle S \perp R \rangle$ — *Independence (also 'Marginal Independence')*

Now suppose you observe D

$$P(R, S, W|D) = \frac{P(R, S, W, D)}{P(D)}$$

$$= \frac{P(S)P(R)P(W|S, R)P(D|W)}{P(D)}$$

No further simplification possible

$$\neq P(S|D)P(R|D)$$

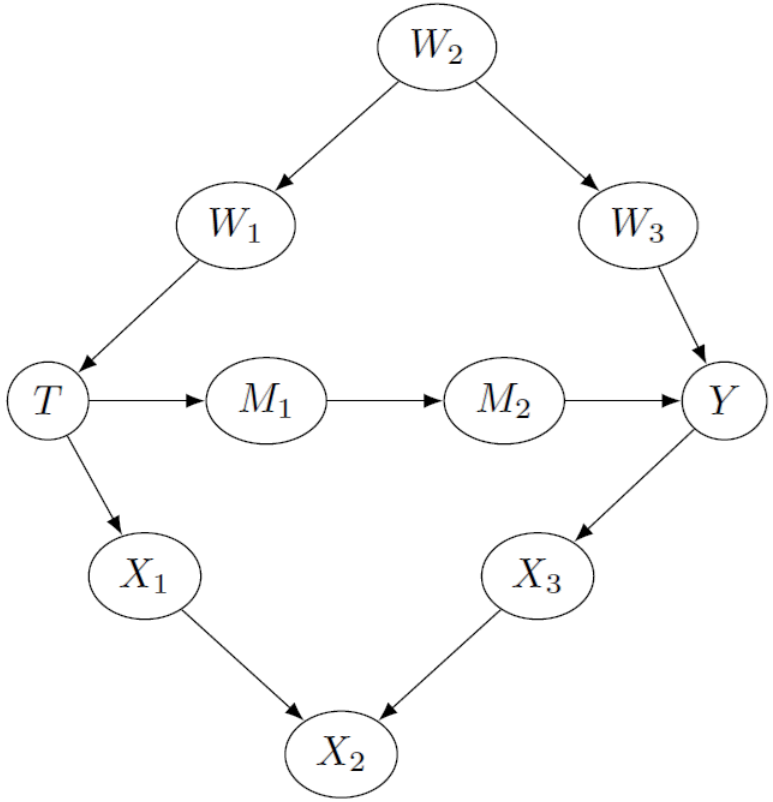
This implies $\langle S \not\perp R \,|D\rangle$

... at any subsequent level of descendance (try yourself)

d-Separation

Reading Graphical Models

Any graphical model represents a univariate factorization



- **Factorization for a graphical model**

  The general formula is:

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

  where *parents* are all nodes having a direct, incoming dependence arc

For the example in figure:

$$P(W_1, W_2, W_3, T, M_1, M_2, Y, X_1, X_2, X_3) =$$
$$P(W_2)P(W_1|W_2)P(W_3|W_2)P(T|W1)P(M_1|T)P(M_2|M_1)P(Y|W_3, M_2)P(X_1|T)P(X_3|Y)P(X_2|X_1, X_3)$$

Paths in Graphical Models

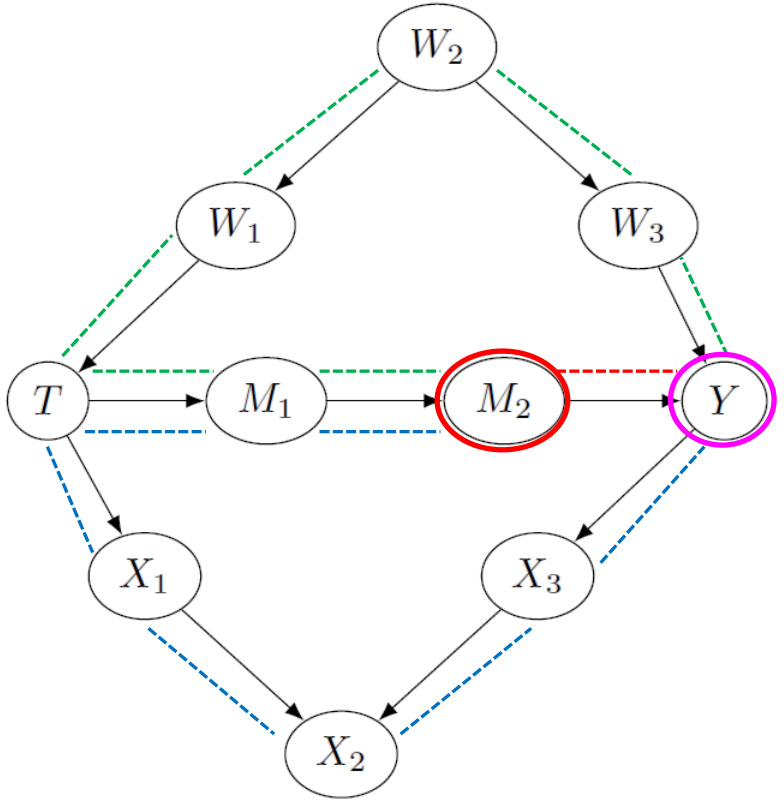
In a graphical model

Consider any two nodes X and Y

A *path* between X and Y
is a path in the graph ignoring orientations (i.e., arrows)

Example:

In the graph on the right,
consider all paths between M_2 and Y



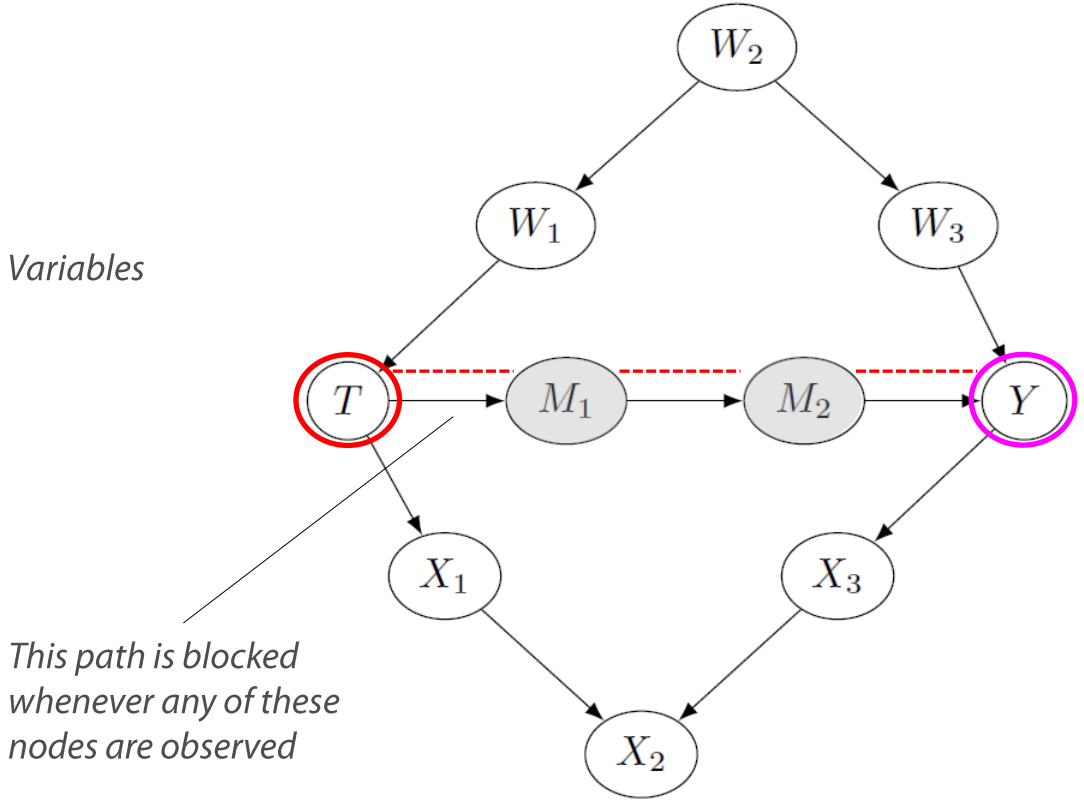
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



