

Module 4: Lecture 8

Bayesian models

Topics

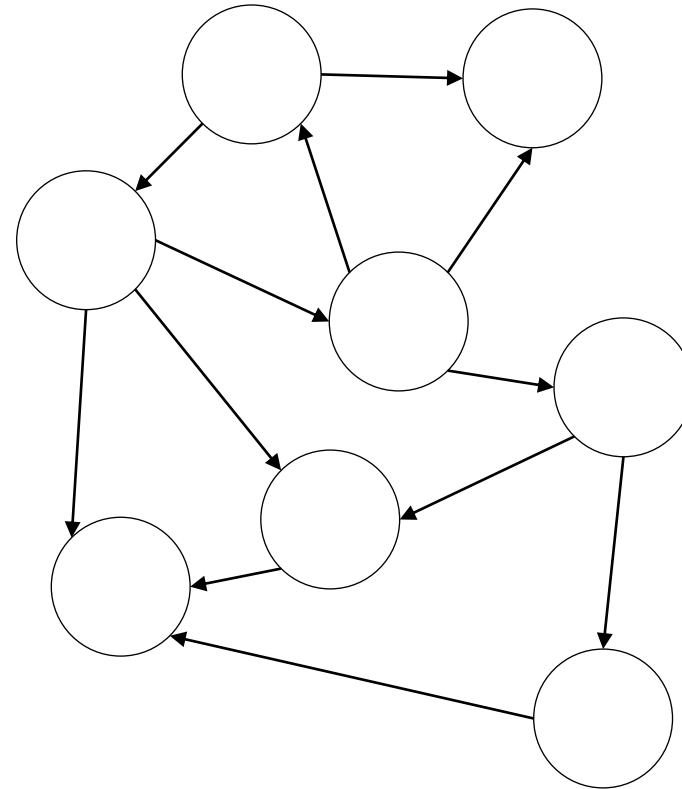
- Probabilistic graphical models
- Bayesian networks
- Computing probabilities
- Estimating probabilities
- The Naïve Bayes classifier
- Fruit example

Probabilistic graphical models

Graphs

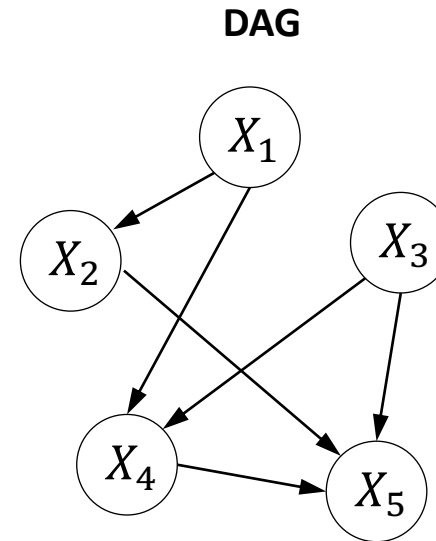
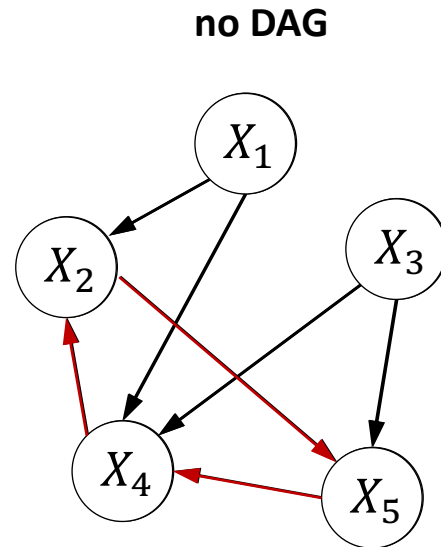
A (directed) *graph* $G = (V, E)$ consists of

- A set V of *vertices* or *nodes*
- A set E of *edges* or *links*



Directed acyclic graphs (DAGs)

Contains no cycles/loops



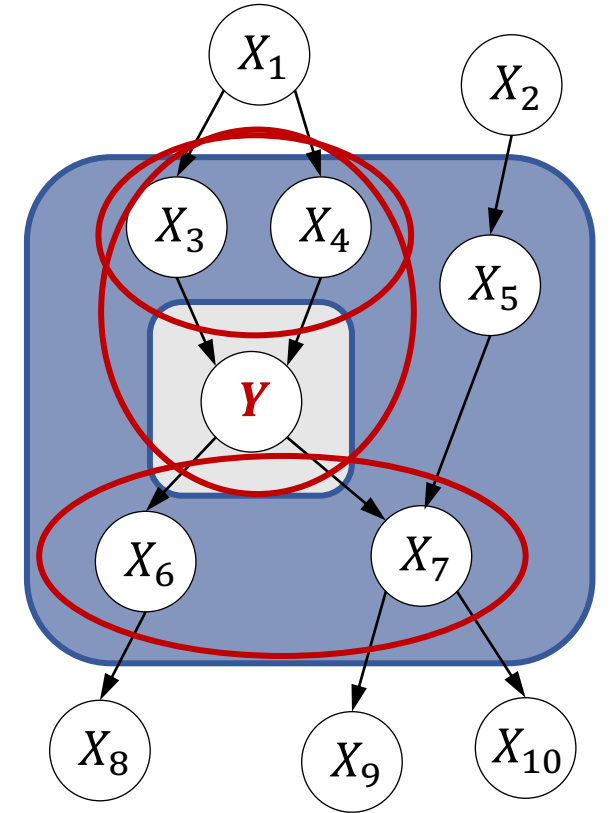
Parents and children

- The *parents* of a node are the nodes with links into it.

$$\text{pa}(Y) = \{X_3, X_4\}$$

- The *children* of a node are the nodes with links to them from that node.

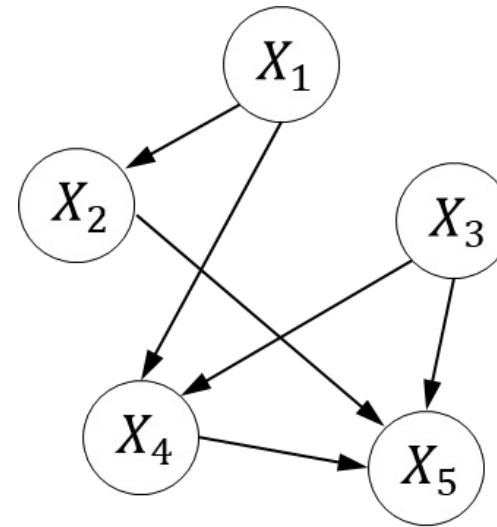
$$\text{ch}(Y) = \{X_6, X_7\}$$



Probabilistic graphical model

A graph that represents a joint distribution of random variables

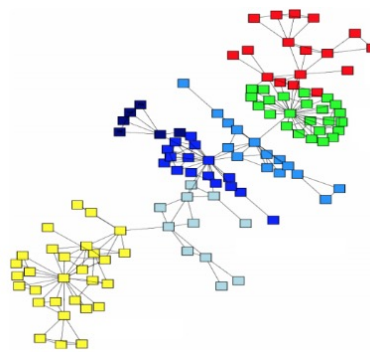
- Vertices: random variables
- Edges: probabilistic relationships



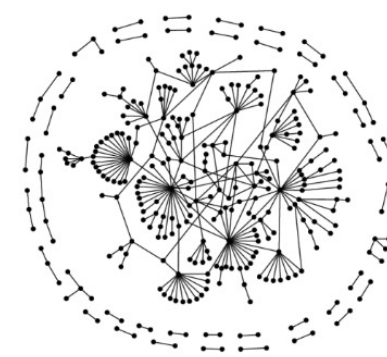
Probabilistic graphical models



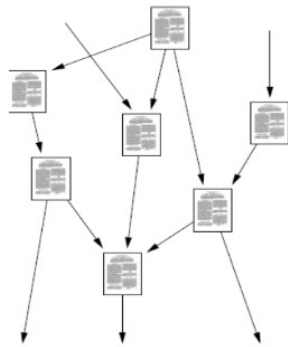
Social networks



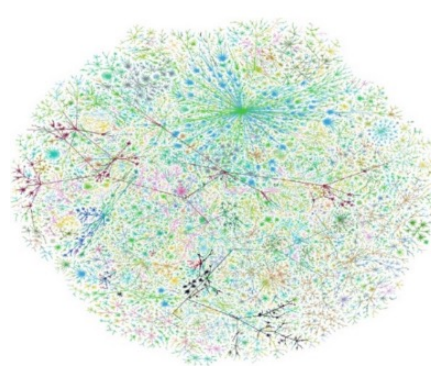
Economic networks



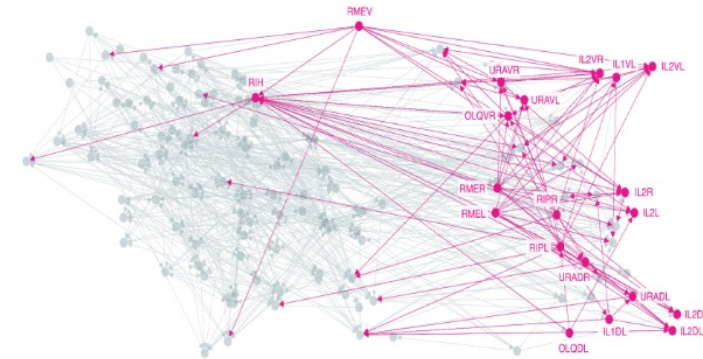
Biomedical networks



Information networks



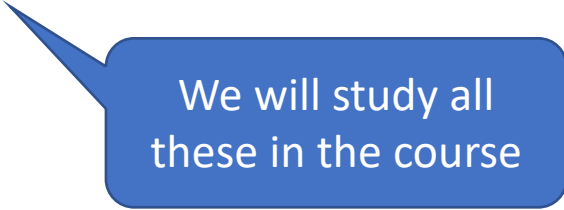
Network of neurons



Internet

Examples of probabilistic graphical models

- Bayesian networks
- Naïve Bayes
- Markov chains
- Neural networks



We will study all
these in the course

A way of representing joint
probability distributions

Bayesian Networks

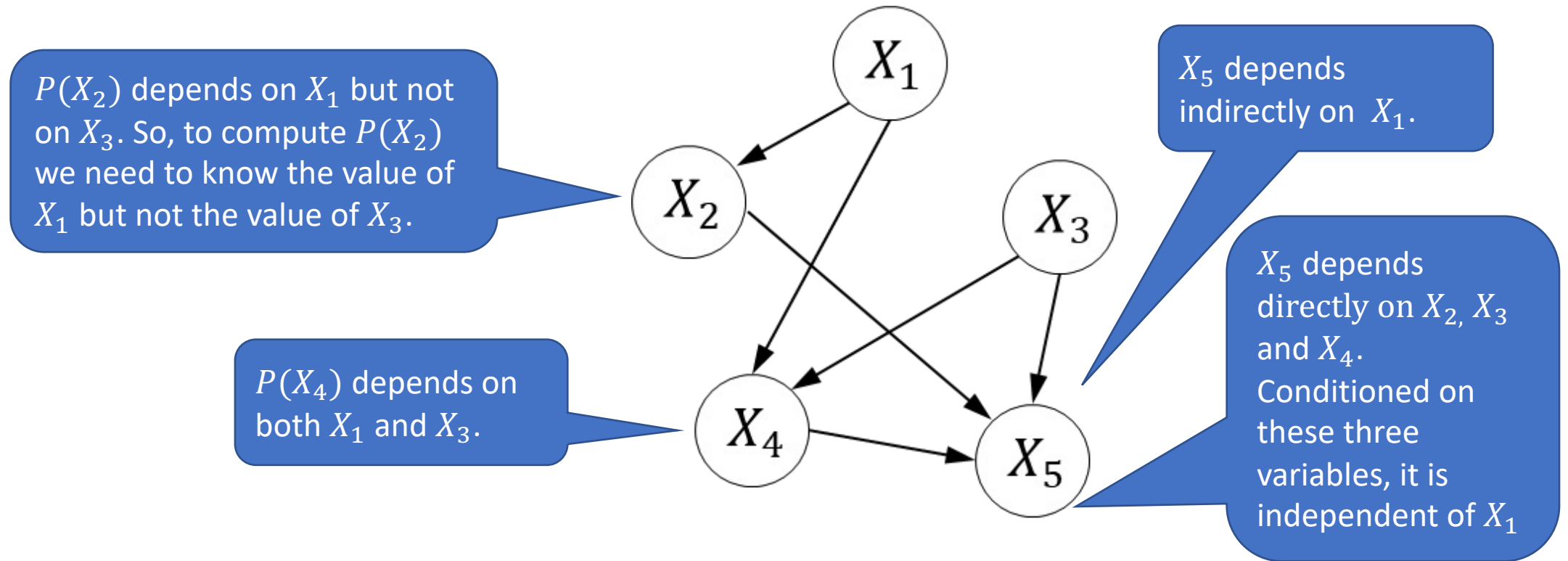
Also known as Bayes nets
and belief networks

Bayesian networks

- A *Bayesian network* consists of
 - A DAG (V, E)
 - A Boolean-valued random variable X_i for each $i \in V$
 - Conditional probability tables $P(X_i | \text{pa}(X_i))$ for each $i \in V$.

