Bayesian Belief

08 January 2021 15:08

Introduction

A belief network is a representation of a particular independence among variables. Belief networks should be viewed as a modelling language.

Many domains are concisely and naturally represented by exploiting the independencies that belief networks compactly represent.

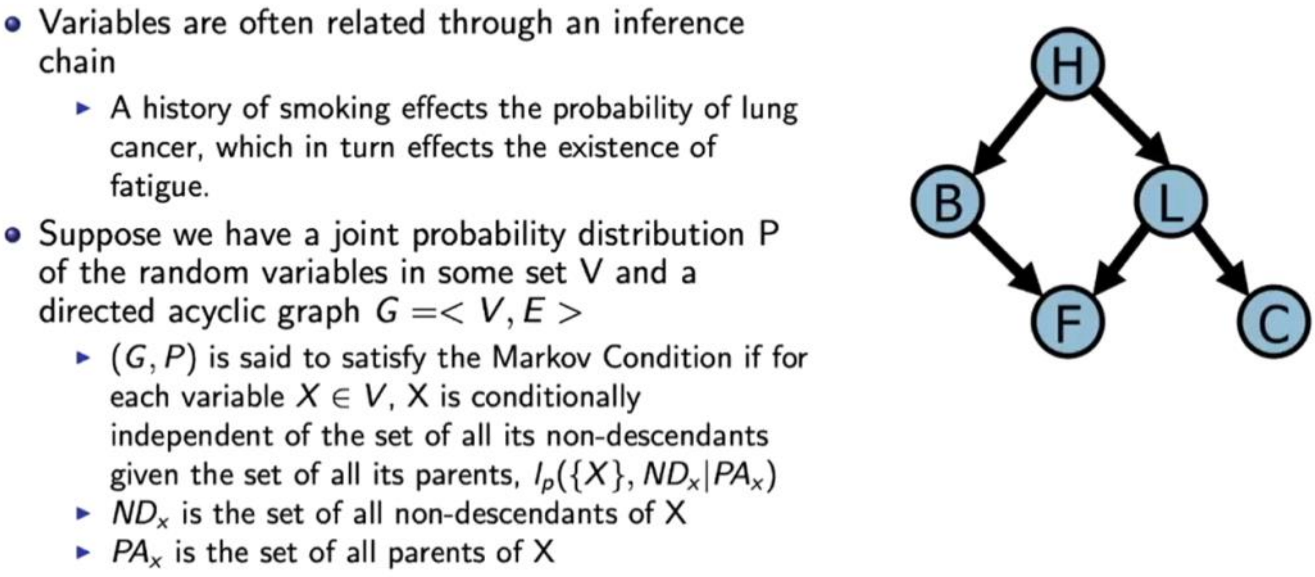
Once the network structure and the domains of the variables for a belief network are defined, which numbers are required (the conditional probabilities) are prescribed. The user cannot simply add arbitrary conditional probabilities but must follow the network's structure. If the numbers required of a belief network are provided and are locally consistent, the whole network will be consistent.

In contrast, the maximum entropy or random worlds approaches infer the most random worlds that are consistent with a probabilistic knowledge base. They form a probabilistic knowledge representation of the second type. For the random worlds approach, any numbers that happen to be available are added and used.

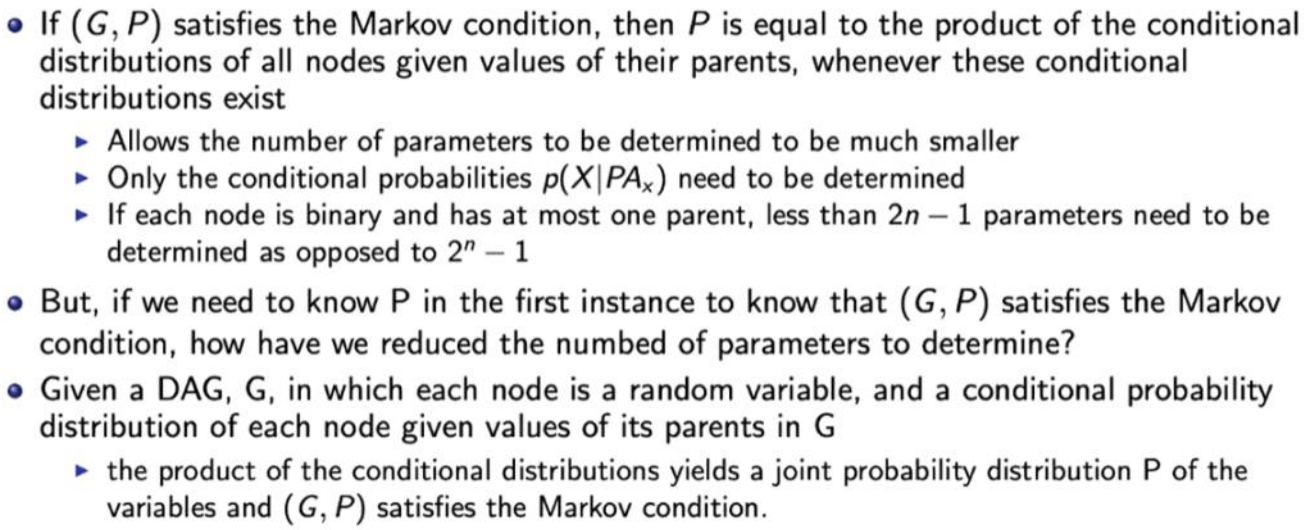
However, if you allow someone to add arbitrary probabilities, it is easy for the knowledge to be inconsistent with the axioms of probability. Moreover, it is difficult to justify an answer as correct if the assumptions are not made explicit.

