Probabilistic reasoning: *inference*

Marco Piastra

Probabilistic Inference

- General setting

  The starting point is a fully-specified joint probability distribution

  $$P(X_1, X_2, \ldots, X_n)$$

  In an *inference* problem, the set of random variables $\{X_1, X_2, \ldots, X_n\}$ is divided into three categories:

  1) *Observed variables* $\{X_e\}$, i.e. having a definite (and certain) value
  2) *Irrelevant variables* $\{X_r\}$, (also *latent variables*) i.e. which are not directly part of the answer
  3) *Relevant variables* $\{X_f\}$, i.e. which are part of the answer we seek for

  In general, the problem is finding:

  $$P(\{X_f\}|\{X_e\}) = \sum_{\{X_r\}} P(\{X_f\},\{X_r\}|\{X_e\})$$

  - "Decidability" (actually "computability") is not an issue (*in a discrete setting)

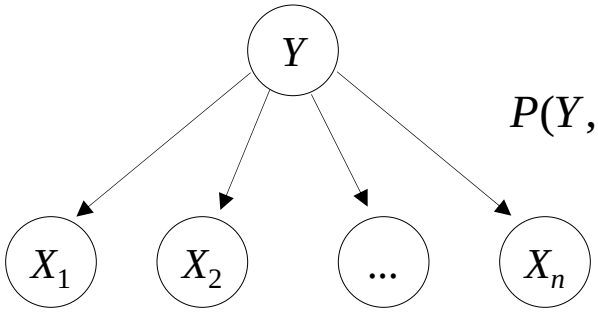
    Given that the joint probability distribution is completely specified

  - Computational efficiency can be a problem

    The number of value combinations grows exponentially with the number of random variables

Example: *anti-spam filter*

Typically (e.g. Mozilla Thunderbird): *'Naive* (*Discrete*) *Bayesian Classifier'*



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

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 *i* word in the message

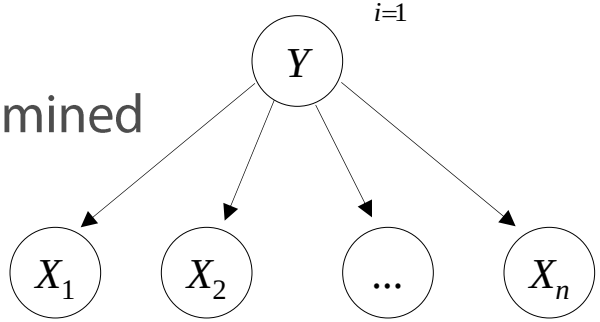
Assume (*for now*) that the probabilities are given

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

Inference in the *anti-spam filter*

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

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



The message is *spam* if $\dfrac{P(Y=1 \mid \{X_k\})}{P(Y=0 \mid \{X_k\})} > \lambda$

Note that:

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

Bayes' Theorem

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:

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

Building a graphical model

- Step 1

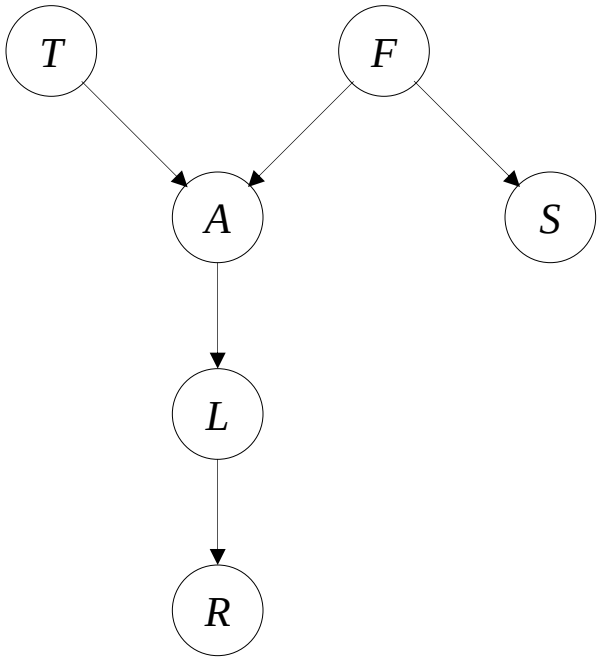
  Defining the nodes, i.e. the random variables