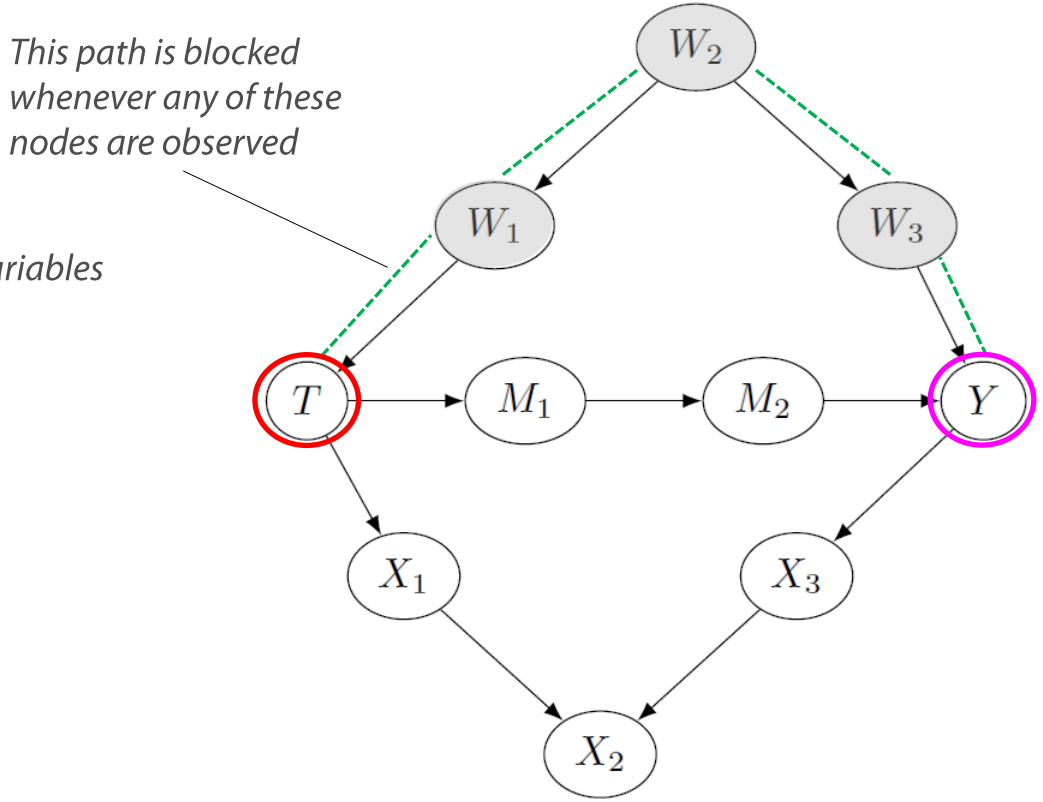
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



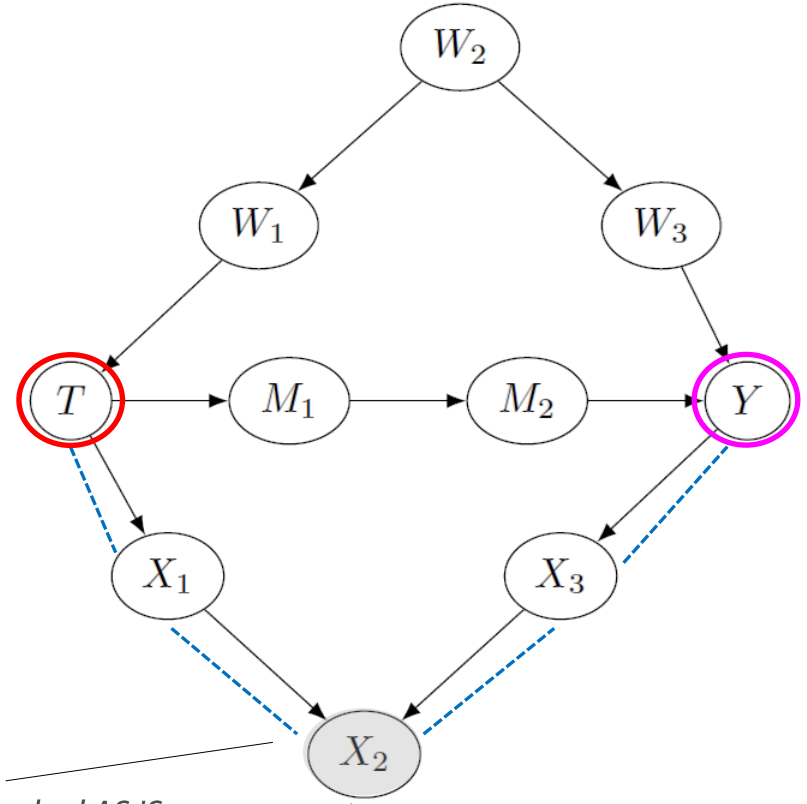
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



This path is <u>blocked</u> AS IS: the collider blocks it

It becomes <u>unblocked</u> when this node is observed...

