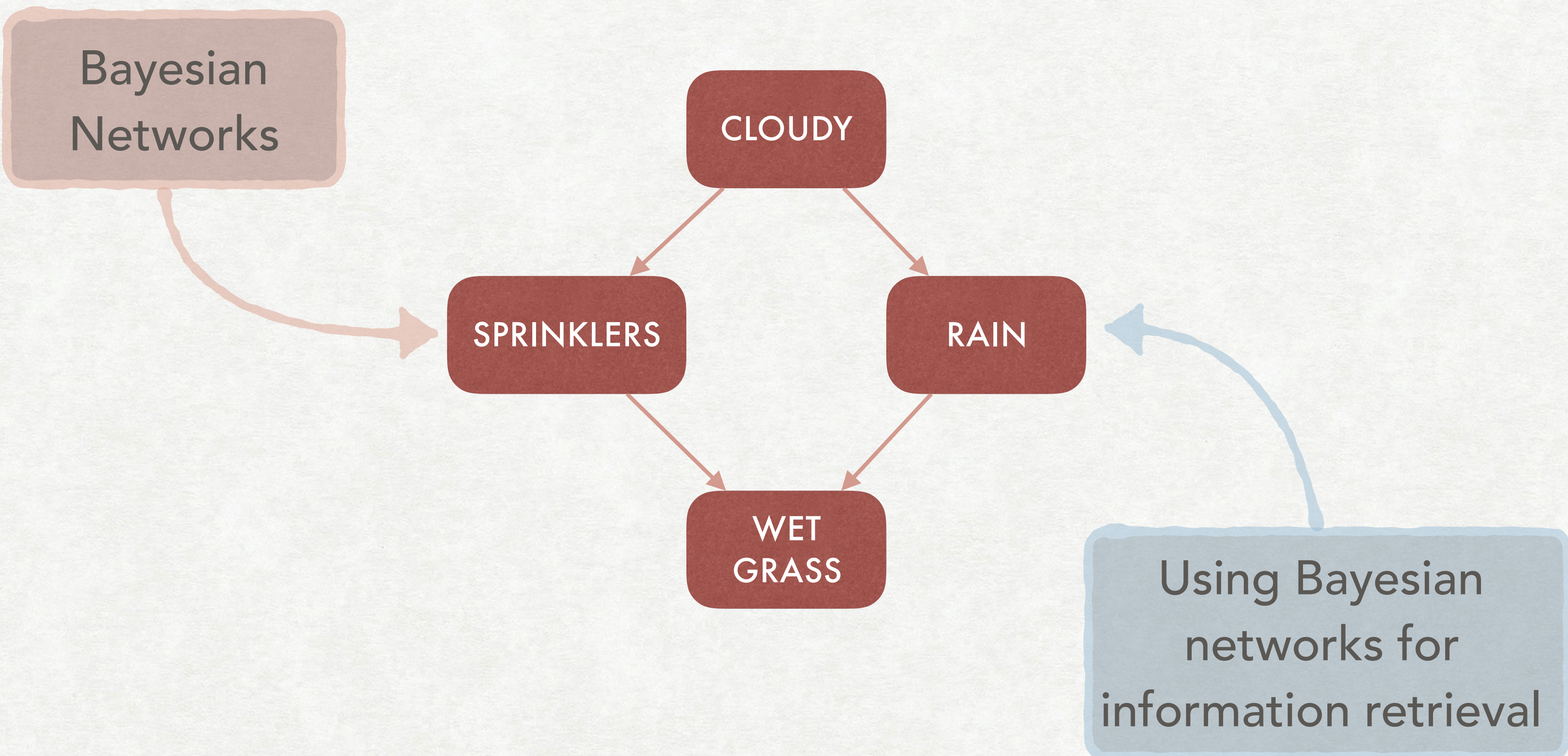
INFORMATION RETRIEVAL

Luca Manzoni
lmanzoni@units.it

Lecture 8

LECTURE OUTLINE

*SUBTITLE INTENTIONALLY LEFT BLANK



BAYESIAN NETWORKS

BAYESIAN NETWORKS

WHAT ARE THEM

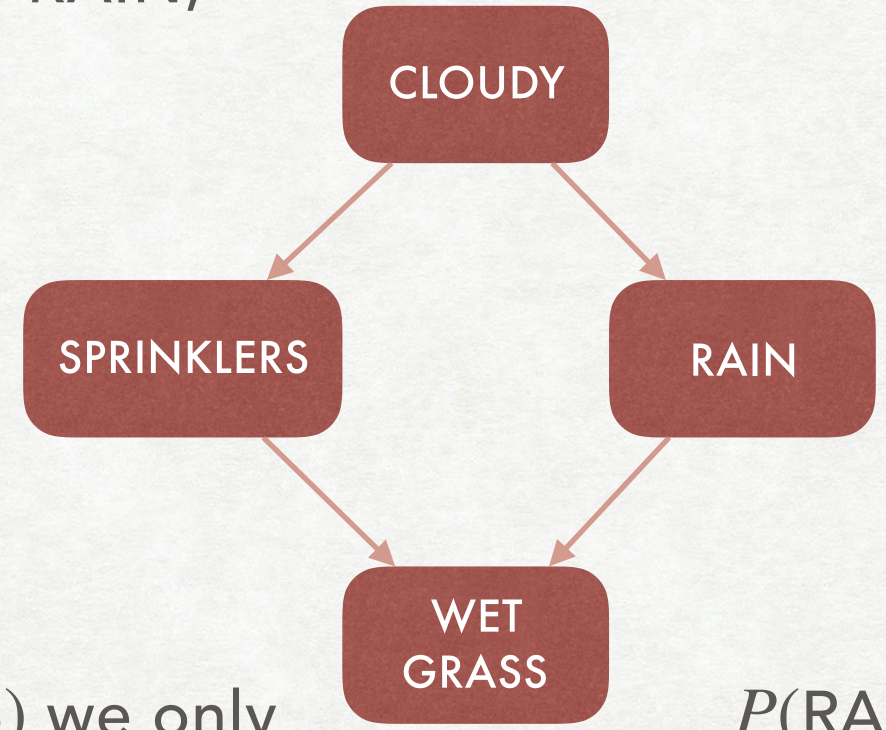
- Also called **Bayesian belief networks, decision network**, etc.
- A **graphical model** is a statistical model using a graph to represent the conditional dependency between random variables.
- BN are a kind **graphical model** using a directed acyclic graph.