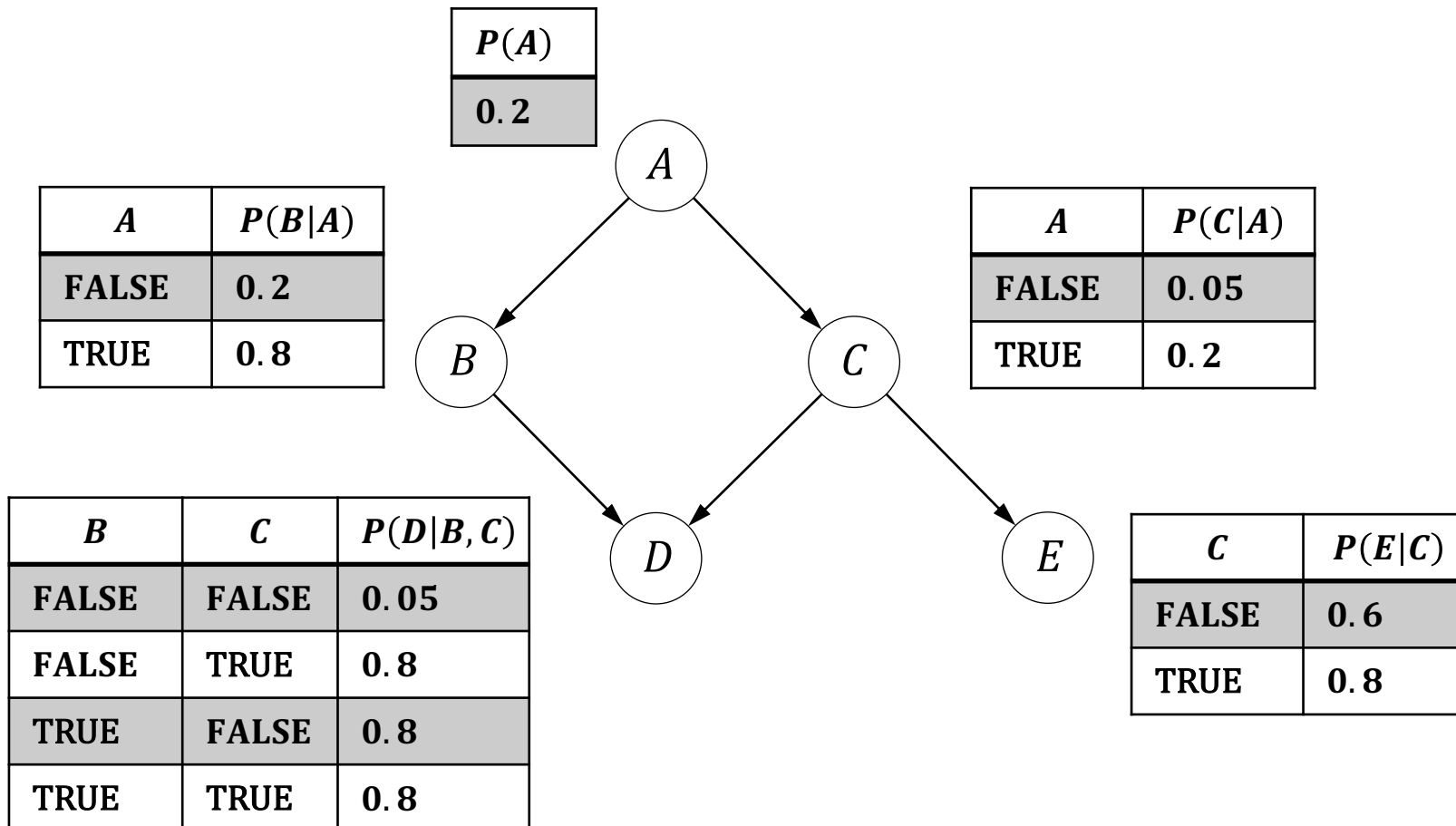
There are more general
definitions of Bayesian networks
but this one will be fine for us!

Intuitive interpretation of the edges



Example of a Bayesian network

Here we see the DAG, the random variables, and the conditional probability tables.



Recall what an ordinary probability table looks like.

Ordinary probability tables

ω	$P(\omega)$
Heads	$1/2$
Tails	$1/2$



ω	$P(\omega)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$



Conditional probability tables

This table describes how the value of $P(C|A)$ depends on the truth-value of A

A	$P(C A)$
FALSE	0.05
TRUE	0.2

If the values in this column were the same, there would be no dependence on A

The values in the column do not sum to 1. Note: $P(C=\text{False} | A) = 1 - P(C=\text{True} | A)$, so latter can be omitted.

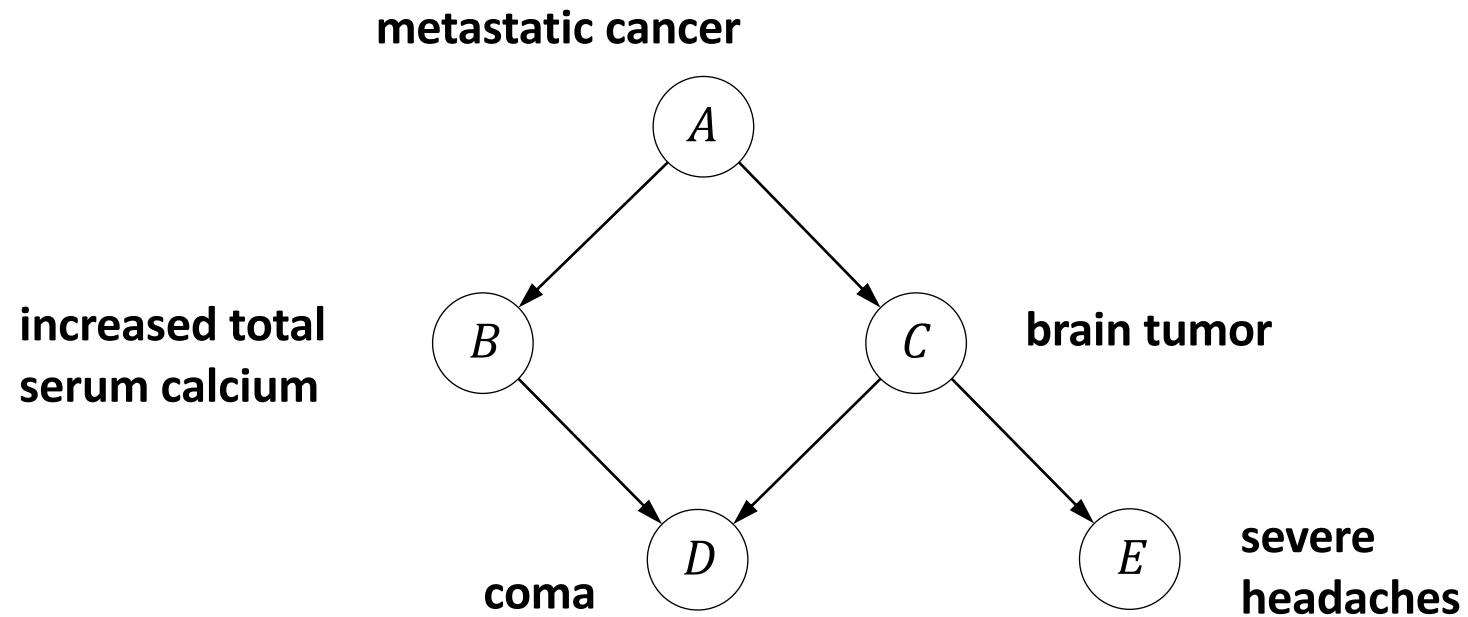
Ordinary truth-tables would have TRUE/FALSE in this column. But this is a probability table

Big tables

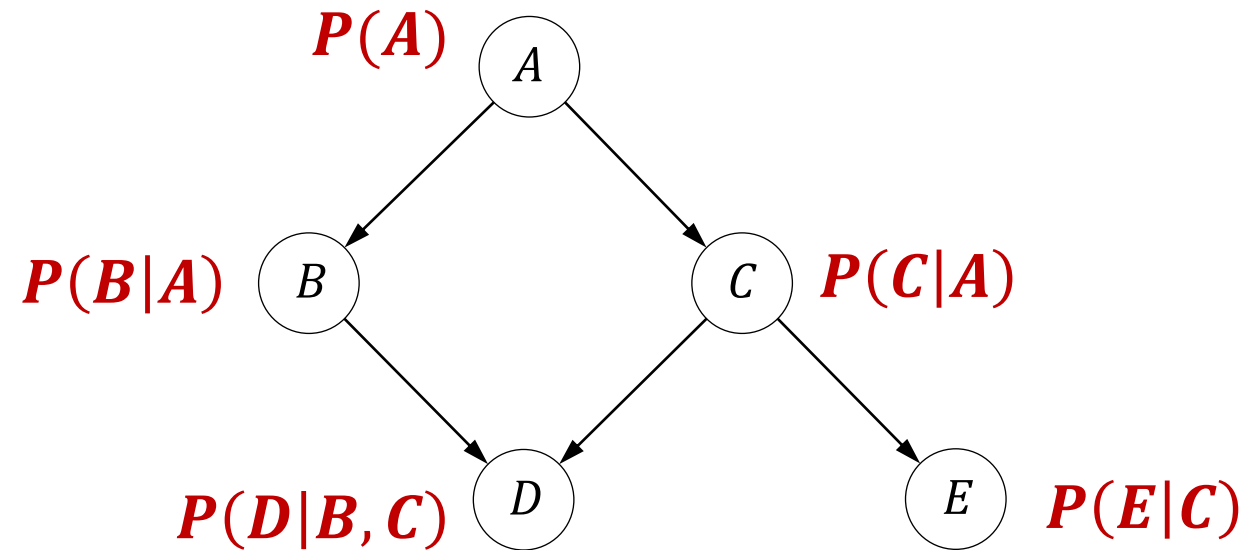
- In general, to describe $P(X_1, X_2, \dots, X_n)$ for all value-combinations of X_1, X_2, \dots, X_n one may use a probability table.
- The size of the full probability table is 2^n . Storing it requires a TB-sized memory when $n = 40$ variables. Collecting data for all combinations gets hard as n grows

But everything does not always depend on everything. Sometimes the tables can be represented much more compactly.