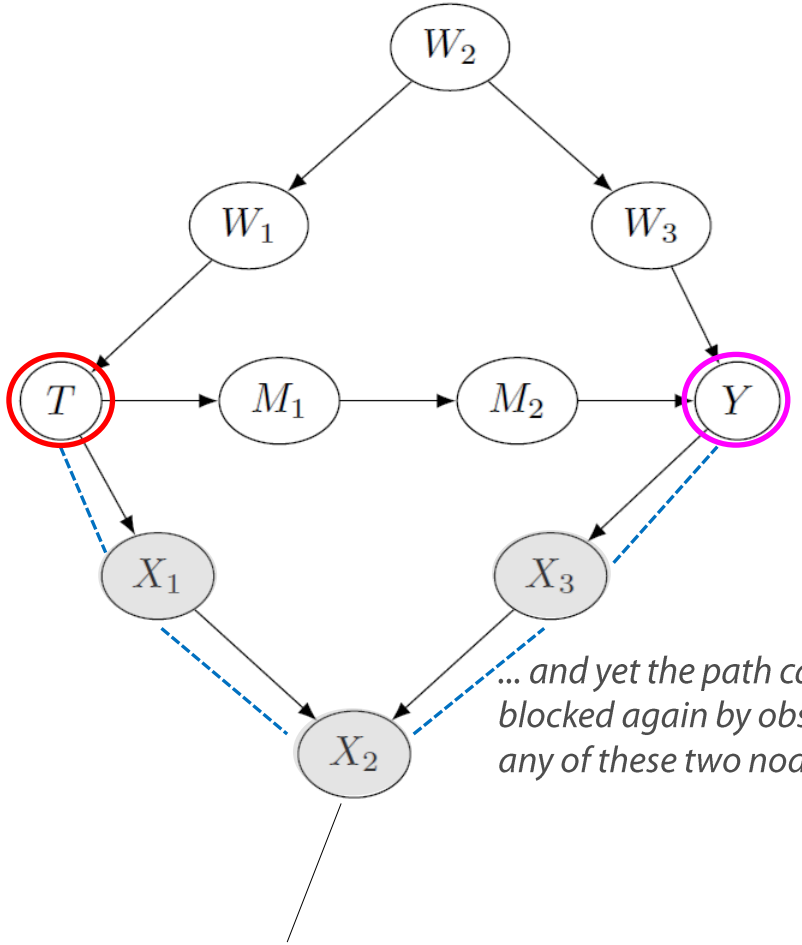
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes A and B is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



D-Separation in Graphical Models

- **Dependency Separation (d-separation)**

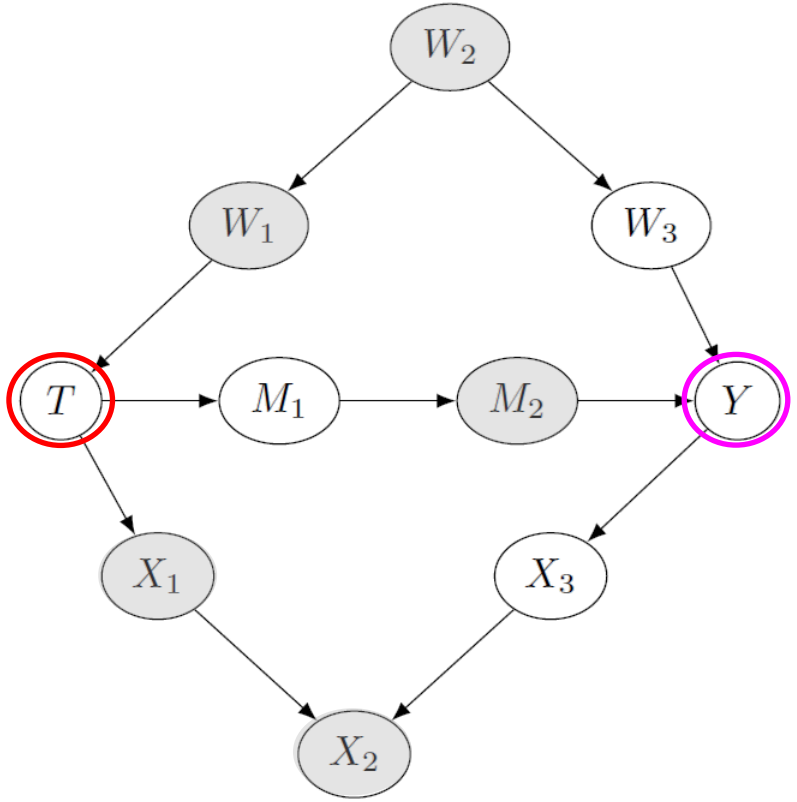
  Any two nodes X and Y in a graphical model are ***d-separated*** whenever the observations $\{Z_o\}$ are such that <u>all paths</u> between X and Y are blocked

  Observed Variables

  In that case we have

  $$\langle X \perp Y \mid \{Z_o\} \rangle$$

  REMEMBER: to have <u>independence</u>, <u>ALL</u> paths must be blocked



These observations make the two nodes <u>d-separated</u>

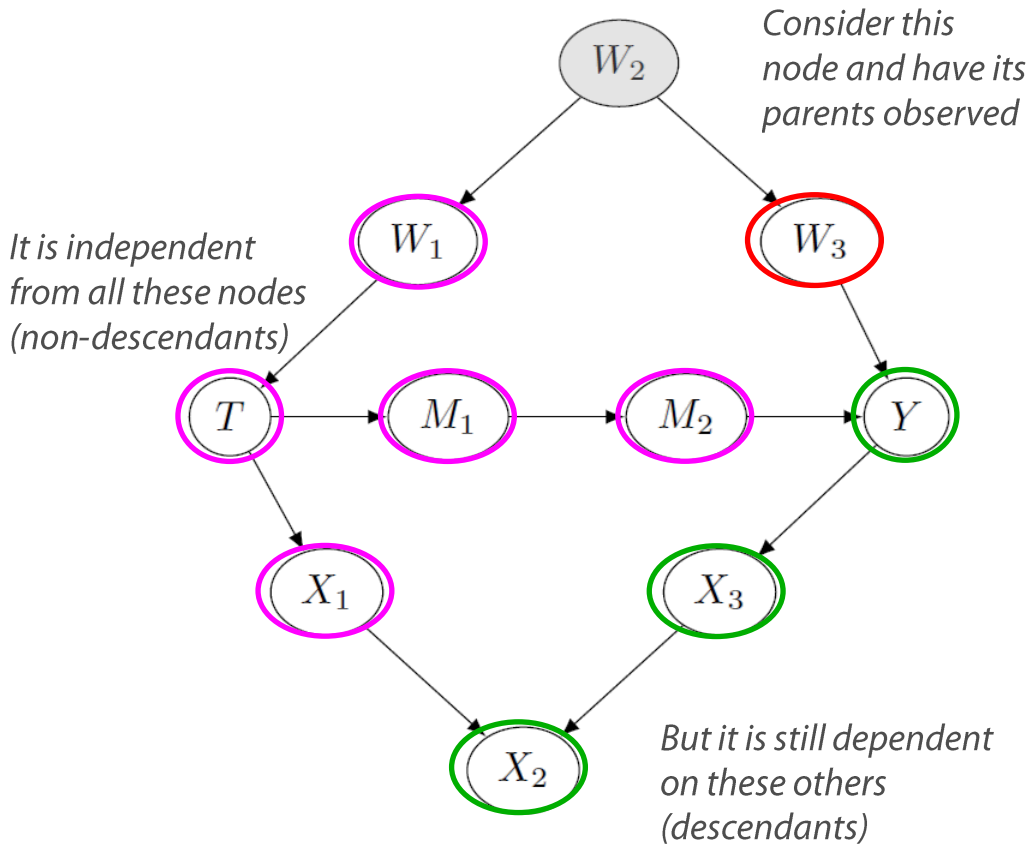
Graphical models: fundamental assumptions