- Intuitively they are useful because when we need to compute $P(y \mid x_1, x_2, \ldots, x_k)$ we actually need to compute only $p(y \mid \mathrm{Pa}(y))$ with $\mathrm{Pa}(y)$ the parent nodes of y.
- An example should clarify this.

BAYESIAN NETWORKS

A SIMPLE EXAMPLE

There are four random variables: CLOUDY, SPRINKLERS, RAIN, and WET GRASS.

The edges represents the conditional dependencies



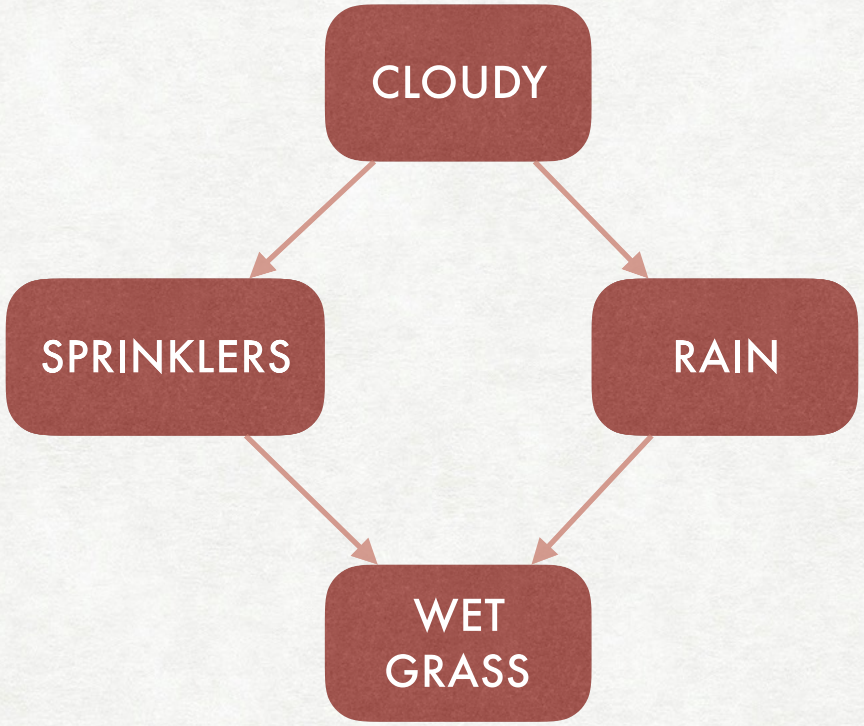
If we want to compute $P(\text{CLOUDY}|\text{SPRINKLES})$ we only compute $P(\text{SPRINKLES}|\text{CLOUDY})$, and we will have to “rewrite” it.

If we want to compute $P(\text{RAIN}|\text{CLOUDY})$ we can find it directly in our table