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*)

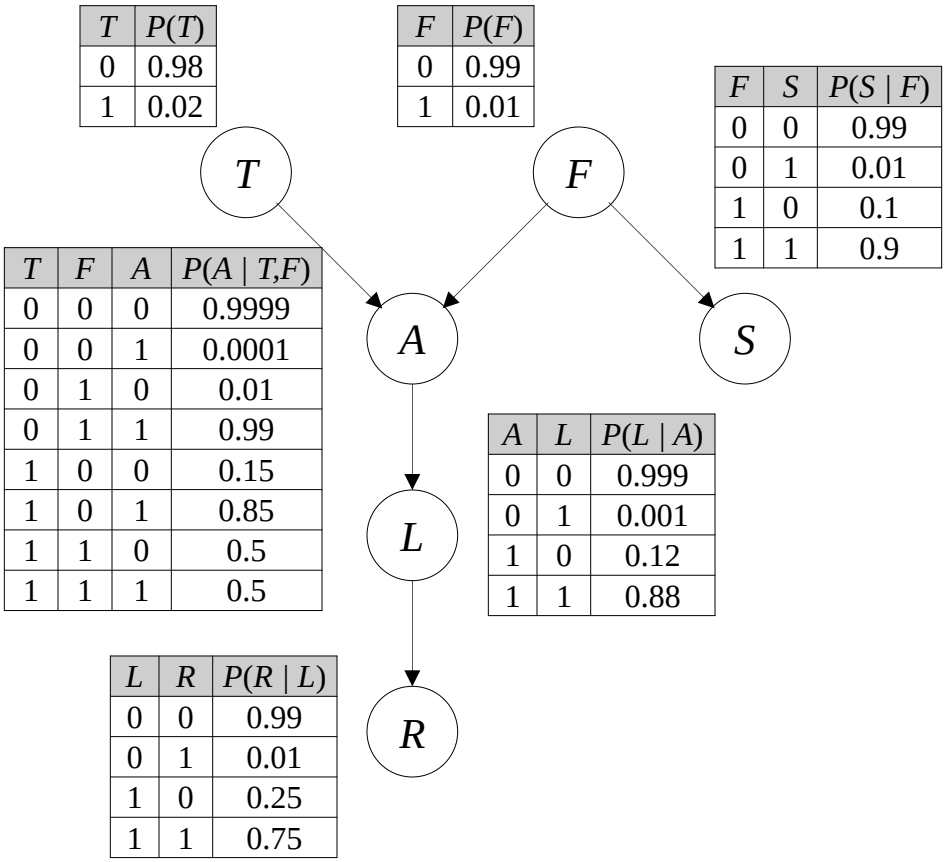
L : (*leaving*)

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

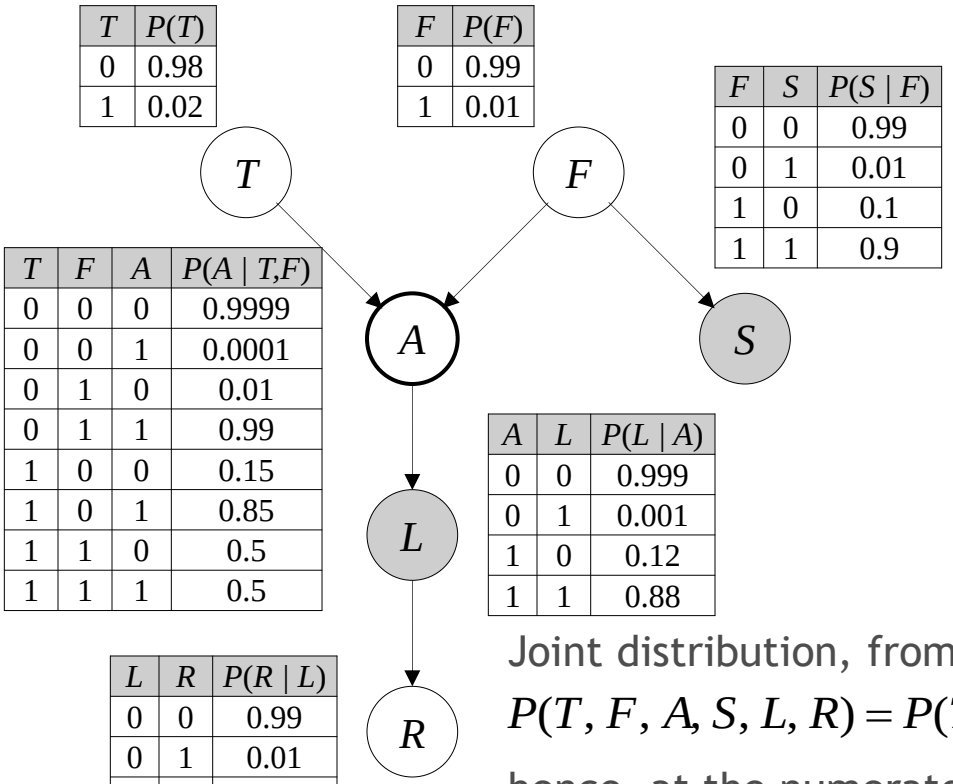
L	R	$P(R \mid L)$
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