Markov Condition

Before we define a Bayesian Belief Network, we need to understand the Markov Condition.



- Variables are often related through an inference chain
  - A history of smoking effects the probability of lung cancer, which in turn effects the existence of fatigue.
- Suppose we have a joint probability distribution P of the random variables in some set V and a directed acyclic graph $G =< V, E >$
  - (G, P) is said to satisfy the Markov Condition if for each variable $X \in V$, X is conditionally independent of the set of all its non-descendants given the set of all its parents, $I_p(\{X\}, ND_x|PA_x)$
  - ND_x is the set of all non-descendants of X
  - PA_x is the set of all parents of X



- If (G, P) satisfies the Markov condition, then P is equal to the product of the conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist
  - Allows the number of parameters to be determined to be much smaller
  - Only the conditional probabilities $p(X|PA_x)$ need to be determined
  - If each node is binary and has at most one parent, less than $2n - 1$ parameters need to be determined as opposed to $2^n - 1$
- But, if we need to know P in the first instance to know that (G, P) satisfies the Markov condition, how have we reduced the numbed of parameters to determine?
- Given a DAG, G, in which each node is a random variable, and a conditional probability distribution of each node given values of its parents in G
  - the product of the conditional distributions yields a joint probability distribution P of the variables and (G, P) satisfies the Markov condition.

Belief Networks

The notion of conditional independence is used to give a concise representation of many domains. The idea is that, given a random variable X, there may be a few variables that DIRECTLY affects the X's value, in the sense that X is conditionally independent of other variables given these variables. The set of locally affecting variables is called the Markov Blanket. This locality is exploited in a belief network.

A belief network is a directed model of conditional dependence among a set of random variables. The conditional independence in a belief network takes in an ordering of the variables and results in a directed graph.

Defining a Belief Network

To define a BN on a set of random variables **{X1, X2, X3}** first select a total ordering of the variables, say **X1, X2, X3**. The chain rule shows how we can then decompose a conjunction into conditional probabilities.

$$
\begin{aligned}
P(X_1 &= v_1 \wedge X_2 = v_2 \wedge \cdots \wedge X_n = v_n) \\
&= \prod_{i=1}^{n} P\left(X_i = v_i \mid X_1 = v_1 \wedge \cdots \wedge X_{i-1} = v_{i-1}\right).
\end{aligned}
$$

In terms of random variables and probability distributions..

$$P\left(X_1, X_2, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i \mid X_1, \ldots, X_{i-1}\right).$$

Define the parents of random variable Xi (written parents(Xi) to be a minimal set of predecessors of Xi in the total ordering such that the other predecessors of Xi are conditionally independent of Xi given parents(Xi). Thus Xi probabilistically depends on each of its parents but is independent of its other predecessors.

$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid parents(X_i)) .$$

Putting the chain rule and the parents definition together, we arrive at

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

The probability over all of the variables P(X1, X2, ... Xn) is called the joint probability distribution. A BN defines a factorization of the joint probability distribution into a product of conditional probabilities.

A belief network, also called a Bayesian Network, is an acyclic direct graph where the nodes are random variabels. There is an arc from each element of parents(Xi) into Xi. Associated with the belief network is a set of conditional probability distributions that specify the conditional probability of each variable given its parents which includes the prior probability of those variables with no parents.