BAYESIAN NETWORKS

A SIMPLE EXAMPLE

C = 0	C =1
0.5	0.5



	S = 0	S =1
C = 0	0.5	0.5
C = 1	0.9	0.1

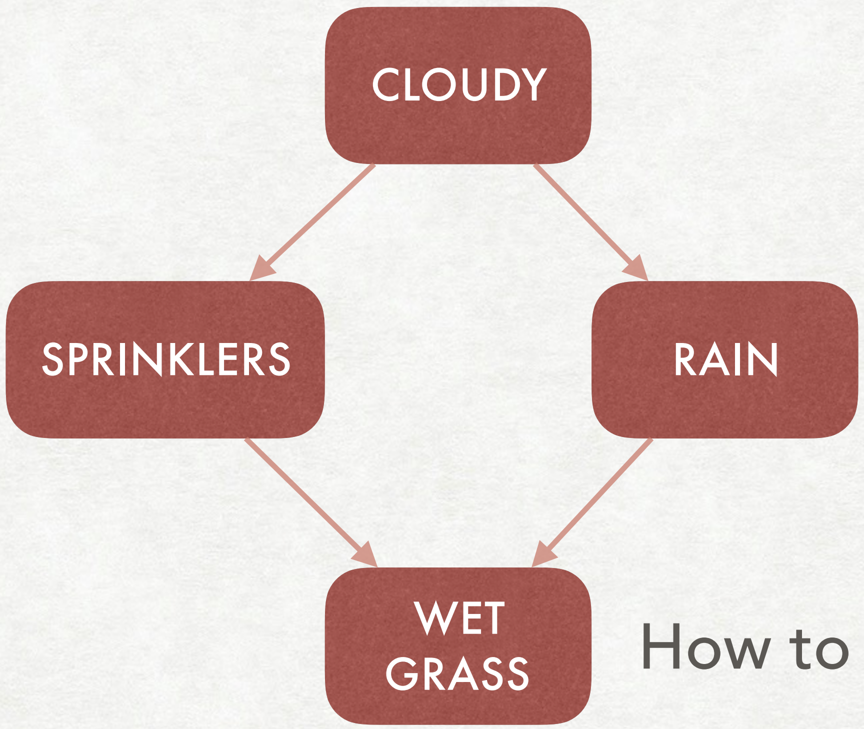
	R = 0	R =1
C = 0	0.8	0.2
C = 1	0.2	0.8

		W = 0	W =1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99

BAYESIAN NETWORKS

A SIMPLE EXAMPLE

C = 0	C =1
0.5	0.5



	S = 0	S =1
C = 0	0.5	0.5
C = 1	0.9	0.1

	R = 0	R =1
C = 0	0.8	0.2
C = 1	0.2	0.8

How to find $P(\mathrm{W} = 1 \,|\, \mathrm{S} = 1, \mathrm{R} = 0)$?

		W = 0	W =1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99

BAYESIAN NETWORKS

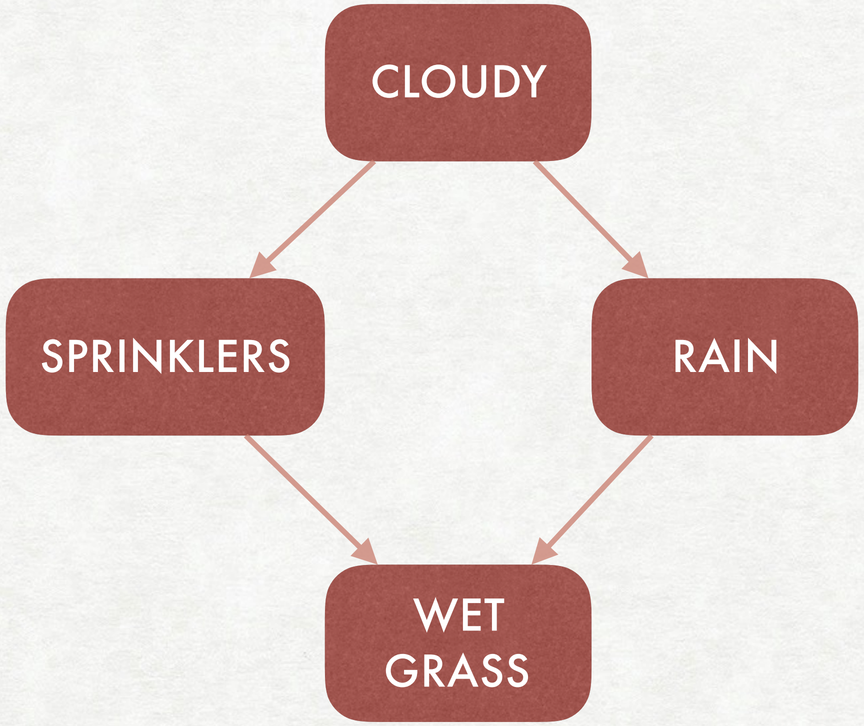
A SIMPLE EXAMPLE

C = 0	C =1
0.5	0.5

	S = 0	S =1
C = 0	0.5	0.5
C = 1	0.9	0.1

	R = 0	R =1
C = 0	0.8	0.2
C = 1	0.2	0.8

		W = 0	W =1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99



How to find

$P(\mathrm{W}=1\,|\,\mathrm{C}=1,\mathrm{R}=0)$?

$$P(\mathrm{W}=1\,|\,\mathrm{C}=1,\mathrm{R}=0)$$

$$= P(\mathrm{W}=1\,|\,\mathrm{R}=0,\mathrm{S}=1)\cdot P(\mathrm{S}=1\,|\,\mathrm{C}=1)$$
$$+P(\mathrm{W}=1\,|\,\mathrm{R}=0,\mathrm{S}=0)\cdot P(\mathrm{S}=0\,|\,\mathrm{C}=1)$$