- **Minimality**

  Adjacent nodes in the graph are dependent

- **Local Markov Assumption**

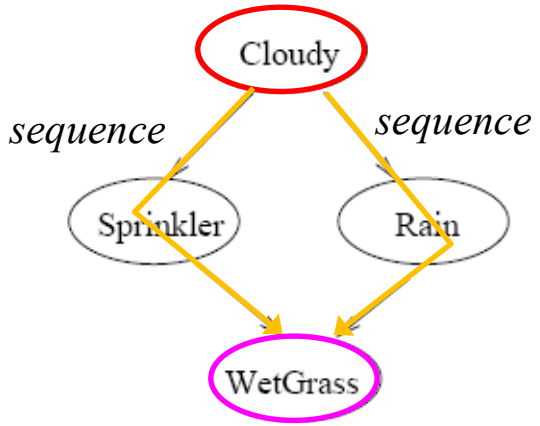
  Given its parents in the graph, a node X is independent of all its non-descendants



D-Separation in Graphical Models

Example:

Cloudy and WetGrass are independent when both paths in color are blocked



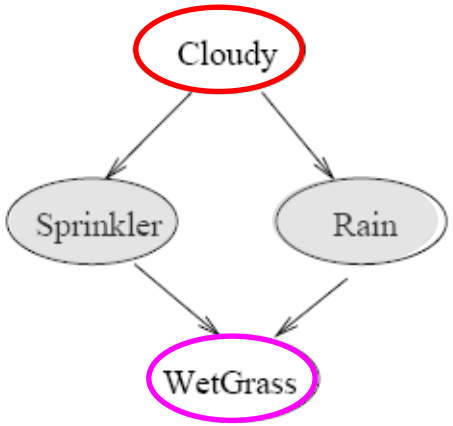
D-Separation in Graphical Models

Example:

Cloudy and WetGrass are independent when both paths in color are blocked

These are two *sequences*:
Sprinkler and Rain must be known

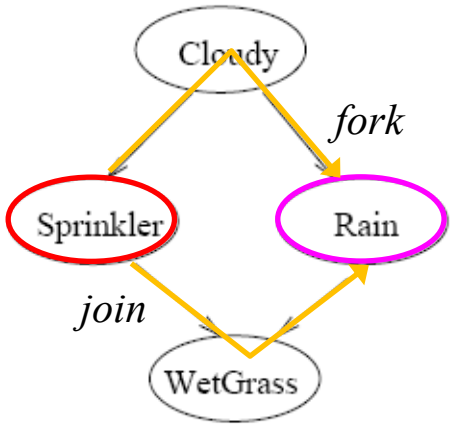
<Cloudy ⊥ WetGrass | Sprinkler, Rain>



D-Separation in Graphical Models

Example:

Sprinkler and Rain are independent when both paths in color are blocked



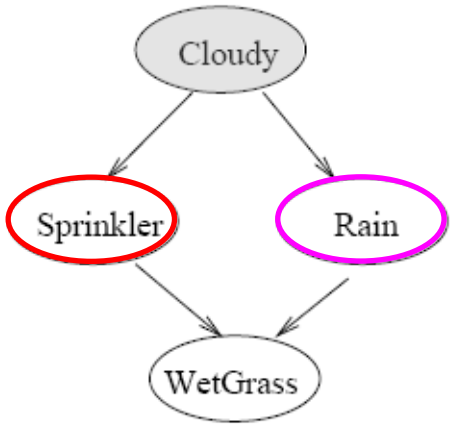
D-Separation in Graphical Models

Example:

Sprinkler and Rain are independent when
both paths in color are blocked

One *fork* and one *collider*:
Cloudy must be known whereas WetGrass must be unknown

$< \text{Sprinkler} \perp \text{Rain} \mid \text{Cloudy} >$



Graphical Models in general

Graphical models as a probabilistic method

- ***Advantages***

  Correctness (of representation)

  Independence in the graph model
  implies independence in the joint probability distribution

$$\langle X \perp Y \mid \{Z\} \rangle_{GM} \Rightarrow \langle X \perp Y \mid \{Z\} \rangle_{JPD}$$

  In a *finitary setting*, probabilistic inference is *always computable*

  Graph models are easy to draw and read (compared to JPDs)

Graphical models as a probabilistic method

- ***Limitations***

  No completeness

  Not all independence models can be captured via univariate factorization

  Counterexample (*Verma constraint*):

$$\langle X \perp Y \rangle$$
$$\langle X \perp Z \mid Y \rangle$$
$$\langle X \not\perp Z \rangle$$

  No DAG can represent this case without introducing some further dependence relations

  However, DAG are complete for all independence models that are decomposable via univariate factorization

Computing Inference in a Graphical Model

Building a graphical model

- Step 1

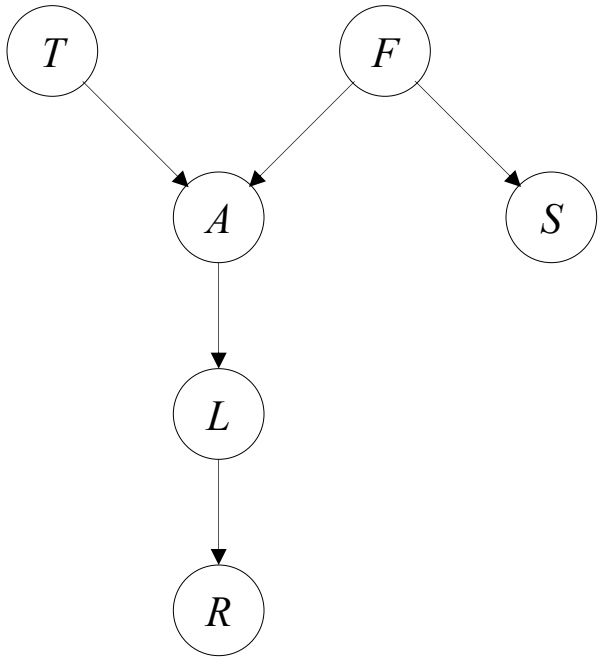
  Defining the nodes, i.e. the random variables

T : (*tampering*)
F : (*fire*)
A : (*alarm*)
S : (*smoke*)
L : (*leaving*)
R : (*report*)

Building a graphical model

- Step 2

  Defining the structure, i.e. the graph



We are thus saying that:

$<T \perp F>$ (but they become dependent when any of A, L or R are known)

$<A \perp S \mid F>$

$<L \perp T \mid A>$

$<L \perp F \mid A>$

$<A \perp R \mid L>$

T : (*tampering*)