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

Probabilistic inference

- Step 4

  Defining 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:

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

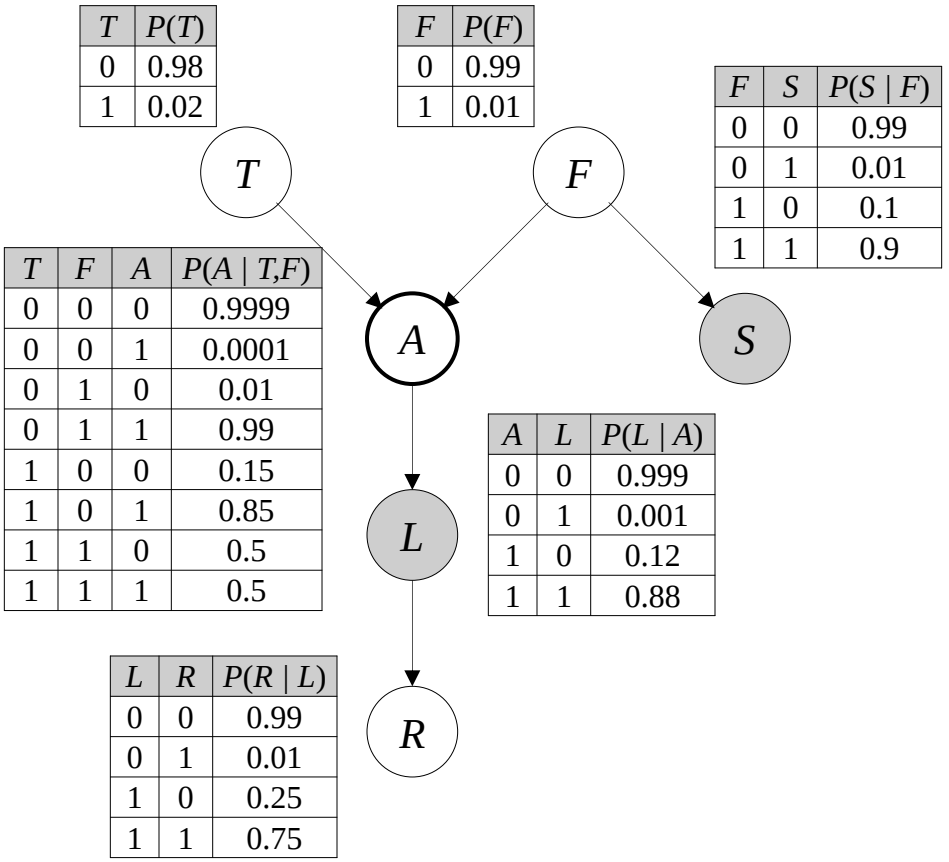
hence, at the numerator:

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

Probabilistic inference

- Step 5

  Computing the answer



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

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:

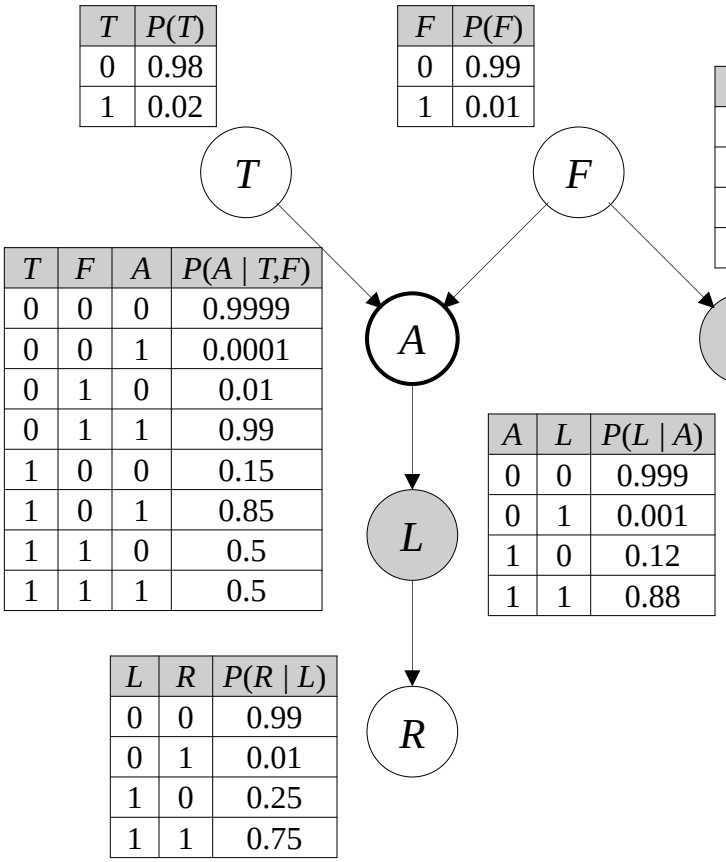
$$P(L=1,S=0)=\sum_{A}P(A,L=1,S=0)$$

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

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

By rewriting:

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

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

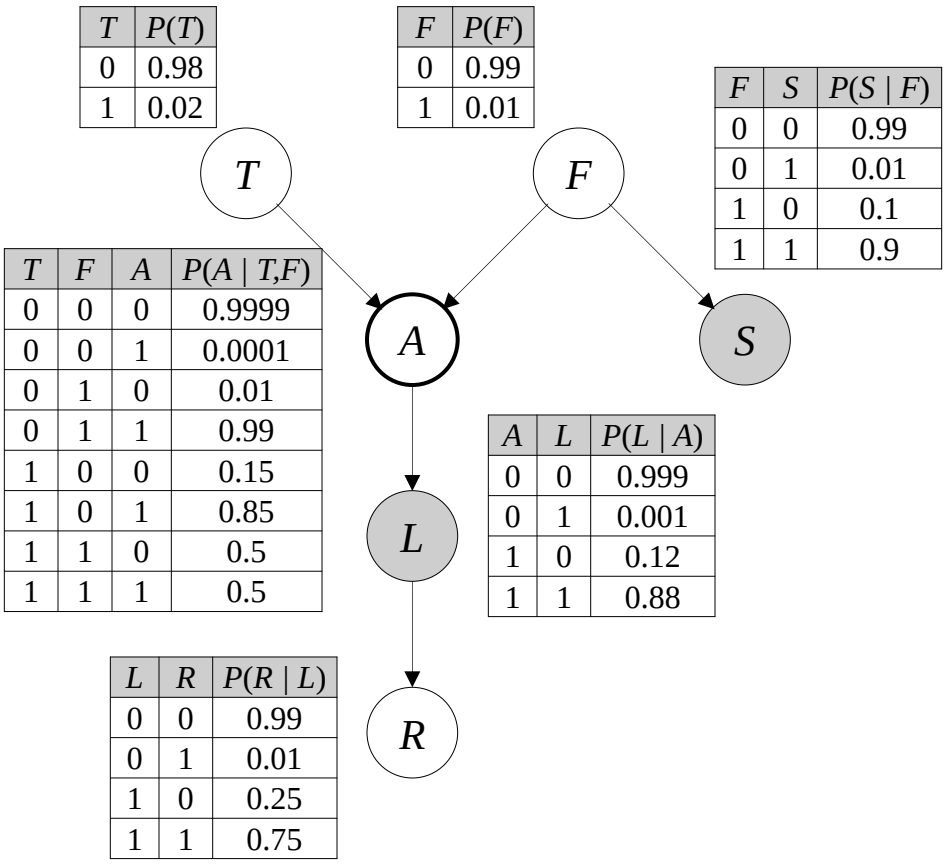
$$= P(L=1|A) \sum_T \sum_F P(A|T,F)P(T)P(F)P(S=0|F) \sum_R P(R|L=1)$$

This sum has value 1
This is not surprising
given that $<A \perp R \mid L>$

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

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*.