$$= 0.9\cdot 0.1 + 0\cdot 0.9$$

$$= 0.09$$

BAYESIAN NETWORKS

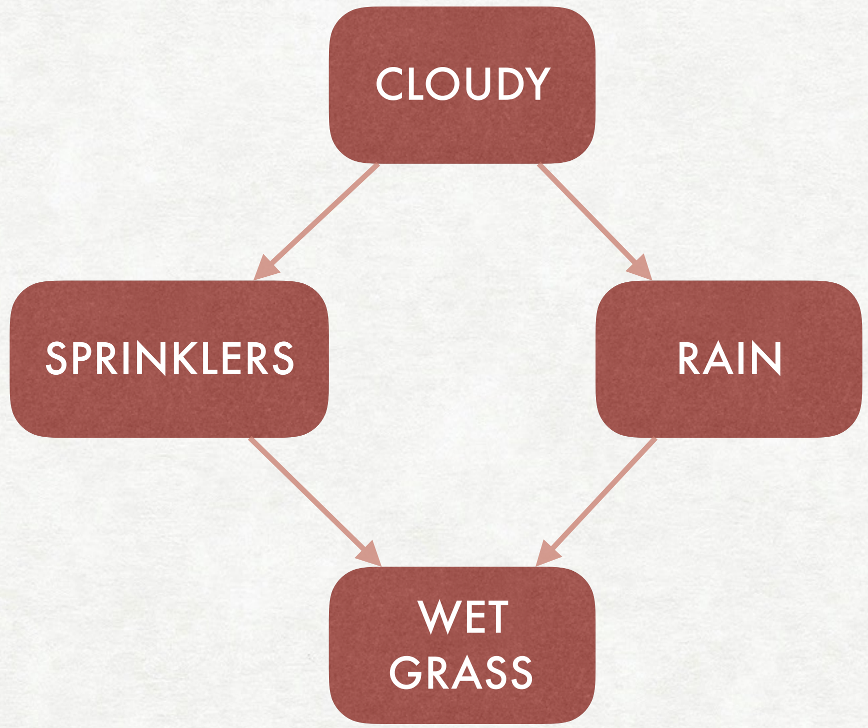
A SIMPLE EXAMPLE

C = 0	C = 1
0.5	0.5

	S = 0	S = 1
C = 0	0.5	0.5
C = 1	0.9	0.1

	R = 0	R = 1
C = 0	0.8	0.2
C = 1	0.2	0.8

		W = 0	W = 1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99



How to find

$$P(S = 1 \mid C = 1, W = 1)?$$

$$P(S = 1 \mid C = 1, W = 1)$$

$$= \frac{P(W = 1 \mid C = 1, S = 1)}{P(W = 1 \mid C = 1)} \cdot P(S = 1 \mid C = 1)$$