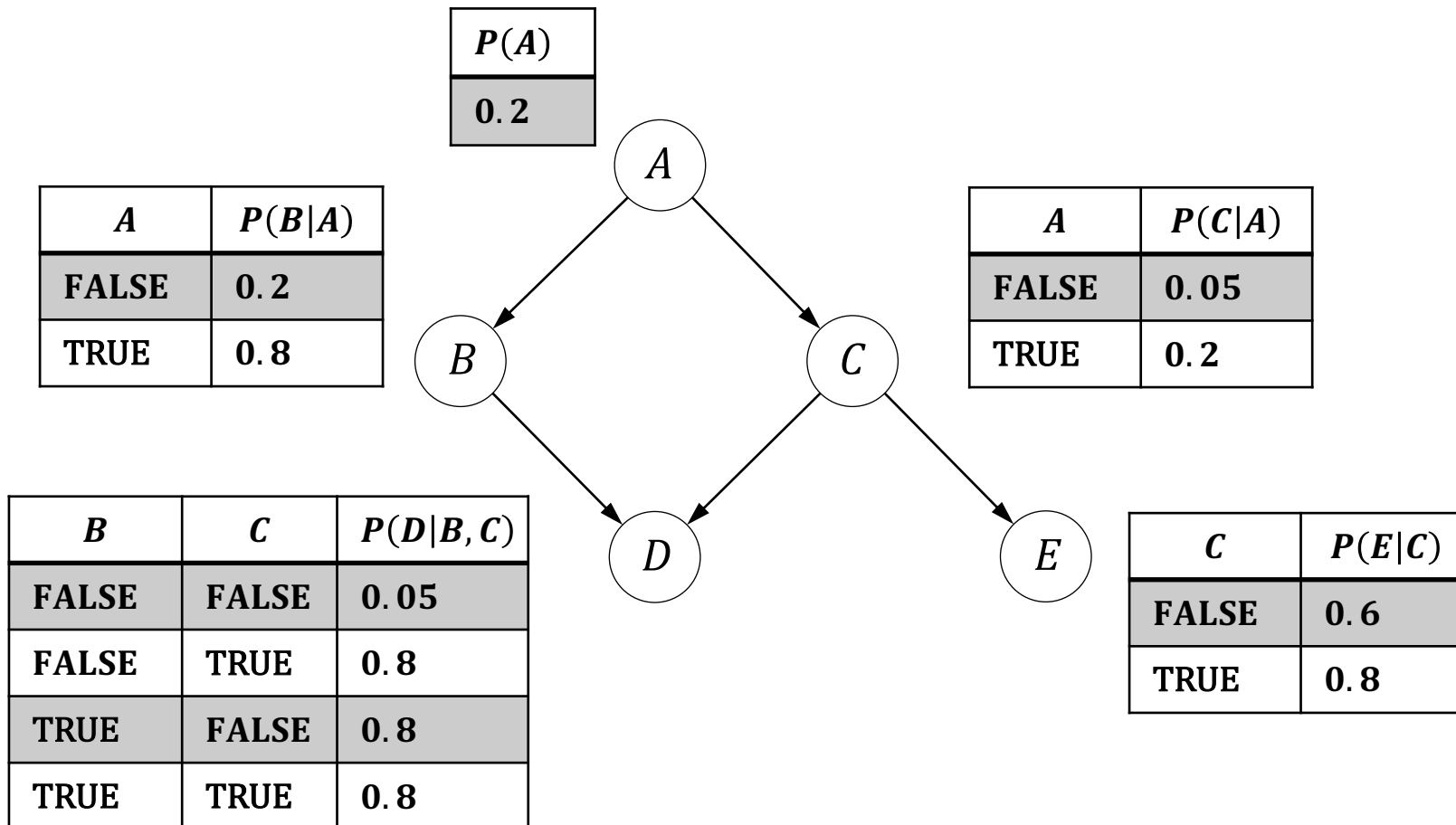
Example (part 1)



Example (part 2)



Example (part 3)



Chain rule in general

By the definition of conditional probability we have:

$$P(X_1, X_2) = P(X_1)P(X_2|X_1)$$

Thus we get this for n variables:

$$P(X_1, X_2, \dots, X_n) = P(X_2, \dots, X_n|X_1)P(X_1) = P(X_n|X_1, \dots, X_{n-1})P(X_1, \dots, X_{n-1})$$

Iterative application of conditional probability

For example:

$$P(X_1, X_2, X_3) = P(X_3|X_1, X_2)P(X_1, X_2) = P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)$$

Chain rule in general

The chain rule exists in $n!$ versions depending on the variable ordering

We can generalize this to any value of n giving us the **Chain Rule**

$$\begin{aligned} P(X_1, X_2, \dots, X_n) = \\ = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3) \cdots P(X_n|X_1, \dots, X_{n-1}) \end{aligned}$$

This factor requires a table with 2^{n-1} rows: one for each truth-value combination of X_2, \dots, X_n . To store it we still need a TB-sized memory when $n = 40$

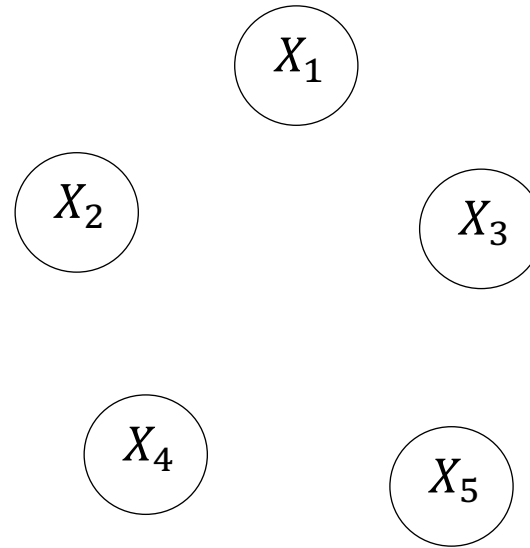
Chain rule for DAGs

This rule will often enable us to represent the table more compactly (as many small tables rather than one big)

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

Here we only need to list the variables that X_i actually depends on. In a Bayesian network, those are the variables in $\text{pa}(X_i)$

Example 1



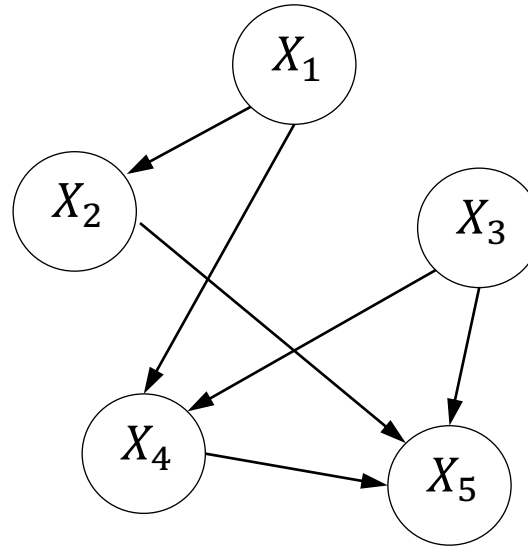
No edges: Independent random variables!

The chain rule for this DAG gives:

$$\begin{aligned} P(X_1, X_2, \dots, X_5) \\ = P(X_1) \dots P(X_5) \end{aligned}$$

We recognize this formula for several independent events

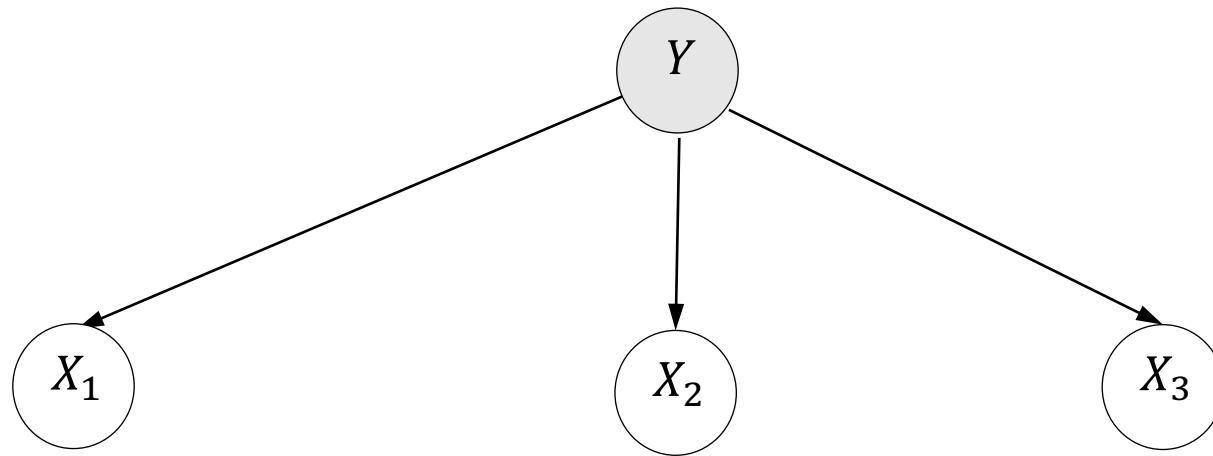
Example 2



The chain rule for this DAG gives:

$$\begin{aligned} & P(X_1, X_2, \dots, X_5) && \text{These terms don't matter} \\ = & P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)P(X_5|X_1, X_2, X_3, X_4) \\ = & P(X_1)P(X_3)P(X_2|X_1)P(X_4|X_1, X_3)P(X_5|X_2, X_3, X_4) \end{aligned}$$

Example 3



The chain rule for this DAG gives:

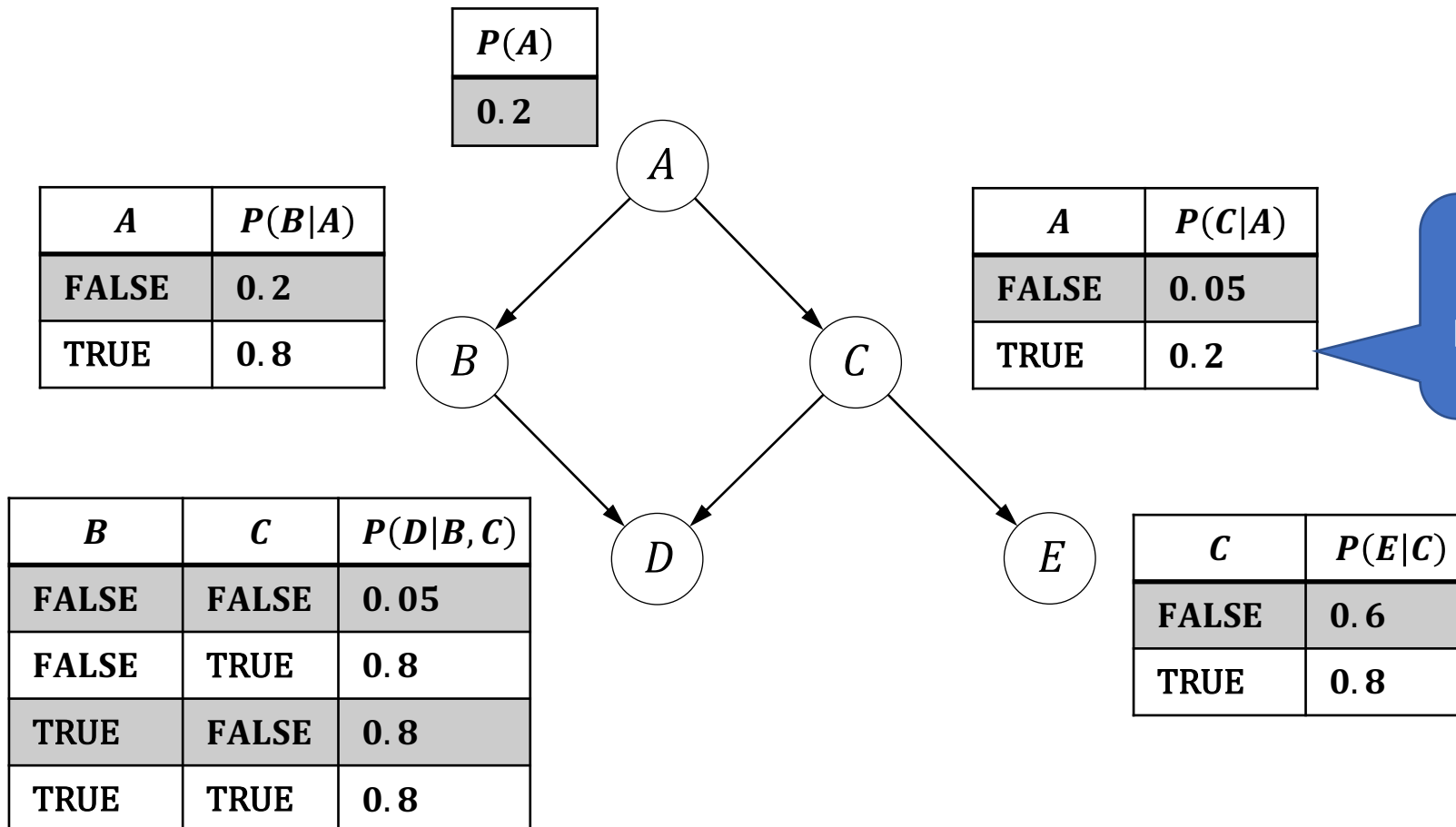
$$\begin{aligned} &P(Y, X_1, X_2, X_3) \\ &= P(Y) * P(X_1|Y) * P(X_2|Y) * P(X_3|Y) \end{aligned}$$

Computing probabilities

Probability of a combination

Let's see how we can compute the probability of any combination, for example $P(A^+, B^-, C^+, D^-, E^+)$.