F : (*fire*)

A : (*alarm*)

S : (*smoke*)

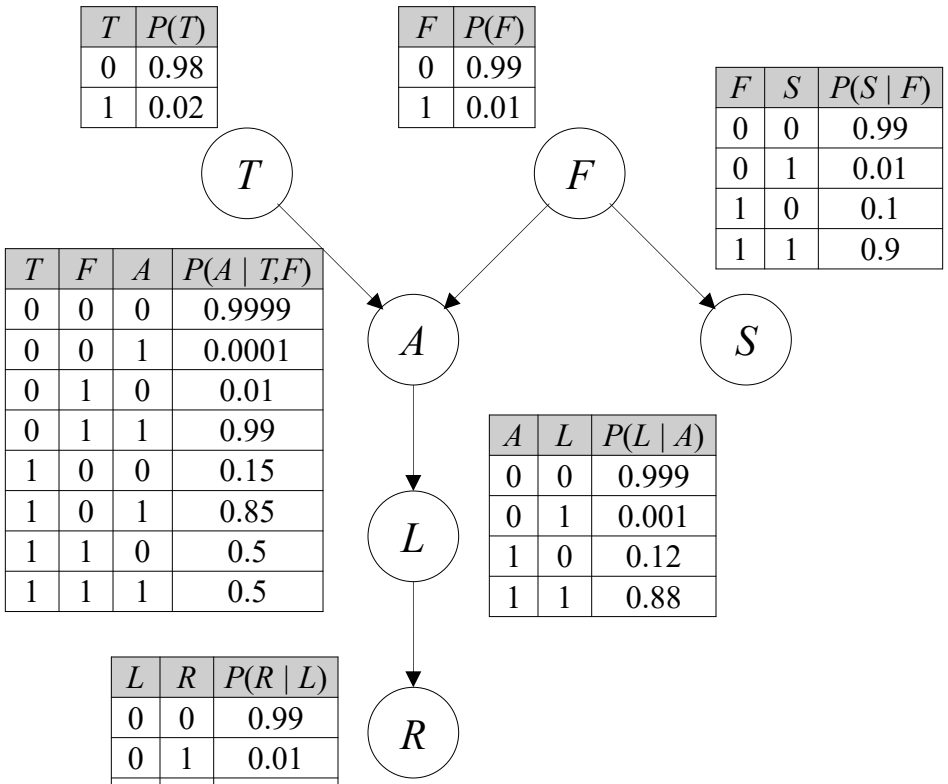
L : (*leaving*)

R : (*report*)

Building a graphical model

- Step 3

  Defining *conditional probability tables – CPTs*



T	$P(T)$
0	0.98
1	0.02

F	$P(F)$
0	0.99
1	0.01

F	S	$P(S \mid F)$
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	$P(A \mid T,F)$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	$P(L \mid A)$
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

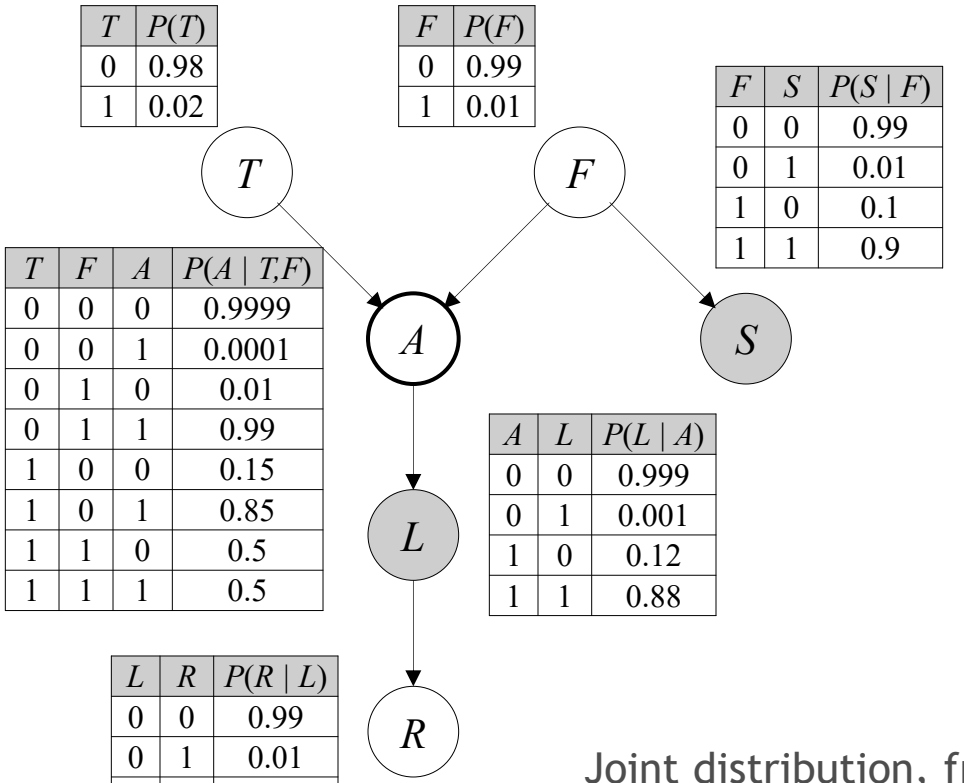
L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

T : (*tampering*)
F : (*fire*)
A : (*alarm*)
S : (*smoke*)
L : (*leaving*)
R : (*report*)

Probabilistic inference

- Step 4

  Consider a specific problem



T	P(T)
0	0.98
1	0.02

F	P(F)
0	0.99
1	0.01

F	S	P(S | F)
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	P(A | T,F)
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	P(L | A)
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	P(R | L)
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