$$= \frac{P(W = 1 \mid C = 1, S = 1)}{P(W = 1 \mid C = 1)} \cdot 0.1$$

$$P(W = 1 \mid C = 1, S = 1) \qquad P(W = 1 \mid C = 1)$$

BAYESIAN NETWORKS

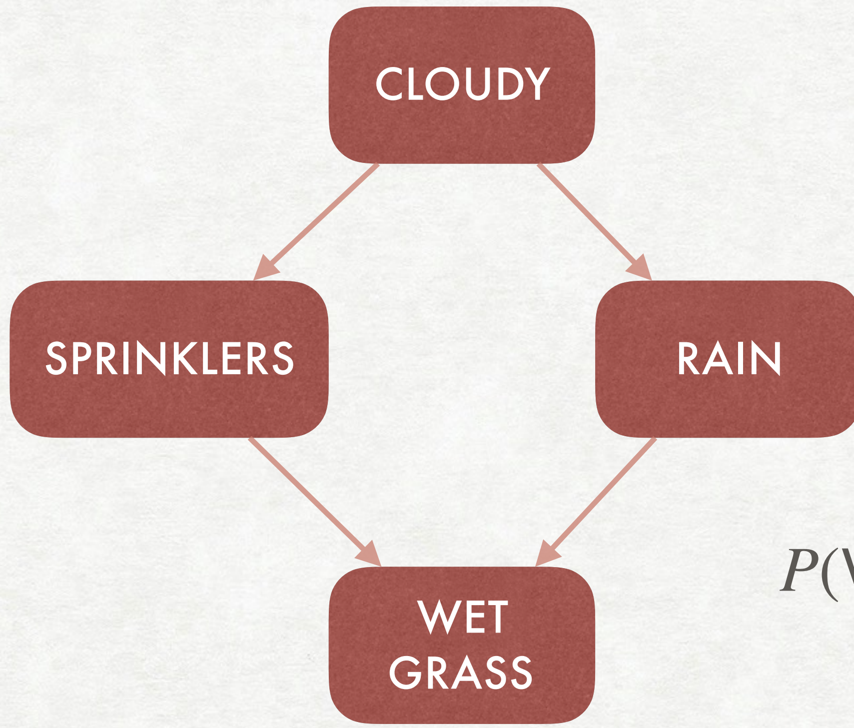
A SIMPLE EXAMPLE

C = 0	C =1
0.5	0.5

	S = 0	S =1
C = 0	0.5	0.5
C = 1	0.9	0.1

	R = 0	R =1
C = 0	0.8	0.2
C = 1	0.2	0.8

		W = 0	W =1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99



$$P(\mathrm{W} = 1 \mid \mathrm{C} = 1, \mathrm{S} = 1) = 0.0972$$

$$P(\mathrm{W} = 1 \mid \mathrm{C} = 1)$$

$$\begin{aligned}
&P(\mathrm{W} = 1 \mid \mathrm{S} = 0, \mathrm{R} = 0) \cdot P(\mathrm{S} = 0 \mid \mathrm{C} = 1) \cdot P(\mathrm{R} = 0 \mid \mathrm{C} = 1) + \\
&P(\mathrm{W} = 1 \mid \mathrm{S} = 0, \mathrm{R} = 1) \cdot P(\mathrm{S} = 0 \mid \mathrm{C} = 1) \cdot P(\mathrm{R} = 1 \mid \mathrm{C} = 1) + \\
&P(\mathrm{W} = 1 \mid \mathrm{S} = 1, \mathrm{R} = 0) \cdot P(\mathrm{S} = 1 \mid \mathrm{C} = 1) \cdot P(\mathrm{R} = 0 \mid \mathrm{C} = 1) + \\
&P(\mathrm{W} = 1 \mid \mathrm{S} = 1, \mathrm{R} = 1) \cdot P(\mathrm{S} = 1 \mid \mathrm{C} = 1) \cdot P(\mathrm{R} = 1 \mid \mathrm{C} = 1)
\end{aligned}$$

$$0 \cdot 0.9 \cdot 0.2 + 0.9 \cdot 0.9 \cdot 0.8 + 0.9 \cdot 0.1 \cdot 0.2 + 0.99 \cdot 0.1 \cdot 0.8 = 0.7452$$

BAYESIAN NETWORKS

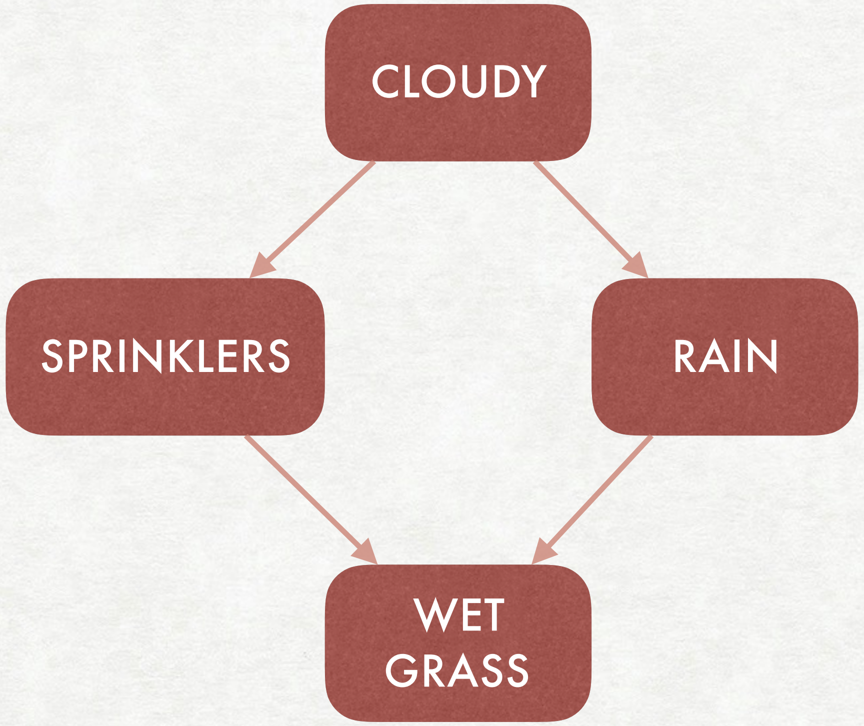
A SIMPLE EXAMPLE

C = 0	C =1
0.5	0.5

	S = 0	S =1
C = 0	0.5	0.5
C = 1	0.9	0.1

	R = 0	R =1
C = 0	0.8	0.2
C = 1	0.2	0.8

		W = 0	W =1
S = 0	R = 0	1	0
S = 1	R = 0	0.1	0.9
S = 0	R = 1	0.1	0.9
S = 1	R = 1	0.01	0.99



How to find

$$P(S = 1 \mid C = 1, W = 1)?$$

$$P(S = 1 \mid C = 1, W = 1)$$

$$= \frac{P(W = 1 \mid C = 1, S = 1)}{P(W = 1 \mid C = 1)} \cdot P(S = 1 \mid C = 1)$$

$$= \frac{P(W = 1 \mid C = 1, S = 1)}{P(W = 1 \mid C = 1)} \cdot 0.1$$

$$= \frac{0.0972}{0.7452} \cdot 0.1 \approx 0.013$$

BAYESIAN NETWORKS
INFERENCE

- To find the probability of an event we can use the tables of conditional probabilities of the network.

- We can have more than binary variables by making larger tables.

- The size of the table depends on the number of edges entering the node. For binary variables it is 2^k with k the in-degree of the node.

- Inference in Bayesian networks is, in the general case, intractable from a computational point of view…

- …but for specific cases it can still be performed efficiently.