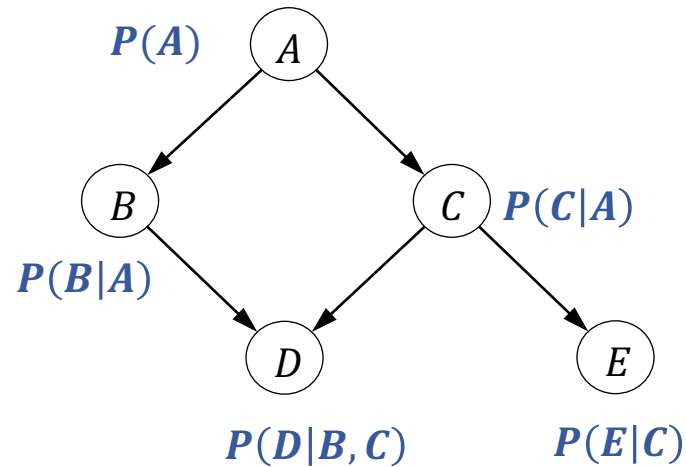
A true, B false, etc.



Here we see that $P(C^+|A^+) = 0.2$. Hence $P(C^-|A^+) = 1 - 0.2 = 0.8$

Probability of a combination

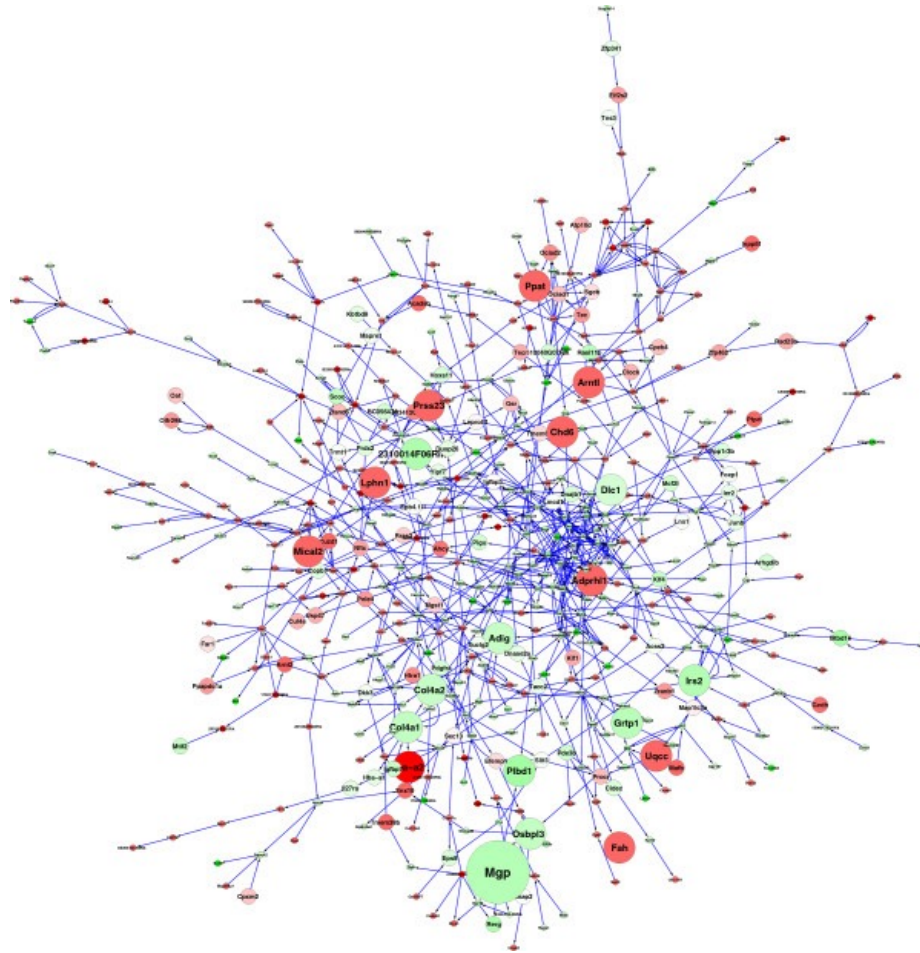
We can compute $P(A^+, B^-, C^+, D^-, E^+)$ by using the chain rule for DAGs.



$$\begin{aligned} P(A^+, B^-, C^+, D^-, E^+) &= \\ &= P(A^+)P(B^-|A^+)P(C^+|A^+)P(D^-|B^-, C^+)P(E^+|C^+) \\ &= P(A^+)(1 - P(B^+|A^+))P(C^+|A^+)(1 - P(D^+|B^-, C^+))P(E^+|C^+) \\ &= \dots = \mathbf{0.00128} \end{aligned}$$

Estimating probabilities

Sampling in Bayesian networks

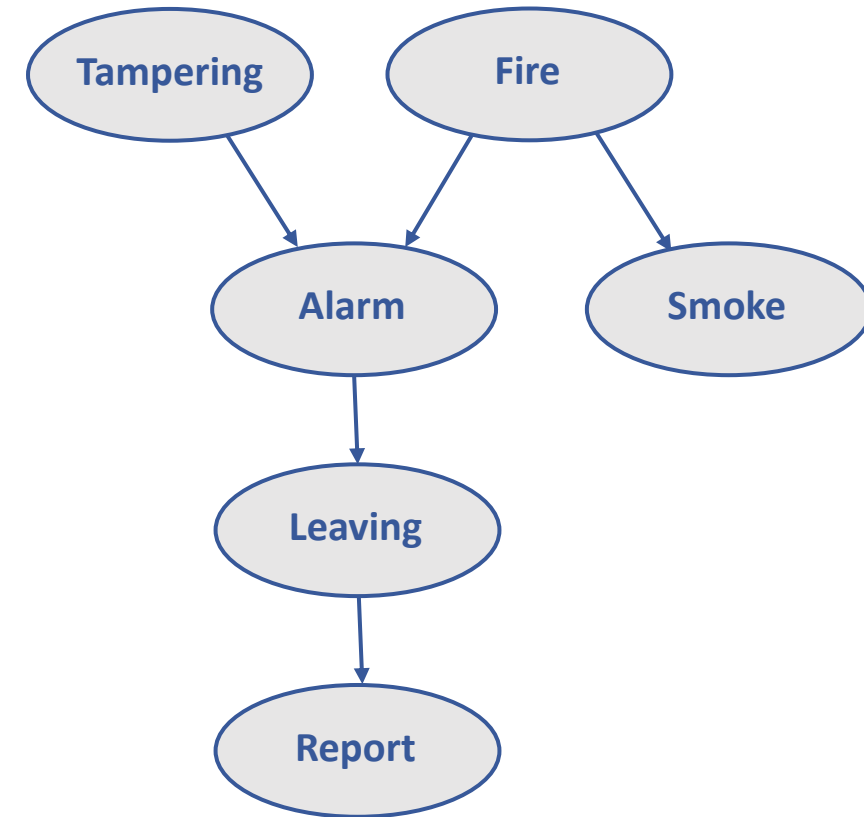


Sampling in Bayesian networks

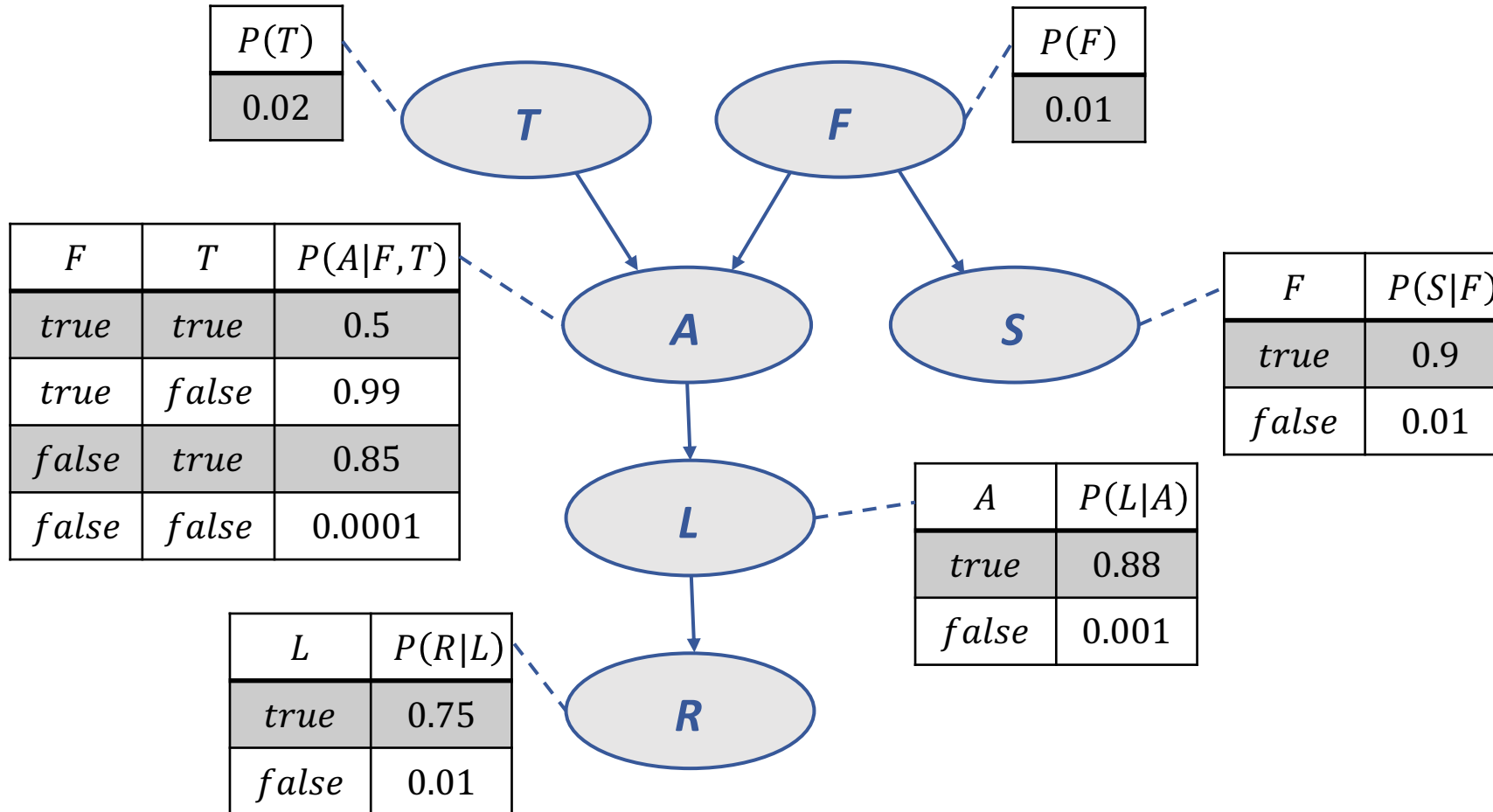
Variables (all true/false):

- *Fire*: true when there is a fire
- *Alarm*: true when the alarm sounds
- *Smoke*: true when there is smoke
- *Leaving*: true if many people leave the building
- *Report*: true if reports of people leaving
- *Tampering*: true when alarm were tampered with

Conditional dependencies are given by the DAG.