Note in passing that *factors* f are not probabilities (i.e. they do not sum to 1). For instance:

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

In general:
By summing w.r.t. a condition*ed* variable we obtain a marginal probability

$$P(A|S) = \sum_L P(A, L|S)$$

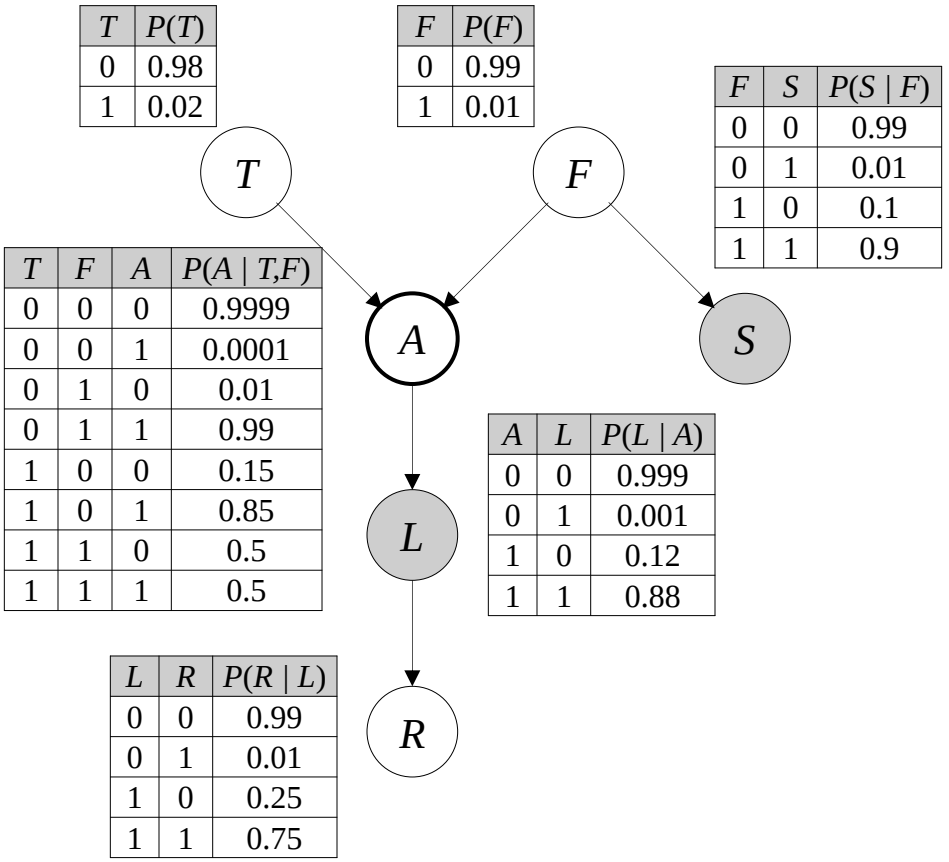
By summing w.r.t. a condition*ing* variable we just obtain a function

$$f_S(A,L) = \sum_S P(A, L|S)$$

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 | 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

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 in general,
whenever node A *d-separates* the graph

Variable elimination for graphical models

- General idea

  Write the joint probability of the query in the form:

$$P(\{X_f\},\{X_e\}) = \sum_{\{X_r\}} \prod_{X_i} P(X_i \mid parents(X_i))$$

1) Find the more convenient order for the marginalization w.r.t . the latent 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 $P(\{X_f\},\{X_e\})$
5) Wrap it up and compute the response:

$$P(\{X_f\}|\{X_e\}) = \frac{P(\{X_f\},\{X_e\})}{\sum_{\{X_f\}} P(\{X_f\},\{X_e\})}$$

Remember: the method is NP-complete (anyway)