USE OF BN FOR INFORMATION RETRIEVAL

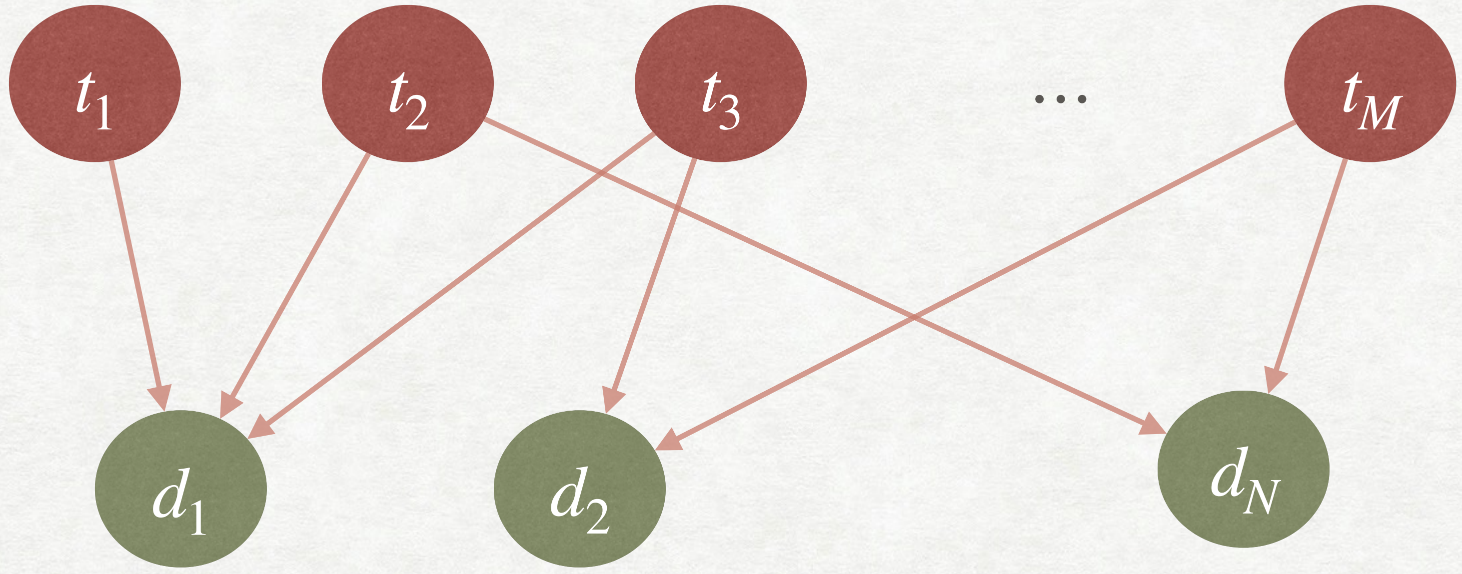
BAYESIAN NETWORKS IN IR

MAIN IDEAS

- Bayesian Networks can model dependencies between terms or documents (contrarily to the assumption of the BIM).
- However, we must always keep an eye to complexity!
- Here we see only one possible model. Other model with different topologies exist.

BN STRUCTURE

A SIMPLE STRUCTURE



Nodes for the terms

Nodes for the documents

Each edge connect a term with a document containing the term.

Both the t_i and d_j are binary random variables with meanings:

- t_i means "the term t_i is relevant"
- d_j means "the document d_j is relevant"

SETTING THE PROBABILITIES

FOR TERMS AND DOCUMENTS

t_i

t_i	not t_i
1/M	1-1/M

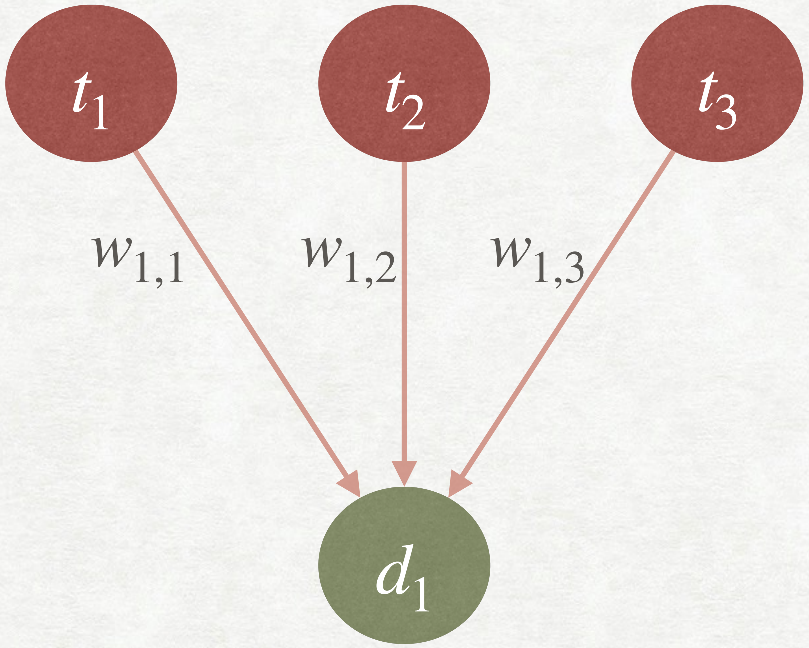
d_j

The size of the table depends *exponentially* by the number of terms in the document: with 50 terms we need a table of 2^{50} entries.