Sampling

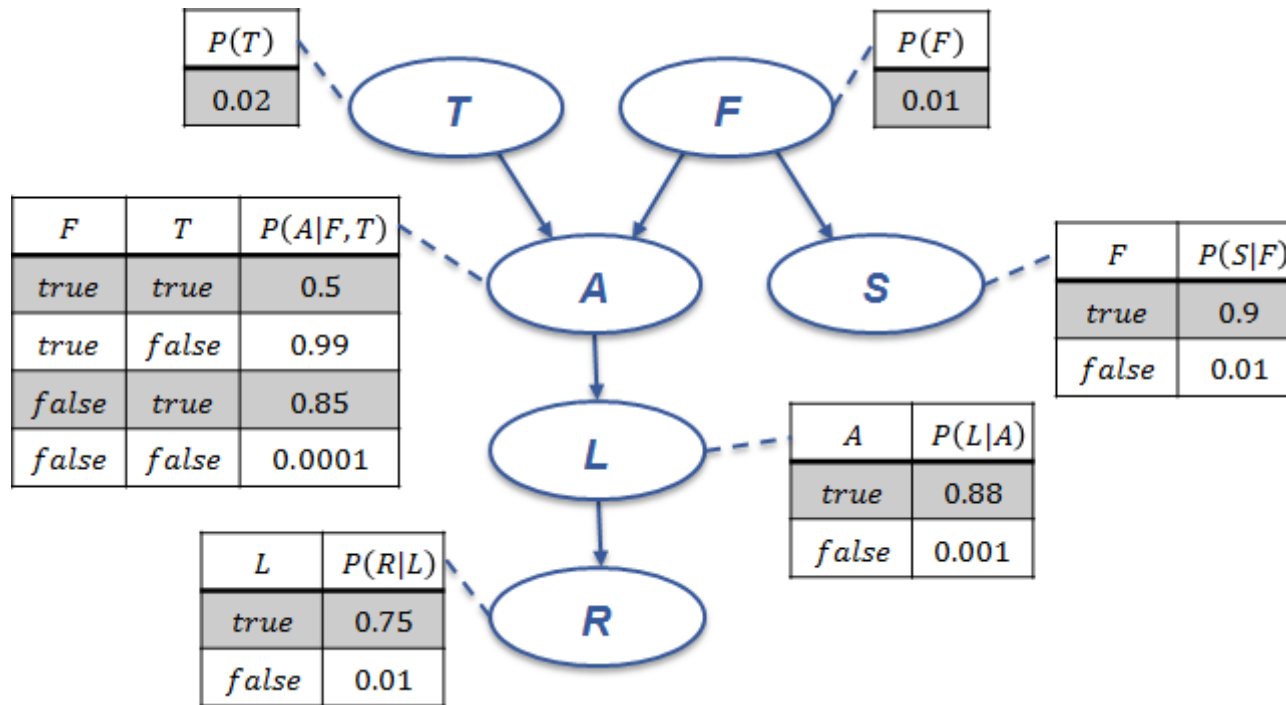


Suppose we have this Bayesian network and want to compute $P(R^+, S^-)$

Then we can compute $P(R^+, S^-, \dots)$ directly for all combinations like we did before and then sum them.

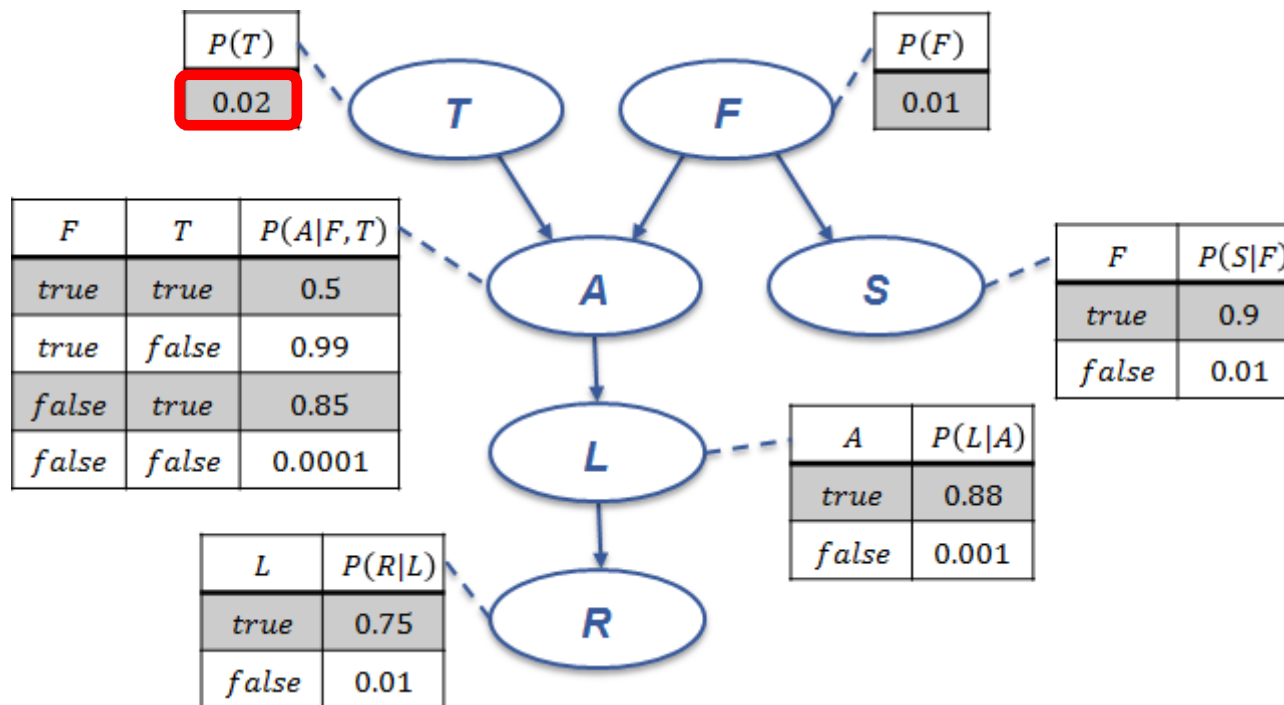
Or we can approximate it using sampling, which scales much better!

Sampling



Example: estimate $P(R^+, S^-)$

Sampling

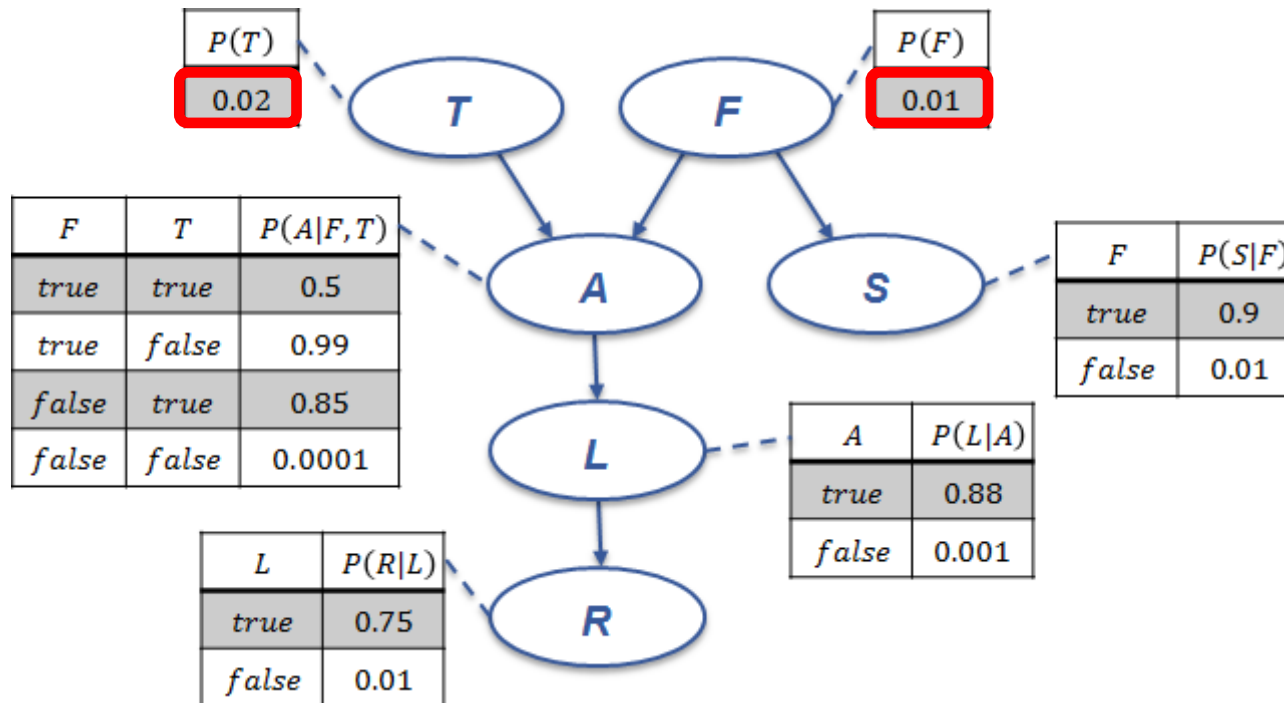


Example: estimate $P(R^+, S^-)$

Generate sample 1:

1. Sample T : e.g $T = false$

Sampling

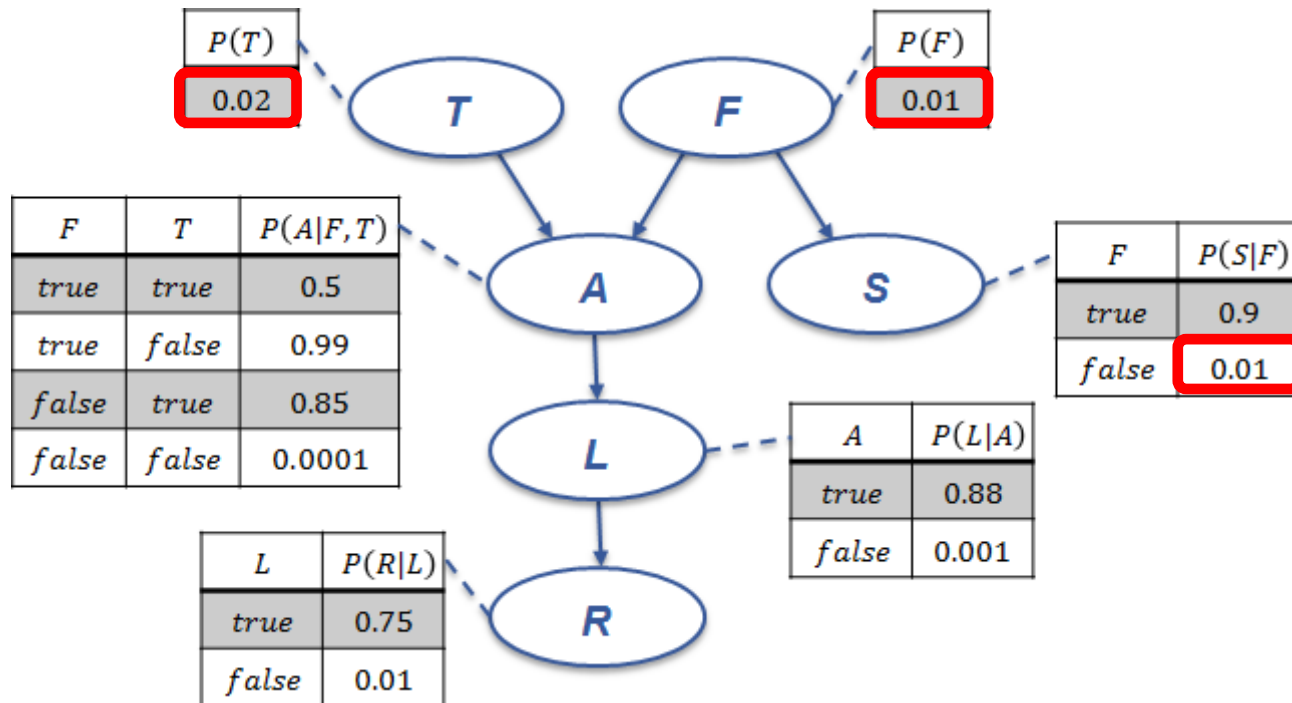


Example: estimate $P(R^+, S^-)$

Generate sample 1:

1. **Sample T :** e.g. $T = \text{false}$
2. **Sample F :** e.g. $F = \text{false}$

Sampling



Example: estimate $P(R^+, S^-)$

Generate sample 1:

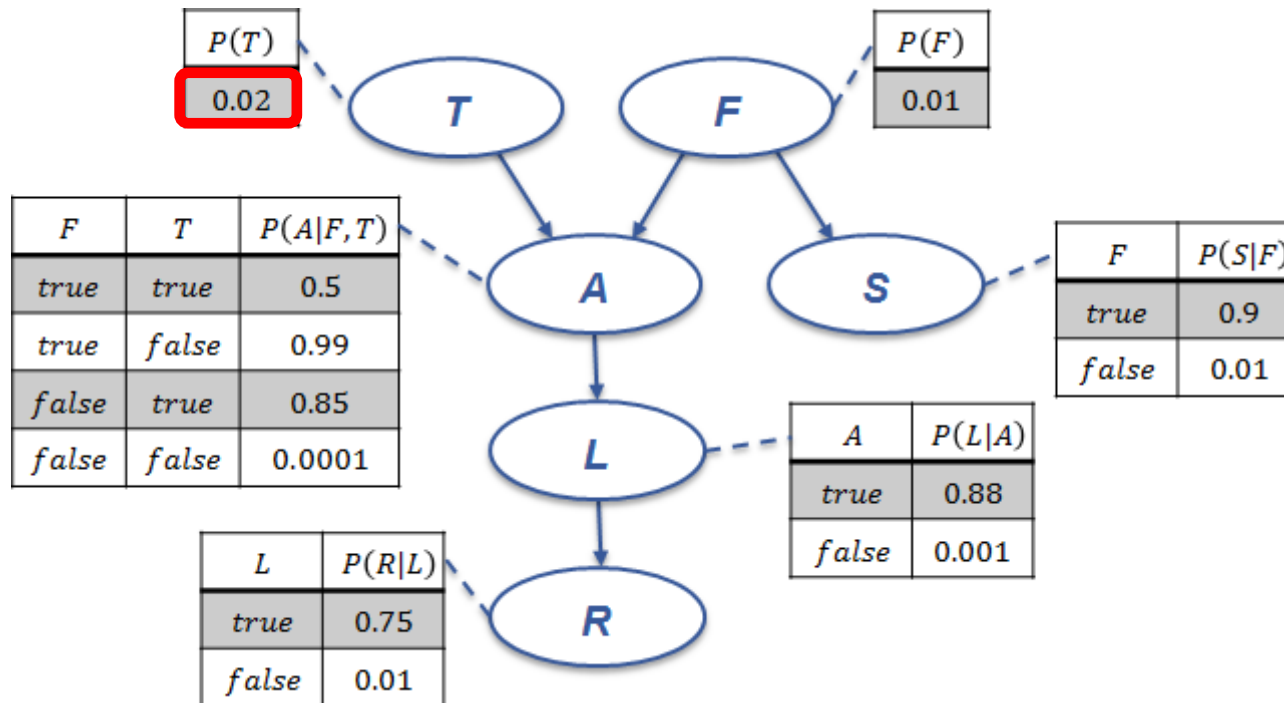
1. Sample T : e.g. $T = false$
2. Sample F : e.g. $F = false$
3. Sample $S|F$: e.g. $S = true$

Score after sample 1:

Hits: 0

Misses: 1

Sampling

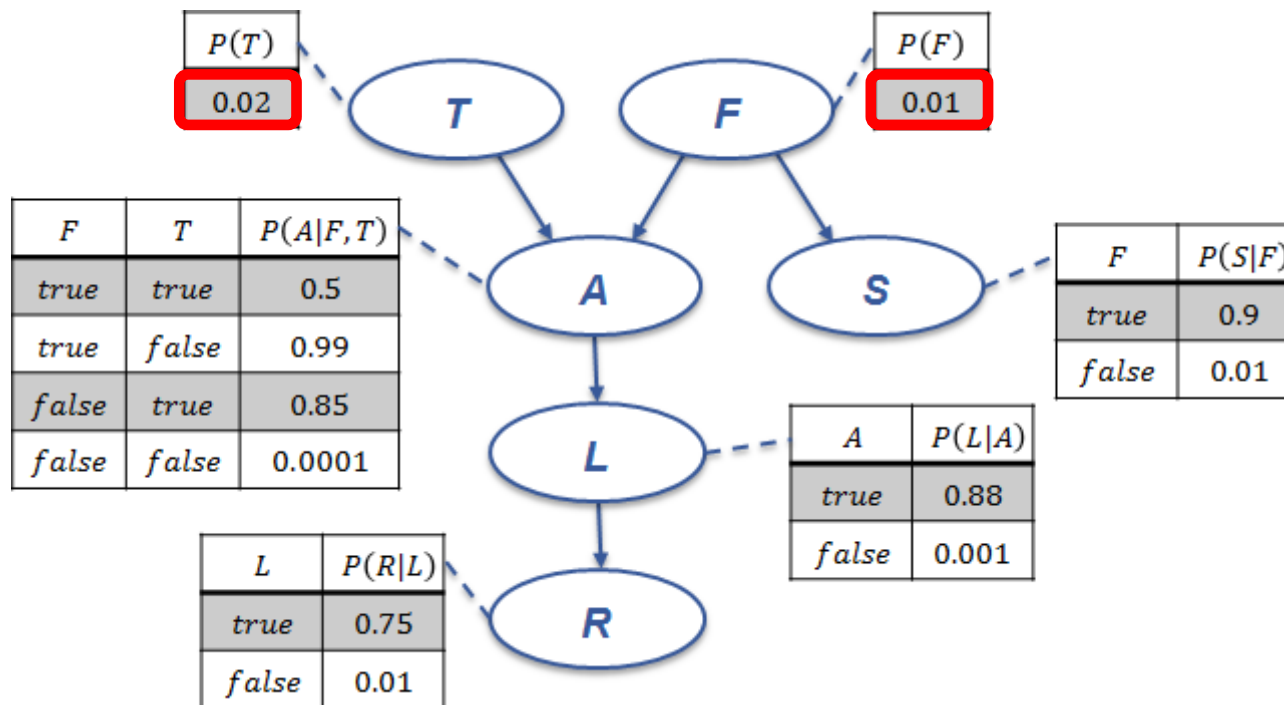


Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. Sample T : e.g $T = false$

Sampling

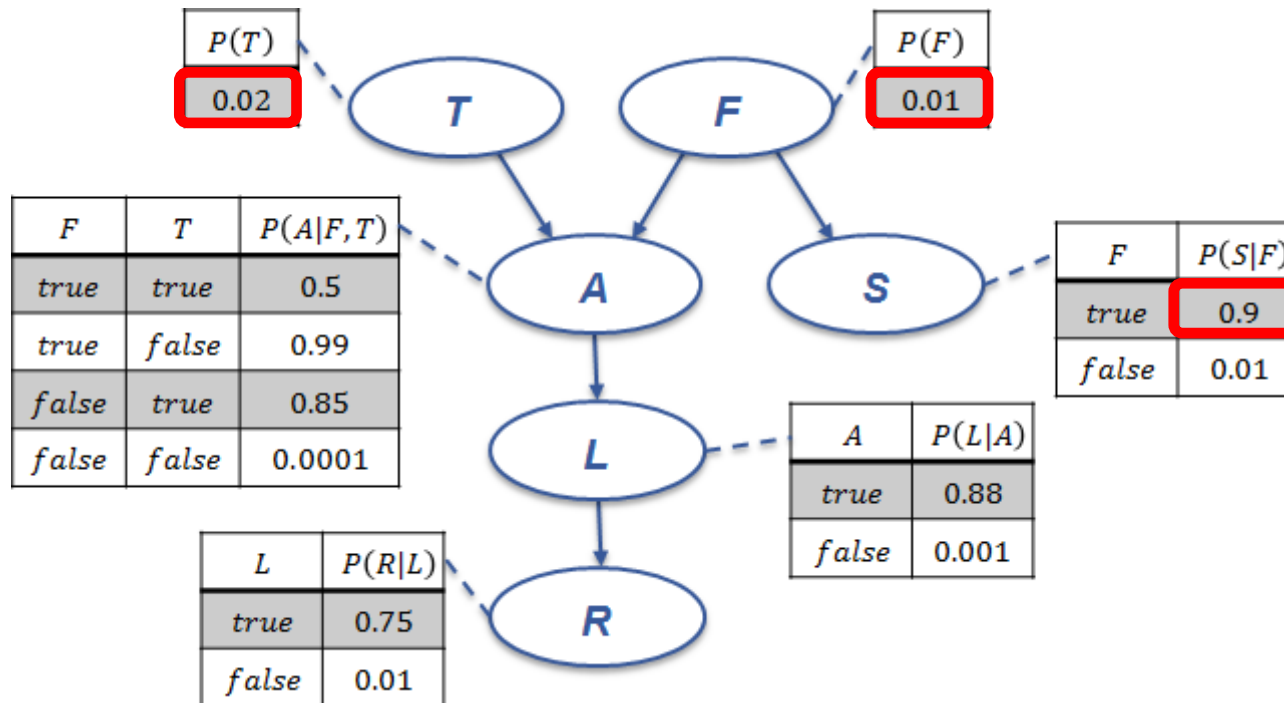


Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. **Sample T :** e.g. $T = \text{false}$
2. **Sample F :** e.g. $F = \text{true}$

Sampling

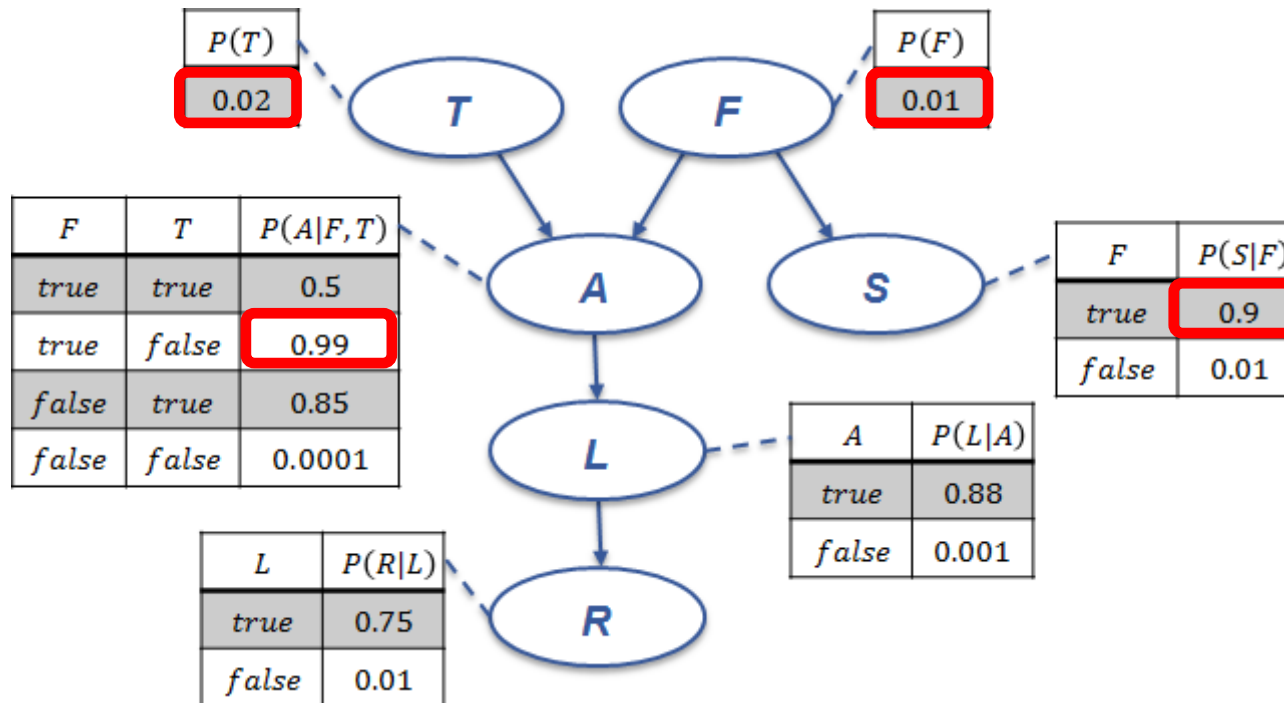


Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. **Sample T :** e.g. $T = \text{false}$
2. **Sample F :** e.g. $F = \text{true}$
3. **Sample $S|F$:** e.g. $S = \text{false}$