Example: finding A given $L=1$ e $S=0$

$$P(A|L=1,S=0) = \frac{P(A,L=1,S=0)}{P(L=1,S=0)}$$

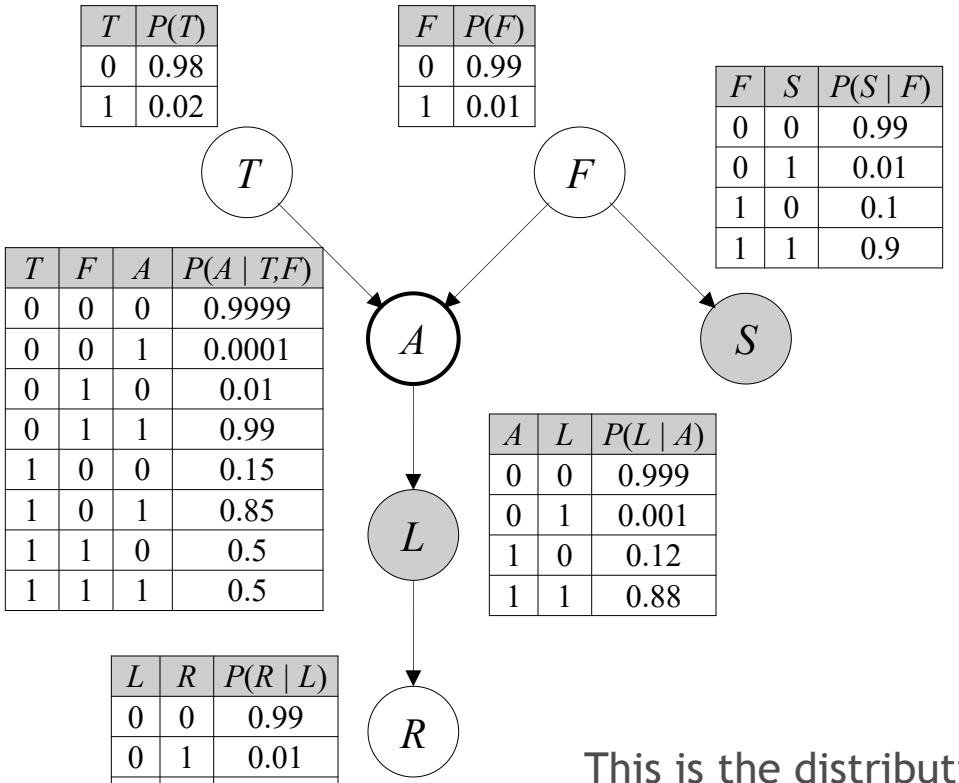
Joint distribution, from the graph:

$$P(T,F,A,S,L,R) = P(T)P(F)P(A|T,F)P(S|F)P(L|A)P(R|L)$$

Probabilistic inference

- Step 5

  Computing the answer



T	$P(T)$
0	0.98
1	0.02

F	$P(F)$
0	0.99
1	0.01

F	S	$P(S \mid F)$
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	$P(A \mid T,F)$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	$P(L \mid A)$
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

Note that:

$$P(A|L=1,S=0) = \frac{P(A,L=1,S=0)}{P(L=1,S=0)}$$

This is a normalizing term:
it can be computed from

$$P(A,L=1,S=0)$$

In fact:

$$P(L=1,S=0) = \sum_{A} P(A,L=1,S=0)$$

Typically, the most time-consuming computations in an inference problem are marginalizations

This is the distribution to be computed:

$$P(A,L=1,S=0) = \sum_{T,F,R} P(T)P(F)P(A|T,F)P(S=0|F)P(L=1|A)P(R|L=1)$$

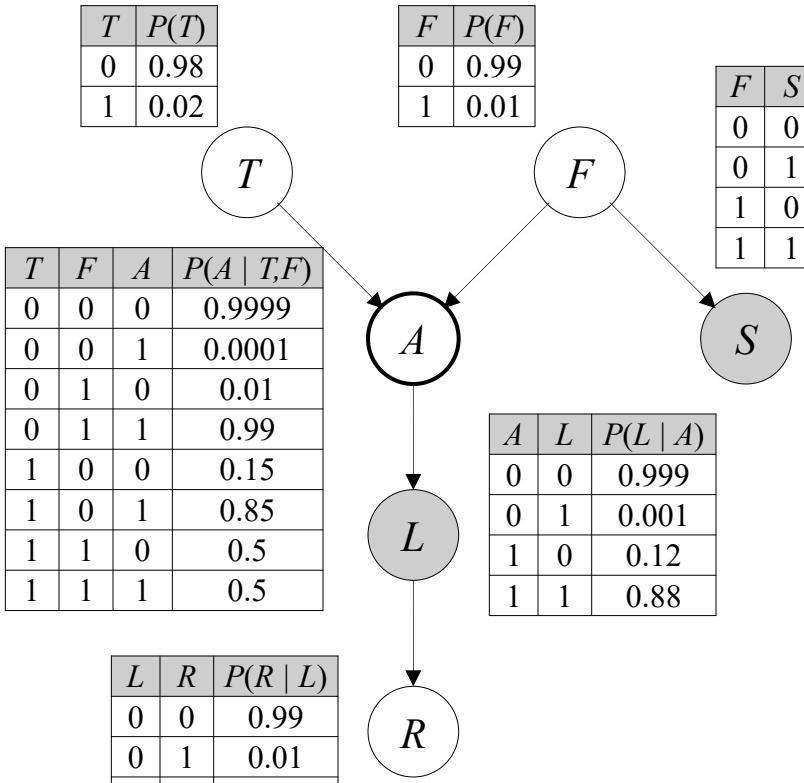
Probabilistic inference

- Step 5

  Computing the answer

This is the distribution to be computed:

$$P(A, L=1, S=0) = \sum_{T,F,R} P(T)P(F)P(A|T,F)P(S=0|F)P(L=1|A)P(R|L=1)$$



T	$P(T)$
0	0.98
1	0.02

F	$P(F)$
0	0.99
1	0.01

F	S	$P(S \mid F)$
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	$P(A \mid T,F)$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	$P(L \mid A)$
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

It can be rewritten as:

$$= P(L=1|A) \sum_{T} \sum_{F} P(A|T,F)P(T)P(F)P(S=0|F) \sum_{R} P(R|L=1)$$

This sum has value 1

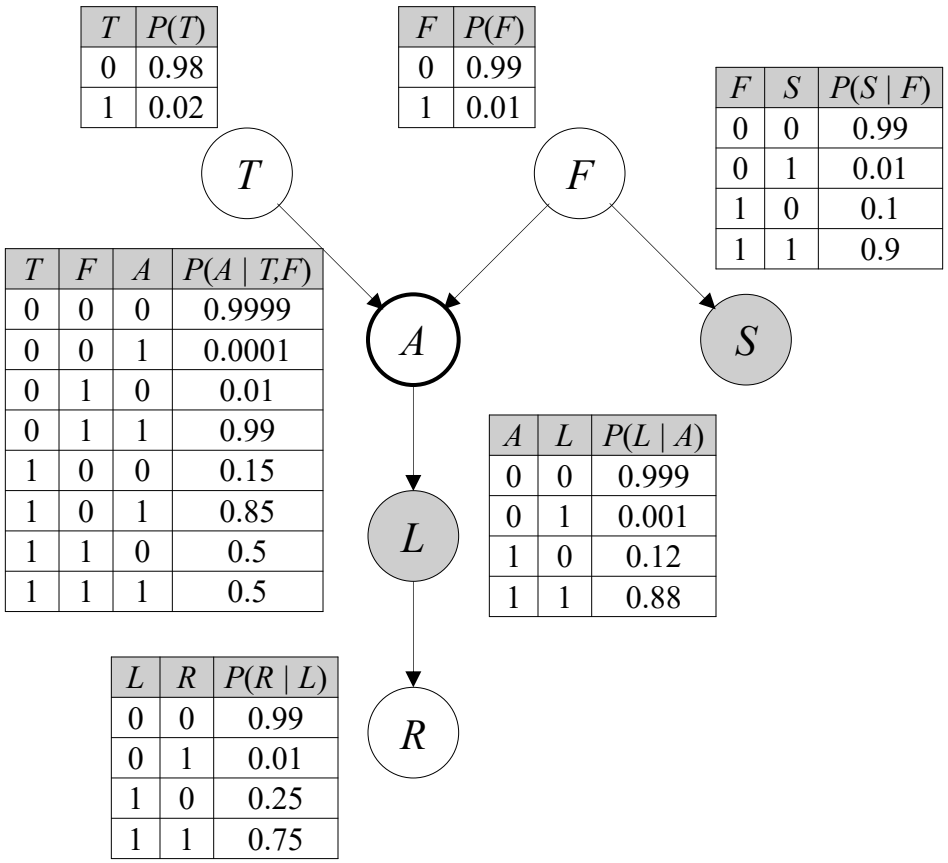
(*It is not surprising given that* $<A \perp R \mid L>$)