A different approach is needed to store the conditional probabilities

SETTING THE PROBABILITIES

FOR TERMS AND DOCUMENTS



We assign weights to each edge

The value $P(d_j \mid \mathrm{Pa}(d_j))$ is now computed as:

$$P(d_j \mid \mathrm{Pa}(d_j)) = \sum_{i : t_i \in \mathrm{Pa}(d_j),\ t_i = 1} w_{i.j}$$

i.e., sum all $w_{i,j}$ for all the parent nodes with state 1 (relevant)

SETTING THE WEIGHTS

ONE METHOD OF WEIGHTING

Multiple weighting methods are possible.
Two conditions to be respected are:

- $w_{i,j} \geq 0$ for all i and j.
- $\sum_{t_i \in d_j} w_{i,j} \leq 1$ for all documents d_j.

MADE TO "RESEMBLE" THE COSINE MEASURE

One possible weighting scheme is

$$w_{i,j} = \alpha^{-1} \frac{\text{tf-idf}_{i,j}^2}{\sum t_k \in d_j \left(\text{tf-idf}_{k,j}\right)^2}$$

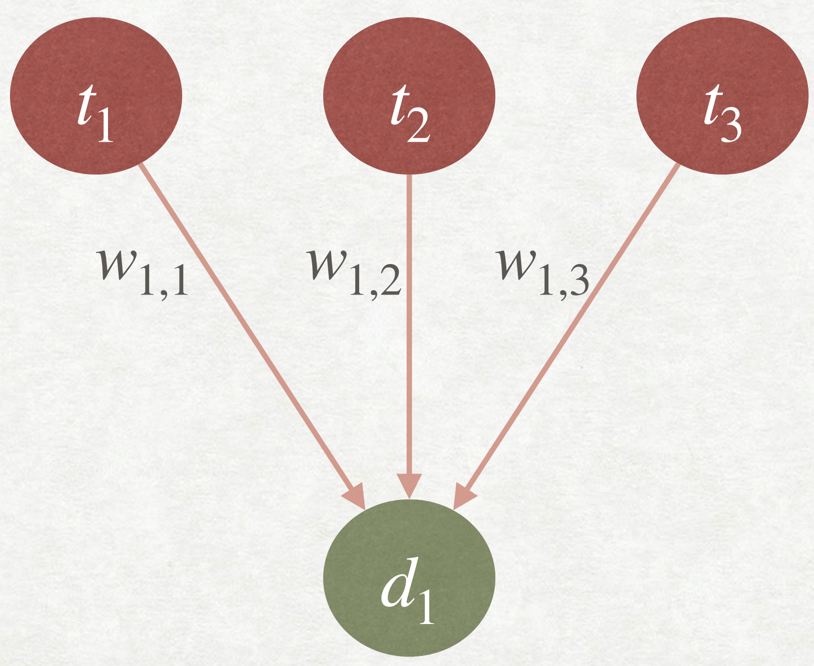
With α a normalising constant

USING A QUERY

HOW THE QUERY SETS THE STATE OF TERMS

Given a query q we assume that all terms in q are relevant (i.e., $t_i = 1$ if $t_i \in q$).
We use the notations $P(t_i \mid q)$ and $P(d_j \mid q)$

Suppose $q = t_1\ t_3$, then $P(d_1 \mid q)$ is:



$$P(d_1 \mid q) = w_{1,1} + w_{1,2} \cdot \frac{1}{M} + w_{1,3}$$

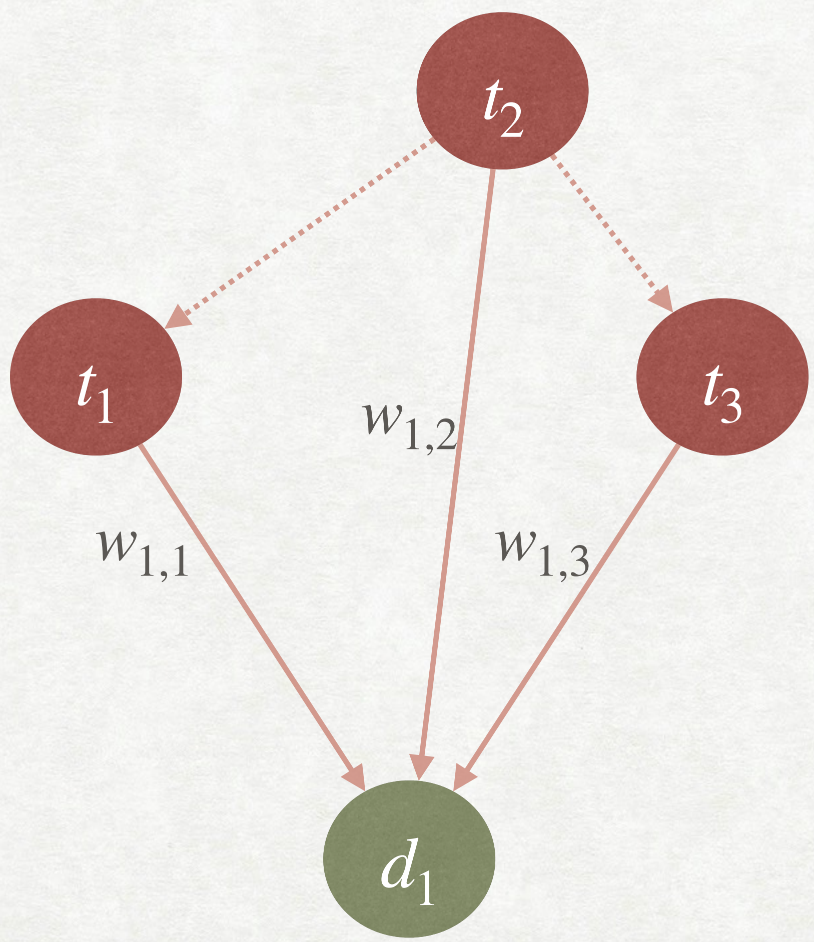
In general:

$$P(d_j \mid q) = \sum_{i: p_i \in \mathrm{Pa}(d_j)} w_{i,j} P(t_j \mid q)$$

ADDING DEPENDENCIES

AT LEAST AMONG TERMS

Until now we have considered the term independent from one another. We can now add some form of dependency between terms while keeping the graph acyclic.



Now we need a way to set the probabilities for root nodes (without any parent) and for nodes with parents.

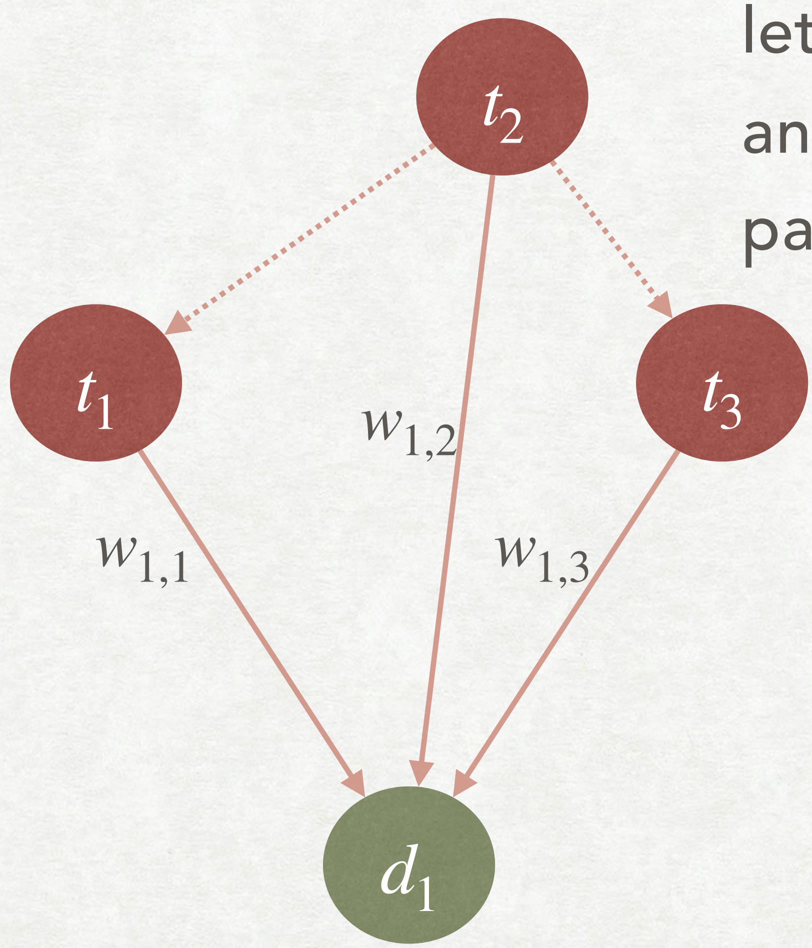
For root nodes we already have:

t_i	not t_i
1/M	1-1/M

ADDING DEPENDENCIES

SETTING THE WEIGHTS

We can use the idea for the Jaccard coefficient of "similarity" among terms



Given a "configuration" x of the parent terms (i.e., which terms are present and which are not) let $A_{\bar{t}_i,x}$ be the set of documents not containing t_i and containing the exact "configuration" x of the parent node. Similarly, define $A_{\bar{t}_i}$ and A_x. Then:

$$P(t_i = 0 \,|\, \mathrm{Pa}(t_i) = x) = \frac{|A_{\bar{t}_i,x}|}{|A_{\bar{t}_i}| + |A_x| - |A_{\bar{t}_i,x}|}$$

$$P(t_i = 1 \,|\, \mathrm{Pa}(t_i) = x) = 1 - P(t_i = 0 \,|\, \mathrm{Pa}(t_i) = x)$$

BAYESIAN NETWORKS

FINAL REMARKS

- We have seen only one model of IR using Bayesian networks.
- We can actually also add some dependencies between documents.
- In any case we must find a way to design or learn the dependencies. E.g., by estimating $P(d_i | d_j)$ and linking the "top documents"
- Other models are possible, including ones with completely different topologies, like mapping document to terms and then to "general concepts".