Sampling

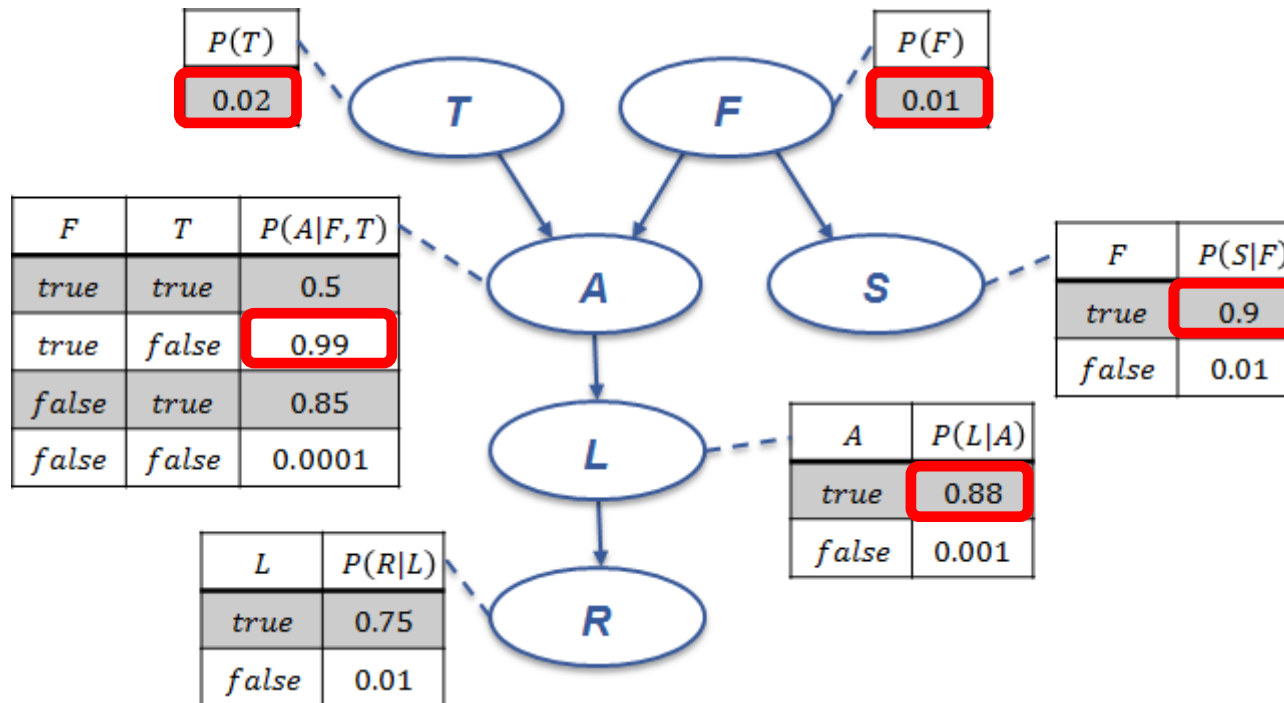


Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. **Sample T :** e.g. $T = false$
2. **Sample F :** e.g. $F = true$
3. **Sample $S|F$:** e.g. $S = false$
4. **Sample $A|F, T^c$:** e.g. $A = true$

Sampling

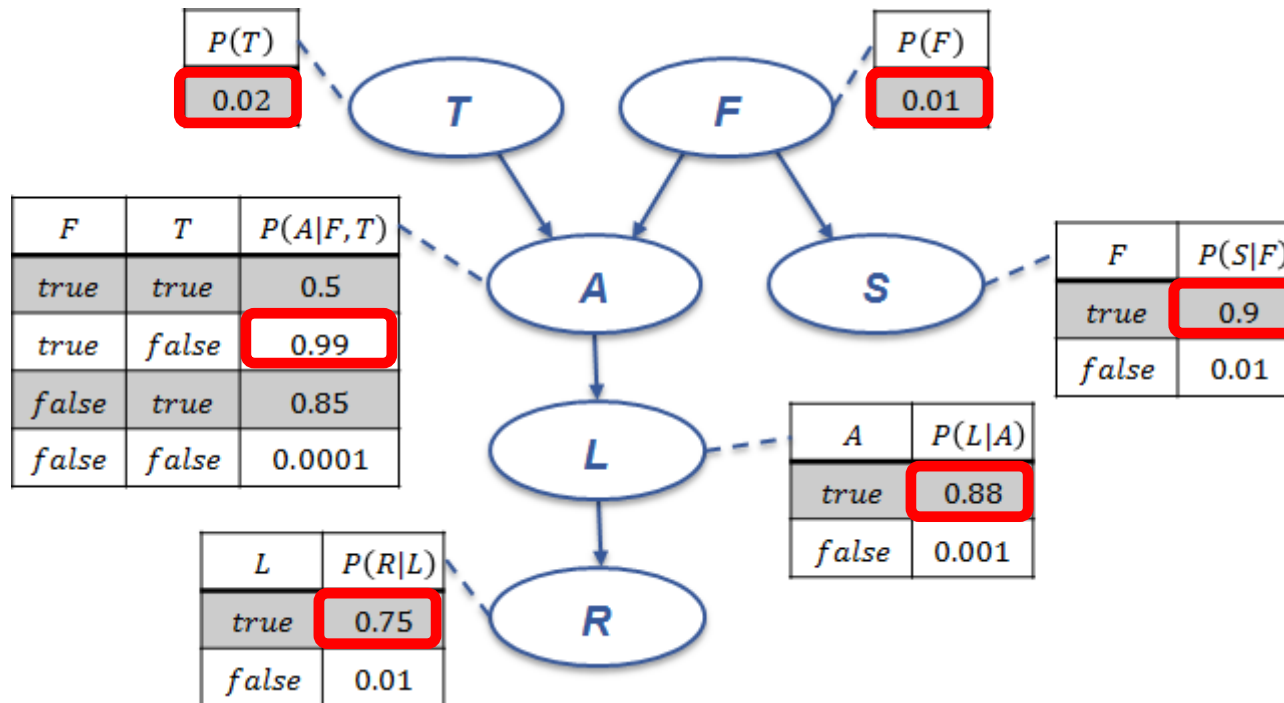


Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. **Sample T : e.g. $T = false$**
2. **Sample F : e.g. $F = true$**
3. **Sample $S|F$: e.g. $S = false$**
4. **Sample $A|F, T^c$: e.g. $A = true$**
5. **Sample $L|A$: e.g. $L = true$**

Sampling



Example: estimate $P(R^+, S^-)$

Generate sample 2:

1. **Sample T :** e.g. $T = false$
2. **Sample F :** e.g. $F = true$
3. **Sample $S|F$:** e.g. $S = false$
4. **Sample $A|F, T^c$:** e.g. $A = true$
5. **Sample $L|A$:** e.g. $L = true$
6. **Sample $R|L$:** e.g. $R = true$

Score after sample 2:

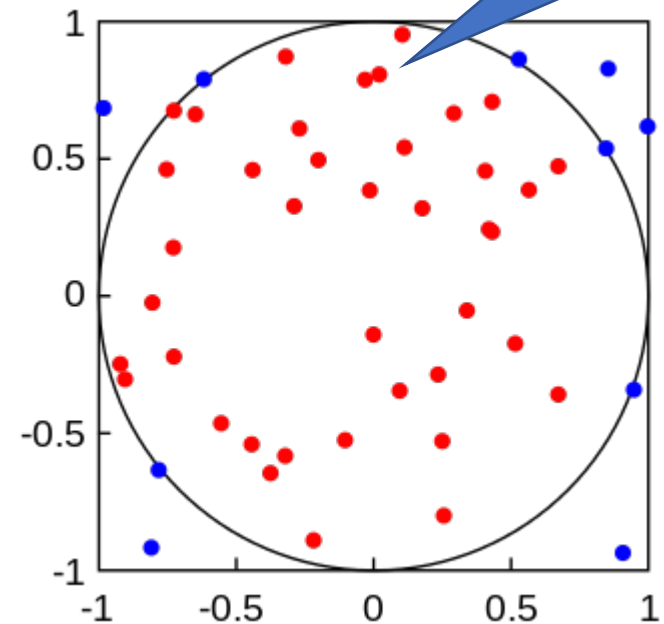
Hits: 1

Misses: 1

Finally compute the result of the sampling process: $P(R, S^c) \approx \frac{\text{Hits}}{\text{Hits} + \text{Misses}}$

Digression: Monte Carlo integration

- Let us approximate π
- Generate 100 random points inside the square ("throw darts")
 - Hits (inside the circle): 80
 - Misses (outside the circle): 20
 - Proportion: 0.8
- circle area $\approx 0.8 * \text{square area}$
 $= 0.8 * 4 = 3.2$
- Also circle area $= \pi * 1 * 1 = \pi$
- So $\pi \approx 3.2$



Applications of Bayesian Networks



<https://data-flair.training/blogs/bayesian-network-applications/>

The Naïve Bayes classifier

Spam or ham?

From: Quantum Method <info@alison1.xyz>
Sent: den 26 januari 2020 21:49
To: Claes Strannegård <claes.strannegard@chalmers.se>
Subject: claes.strannegard, The secret is out. Want to know how to make money?

What is Bitcoin? And should you invest in it?

If you've heard Bitcoin as a buzzword but are at a loss when it comes to the definition, you're not alone.

Bitcoin is a cryptocurrency and worldwide payment system. So basically a Bitcoin is like a Dollar, with the exception that it is rising in value.

The values of Bitcoin, Ripple, and other cryptocurrencies have been crashing lately, but analysts are predicting a huge rise ahead for Bitcoin—with a forecast for it to reach as high as \$100,000 in 2019.

So should you invest? YES. You should!

Spam or ham?

There are several indications of spam in this email!
While reading it we update our belief that it is spam gradually (like in red?). Naïve Bayes tells us how to approximate these probabilities!

0.40

Spam-like sender

0.80

From: Quantum Method <info@alison1.xyz>

Sent: den 26 januari 2020 21:49

To: Claes Strannegård <claes.strannegard@chalmers.se>

0.95

Subject: claes.strannegard, The secret is out. Want to know how to make money?

Spam-like topic

What is Bitcoin? And should you invest in it?

If you've heard Bitcoin as a buzzword but are at a loss when it comes to the definition, you're not alone.

Spam-like word

0.96

Bitcoin is a cryptocurrency and worldwide payment system. So basically a Bitcoin is like a Dollar, with the exception that it is rising in value.

The values of Bitcoin, Ripple, and other cryptocurrencies have been crashing lately, but analysts are predicting a huge rise ahead for Bitcoin—with a forecast for it to reach as high as \$100,000 in 2019.

0.97

So should you invest? YES. You should!