Probabilistic inference

- Step 5

  Computing the answer

$$P(A, L=1, S=0) = P(L=1|A) \sum_T \sum_F P(A|T,F)P(T)P(F)P(S=0|F)$$



T	$P(T)$
0	0.98
1	0.02

F	$P(F)$
0	0.99
1	0.01

F	S	$P(S \mid F)$
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	$P(A \mid T,F)$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	$P(L \mid A)$
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

By convention, we write:

$$P(A, L=1, S=0) = f_{T,F,S=0}(A) f_{L=1}(A)$$

where the f are the *factors* of the method also known as *elimination of variables:*

$$f_{T,F,S=0}(A) := \sum_T \sum_F P(A|T,F)P(T)P(F)P(S=0|F)$$

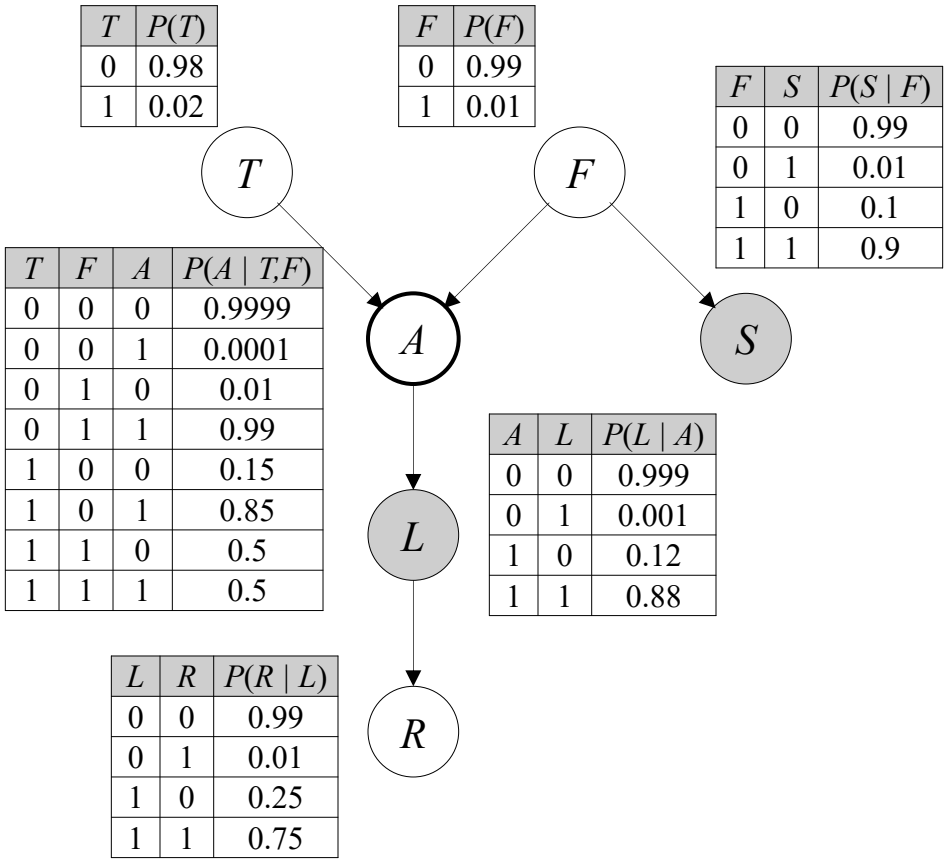
$$f_{L=1}(A) := P(L=1|A)$$

Note in passing that *factors* f may not be probabilities (i.e. they may not sum up to 1).

Probabilistic inference

- Step 5

  Computing the answer



T	$P(T)$
0	0.98
1	0.02

F	$P(F)$
0	0.99
1	0.01

F	S	$P(S \mid F)$
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	$P(A \mid T,F)$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	$P(L \mid A)$
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

Note that:

$$P(A, L=1, S=0) = f_{T,F,S=0}(A) f_{L=1}(A)$$

This factor comes from the *parents* of A

This factor comes from the *descendants* of A

This is true
for any node A that *d-separates* the graph

Variable elimination for graphical models

- General idea

  Write the marginal joint probability from the query in the form:

$$P(\{X_r\}, \{X_o\}) = \sum_{\{X_i\}} \prod_{X} P(X \mid parents(X))$$

1) Find the best ordering of terms for the marginalization of irrelevant variables:
2) Move summations 'inside' the product as much as possible (i.e. find *factors f*)
3) Compute factors (i.e. by sum of products) and obtain numbers (i.e. *terms*)
4) Plug these *terms* into the product and obtain a simpler form for $P(\{X_r\}, \{X_o\})$
5) Wrap it up and compute the response:

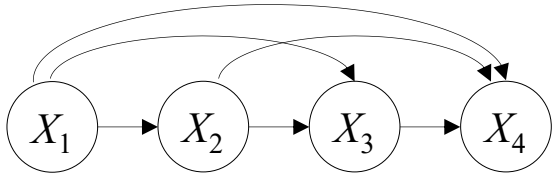
$$P(\{X_r\}|\{X_o\}) = \frac{P(\{X_r\}, \{X_o\})}{\sum_{\{X_r\}} P(\{X_r\}, \{X_o\})}$$

Remember: the method is NP-complete (anyway)

Graphical Models in Action

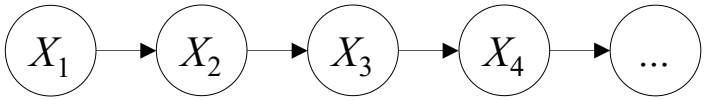
Example of graphical models

- Complete dependency



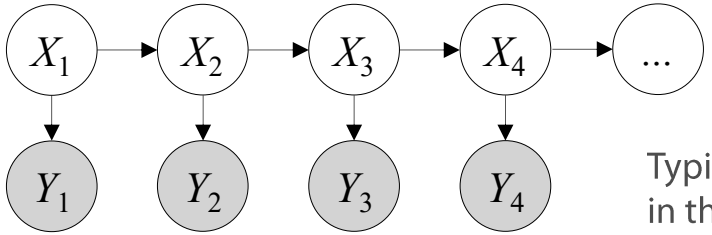
$$P(X_1, X_2, X_3, X_4) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2)\, P(X_4 \mid X_1, X_2, X_3)$$