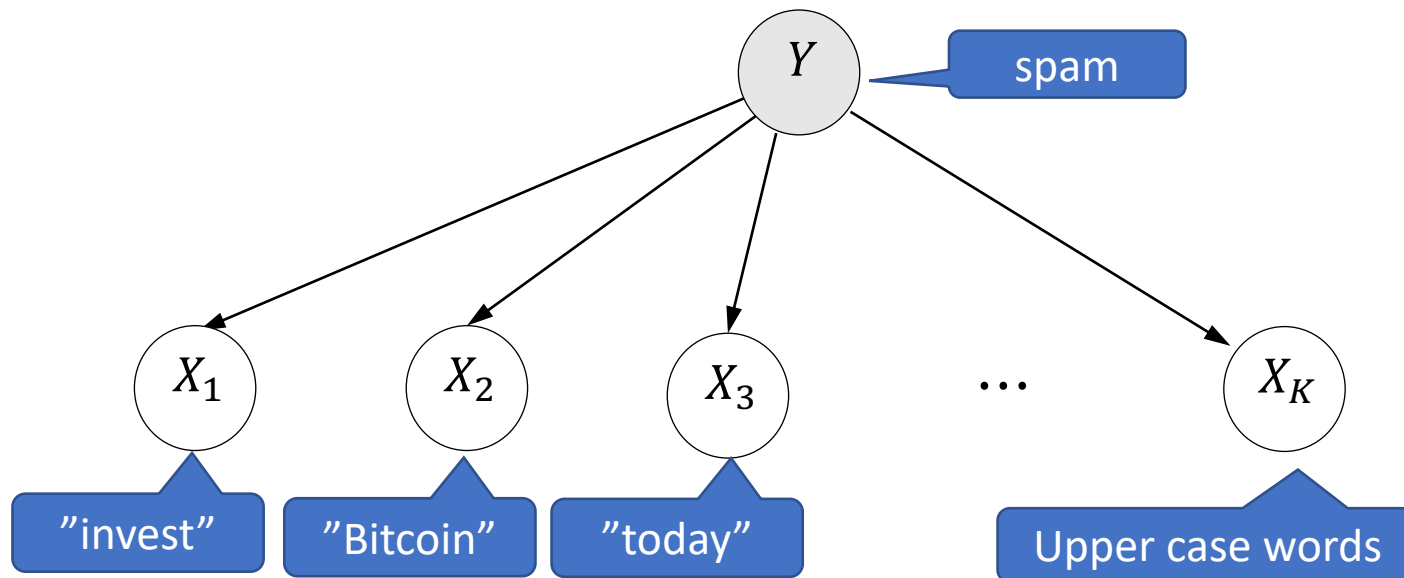
0.98

Spam-like word

Spam-like case

A spam-like word is a word that is more common in spam than in ham!

Simplifying assumption

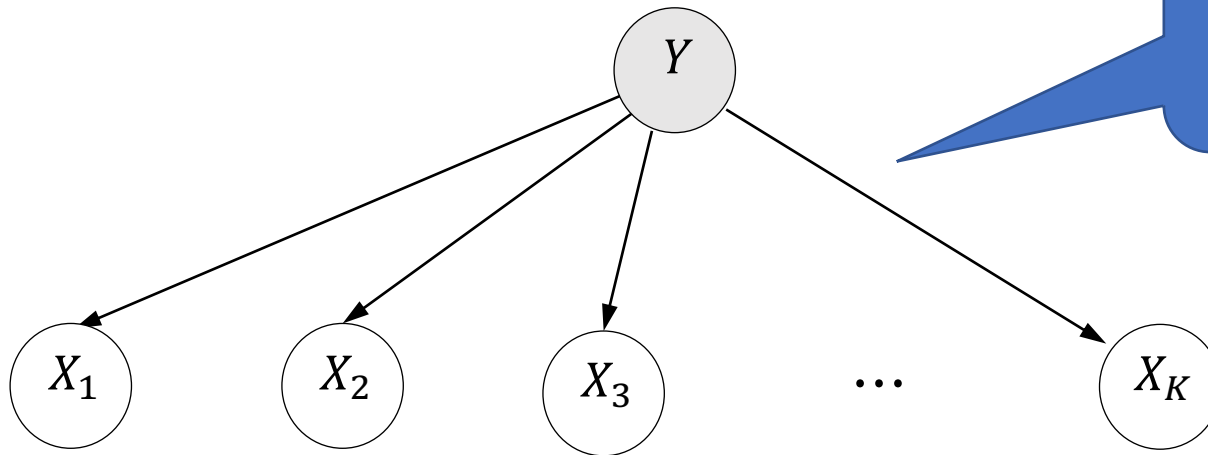


Y influence all the X_i

The X_i do not depend on each other

Naïve Bayes assumption

- Assume that the graph looks like this:



Sometimes the assumption that the X_i do not depend on each other is correct, sometimes not. Then it is "naive" to assume that there are no other dependencies.

Still, this assumption can be helpful also in many situations when it is not 100% correct!

Naïve Bayes formula

- Note that $P(Y|X_1, \dots, X_K) = \frac{P(Y, X_1, \dots, X_K)}{P(X_1, \dots, X_K)}$ by the definition of conditional probability
- The nominator: $P(Y, X_1, \dots, X_K) = P(Y) * \prod_{k=1}^K P(X_k|Y)$ by the graph
- The denominator: $P(X_1, \dots, X_K) = \prod_{k=1}^K P(X_k)$ by the graph
- Hence we get the Posterior

$$P(Y|X_1, \dots, X_K) = \frac{P(Y) \cdot \prod_{k=1}^K P(X_k|Y)}{\prod_{k=1}^K P(X_k)}$$

Note that Bayes' rule is the case $K=1$. This formula allows us to update our beliefs based on several observations rather than one.

Can be used for computing the most likely label given the data. So it can be used as a classifier.

The formula is relatively simple because of the Naïve Bayes assumption!

These numbers are relatively easy to find!
This is why Naïve Bayes is so useful.

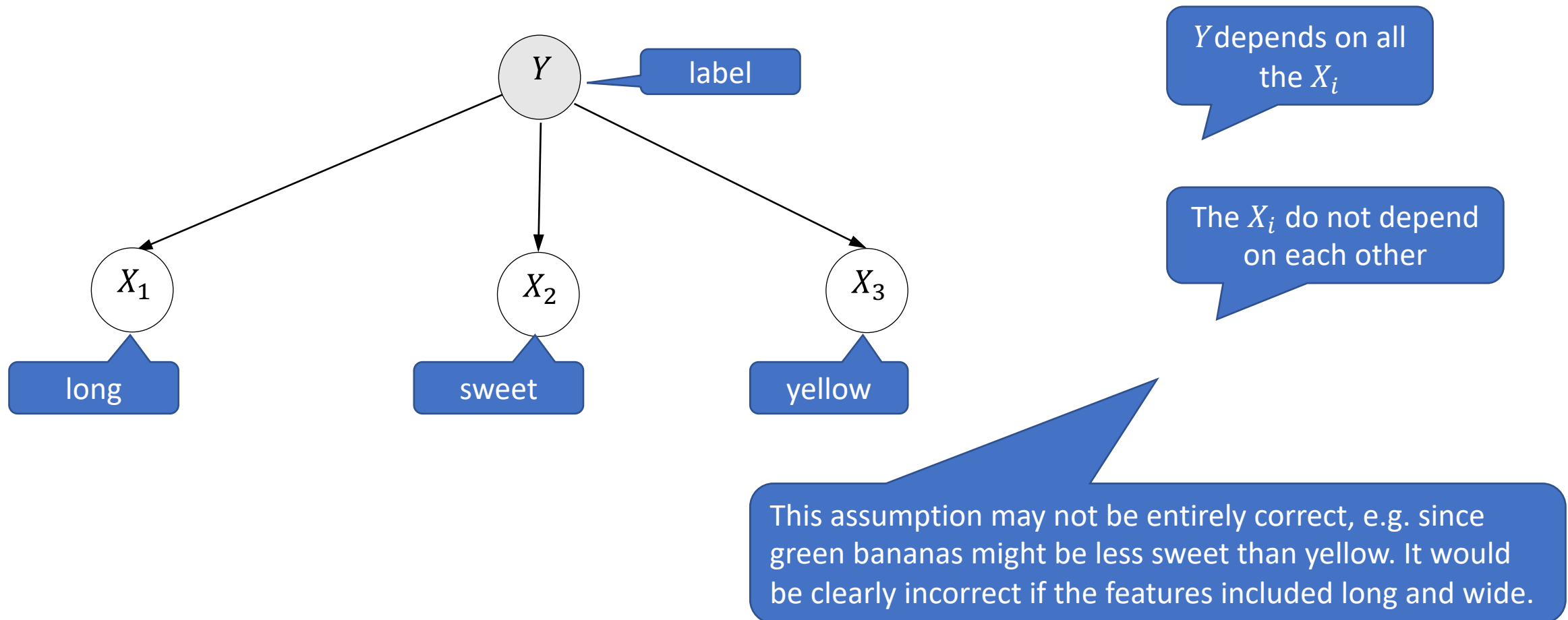
Fruit example

Naïve Bayes: an example

- We have $N = 1000$ fruits with class labels
 - Banana,
 - Orange, or
 - Other
- Three features for each fruit
 - Long
 - Sweet
 - Yellow
- Objective: build a classifier that predicts the class label for a given fruit when only the three features are known



Naïve Bayes assumption



Naïve Bayes: an example

Step 1: Compute the prior probabilities $P(Y)$ for each fruit label

$$P(Y = \text{banana}) = 500/1000 = 0.5$$

$$P(Y = \text{orange}) = 300/1000 = 0.3$$

$$P(Y = \text{other}) = 200/1000 = 0.2$$

Label	Total
Banana	500
Orange	300
Other	200
Total	1000

Naïve Bayes: an example

$$P(Y/X_1, X_2, X_3) = \frac{\prod_{k=1}^3 P(X_k|Y) * P(Y)}{\prod_{k=1}^3 P(X_k)}$$

Step 2: Compute the denominator

$$\prod_{k=1}^3 P(X_k)$$

$$P(X_1 = \text{long}) = 500/1000 = 0.5$$

$$P(X_2 = \text{sweet}) = 650/1000 = 0.65$$

$$P(X_3 = \text{yellow}) = 800/1000 = 0.8$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	500	650	800	1000

Naïve Bayes: an example

Step 3: Compute the likelihood

$$\prod_{k=1}^3 P(X_k|Y) = \prod_{k=1}^3 \frac{\#\{\text{fruits with label } Y \text{ and feature } X_k\}}{\#\{\text{fruits with label } Y\}}$$

$$P(X_1 = \text{long}|\text{banana}) = 400/500 = 0.8$$

$$P(X_2 = \text{sweet}|\text{banana}) = 350/500 = 0.7$$

$$P(X_3 = \text{yellow}|\text{banana}) = 450/500 = 0.9$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500

Naïve Bayes: computing the probabilities

Now we have all the numbers needed for applying the Naïve Bayes formula:

$$P(\text{banana}|\text{long, sweet, yellow})$$

$$= \frac{P(\text{banana})P(\text{long}|\text{banana})P(\text{sweet}|\text{banana})P(\text{yellow}|\text{banana})}{P(\text{long})P(\text{sweet})P(\text{yellow})}$$

$$= \frac{0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9}{0.5 \cdot 0.65 \cdot 0.8} = 0.969$$



Naïve Bayes: finding the most likely label

- Given some features X_1, \dots, X_K , which label is most likely?
- Which is biggest?
 - $P(\textit{banana}|X_1, \dots, X_K)$?
 - $P(\textit{orange}|X_1, \dots, X_K)$?
 - $P(\textit{other}|X_1, \dots, X_K)$?
- To find out, it is enough to compute only the nominator:

$$P(Y|X_1, \dots, X_K) = \frac{P(Y) \cdot \prod_{k=1}^K P(X_k|Y)}{\prod_{k=1}^K P(X_k)}$$

No need to compute the denominator when comparing!

Now we are ready for making the comparison

Naïve Bayes: an example

Step 4: Given that the fruit is long, sweet, and yellow, what is the *most likely label*?

$$P(\text{banana} | \text{long, sweet, yellow})$$

$$\propto P(\text{banana})P(\text{long} | \text{banana})P(\text{sweet} | \text{banana})P(\text{yellow} | \text{banana})$$

$$= 0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9 = 0.252$$

$$P(\text{orange} | \text{long, sweet, yellow}) \propto 0 \text{ because } P(\text{long} | \text{orange}) = 0$$

$$P(\text{other} | \text{long, sweet, yellow}) \propto 0.01875$$

Conclusion: the fruit is most likely a banana!



Applications of Naïve Bayes

- Real-time Prediction (fast, scalable)
- Multi-class Prediction
- Text classification/ Spam Filtering/ Sentiment Analysis
- Recommendation Systems

Assignment 4 :

Naïve Bayes Classifier for Spam Email

- You will download a real dataset of spam/ham emails.
- **Vocabulary:** (subset of) words occurring in training data. Can be many!
- **Features:** words occurring in emails vocabulary.
- Different types of implementation of Naïve Bayes in Sci-kit learn library:
 - Bernoulli: Assumes binary features
 - Is the word “invest” present, Yes/No?
 - Does the word “invest” occur more than twice, Yes/No?
 - Multinomial: Features are discrete counts (frequencies)
 - i.e. how many times does the word invest occur in the text?
 - <https://hub.packtpub.com/implementing-3-naive-bayes-classifiers-in-scikit-learn/>