- Markovian model



$$P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i \mid X_{i-1})$$

- 'Hidden' Markovian model

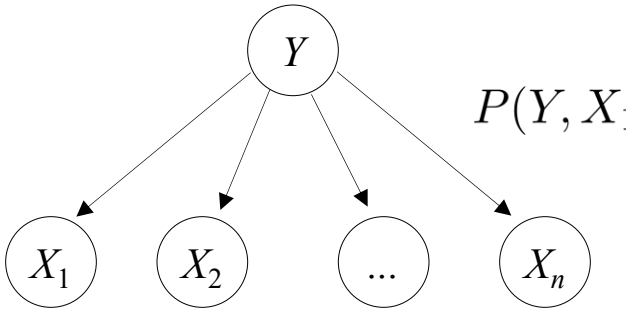


Typically, nodes X_i are *hidden*, in the sense of *non-observable* (see later, about *learning*)

$$P(X_1, \ldots, X_n Y_1, \ldots, Y_n) = P(X_1) P(Y_1 | X_1) \prod_{i=2}^{n} P(X_i \mid X_{i-1}) P(Y_i \mid X_i)$$

Example: *anti-spam filter*

a.k.a. '*Naïve* (*Discrete*) *Bayesian Classifier*'



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

Anti-spam filter:

- All random variables are *binomial* (value: either 0 or 1)
- Y represents the class of the message: 1 *spam*, 0 not-*spam*
- Each X_i represents the occurrence of the word i in the message

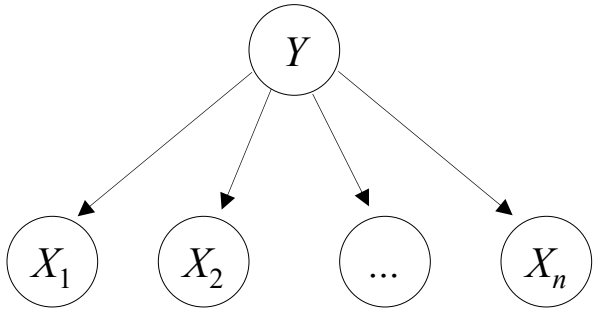
Assume (*for now*) that the probabilities are given

As we will see, finding the 'right' numbers is a *learning* problem (see after)

Inference in the *anti-spam filter*

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

Given a message with occurrence values $\{X_k\}$,
the class with the highest conditional probability is determined



The message is *spam* if

$$\frac{P(Y = 1 \mid \{X_k\})}{P(Y = 0 \mid \{X_k\})} > \lambda$$

Note that:

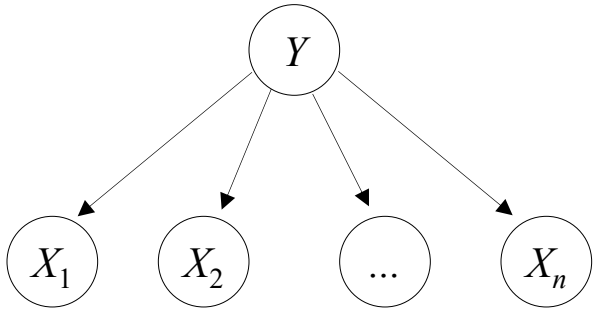
Bayes' Theorem

$$P(Y = 1 \mid \{X_k\}) = \frac{P(\{X_k\} \mid Y = 1)P(Y = 1)}{\sum_Y P(\{X_k\} \mid Y)P(Y)} = \frac{P(Y = 1) \prod_k P(X_k \mid Y = 1)}{\sum_Y P(Y) \prod_k P(X_k \mid Y)}$$

Conditional independency

Inference in the *anti-spam filter*

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



Given a message with occurrence values $\{X_k\}$,
the class with the highest conditional probability is determined

The message is *spam* if

$$\frac{P(Y=1 \mid \{X_k\})}{P(Y=0 \mid \{X_k\})} > \lambda$$

Note that:

Bayes' Theorem

$$P(Y=1 \mid \{X_k\}) = \frac{P(\{X_k\} \mid Y=1)P(Y=1)}{\sum_Y P(\{X_k\} \mid Y)P(Y)} = \frac{P(Y=1)\prod_k P(X_k \mid Y=1)}{\sum_Y P(Y)\prod_k P(X_k \mid Y)}$$

Conditional independency

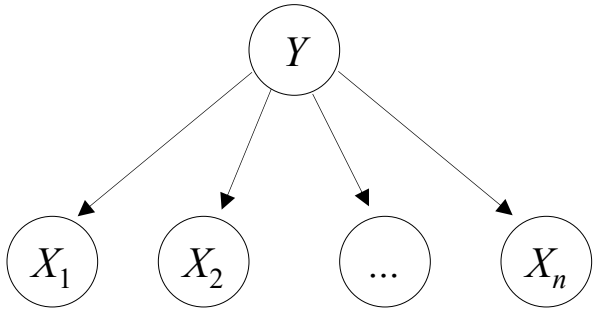
Therefore:

$$\frac{P(Y=1 \mid \{X_k\})}{P(Y=0 \mid \{X_k\})} = \frac{P(Y=1)}{P(Y=0)} \prod_k \frac{P(X_k \mid Y=1)}{P(X_k \mid Y=0)}$$

Inference in the *anti-spam filter*

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

Given a message with occurrence values $\{X_k\}$,
the class with the highest conditional probability is determined



The message is *spam* if

$$\frac{P(Y=1 \mid \{X_k\})}{P(Y=0 \mid \{X_k\})} > \lambda$$

Note that:

Bayes' Theorem

$$P(Y=1 \mid \{X_k\}) = \frac{P(\{X_k\} \mid Y=1)P(Y=1)}{\sum_Y P(\{X_k\} \mid Y)P(Y)} = \frac{P(Y=1)\prod_k P(X_k \mid Y=1)}{\sum_Y P(Y)\prod_k P(X_k \mid Y)}$$

Conditional independency

Therefore:

$$\frac{P(Y=1 \mid \{X_k\})}{P(Y=0 \mid \{X_k\})} = \frac{P(Y=1)}{P(Y=0)} \prod_k \frac{P(X_k \mid Y=1)}{P(X_k \mid Y=0)}$$

The logarithm is used
to simplify computations:

$$\log \frac{P(Y=1|\{X_k\})}{P(Y=0|\{X_k\})} = \log \frac{P(Y=1)}{P(Y=0)} + \sum_k \log \frac{P(X_k|Y=1)}{P(X_k|Y=0)}$$

An aside: plate notation

A shorthand notation for graphical models