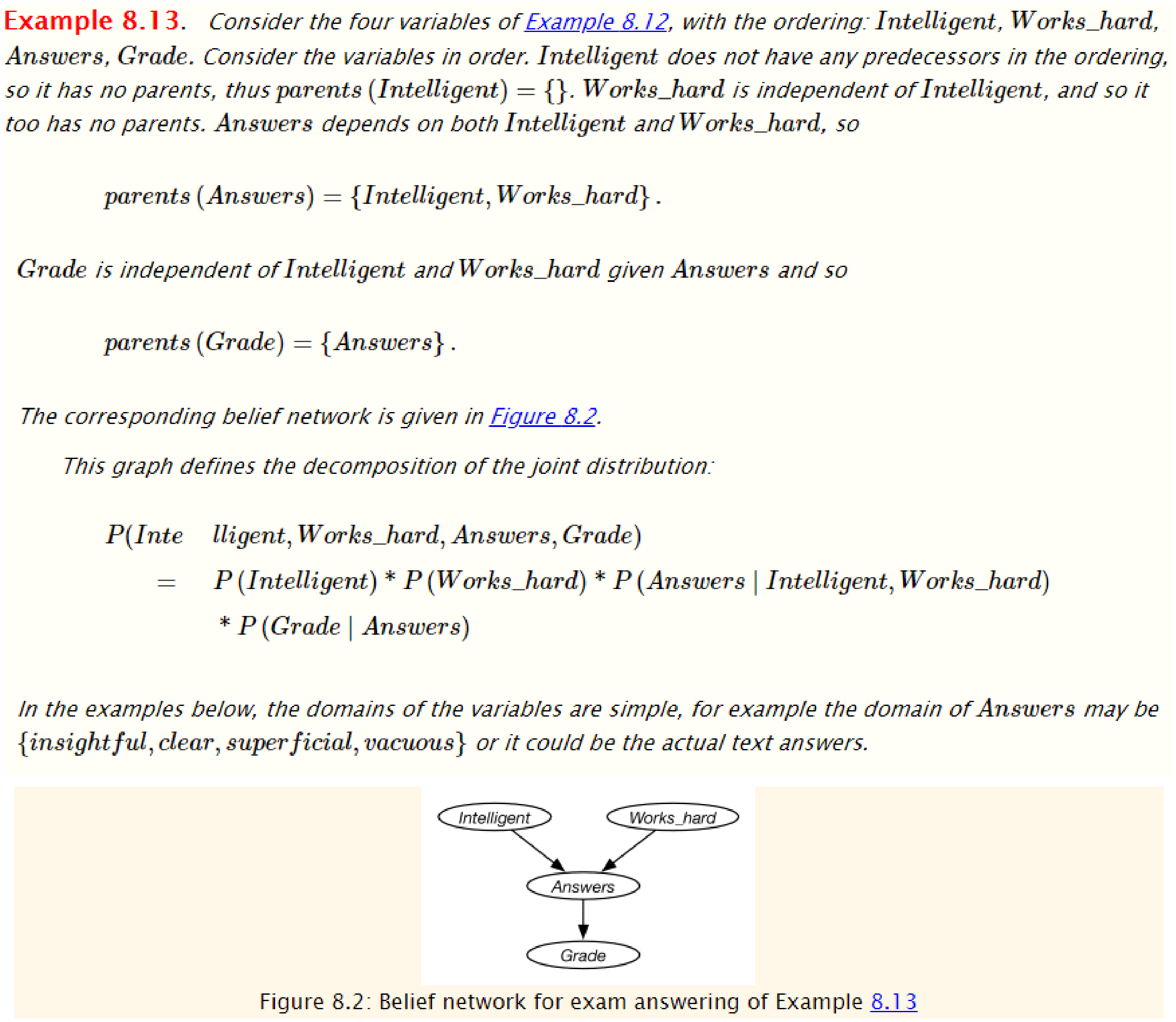
Thus a belief network consists of:

1. A DAG - where each node is labelled by a random variable
2. A domain for each random variables
3. A set of conditional probability distributions giving P(X | parents(X)) for each variable X

A belief network is acyclic by construction.

Remember that different orderings of variables can result in different belief networks. In particular, which variables are parents is dependent on ordering since only predecessor nodes can be parents of a variable. Some of the orderings may result in networks with fewer arcs than others, which generally speaking is a good thing since it simplifies the network.

Example



Example 8.13. *Consider the four variables of Example 8.12, with the ordering:* $Intelligent$, $Works\_hard$, $Answers$, $Grade$. *Consider the variables in order.* $Intelligent$ *does not have any predecessors in the ordering, so it has no parents, thus* $parents\,(Intelligent) = \{\}$. $Works\_hard$ *is independent of* $Intelligent$, *and so it too has no parents.* $Answers$ *depends on both* $Intelligent$ *and* $Works\_hard$, *so*

$$parents\,(Answers) = \{Intelligent, Works\_hard\}\,.$$

$Grade$ *is independent of* $Intelligent$ *and* $Works\_hard$ *given* $Answers$ *and so*

$$parents\,(Grade) = \{Answers\}\,.$$

The corresponding belief network is given in Figure 8.2.

This graph defines the decomposition of the joint distribution:

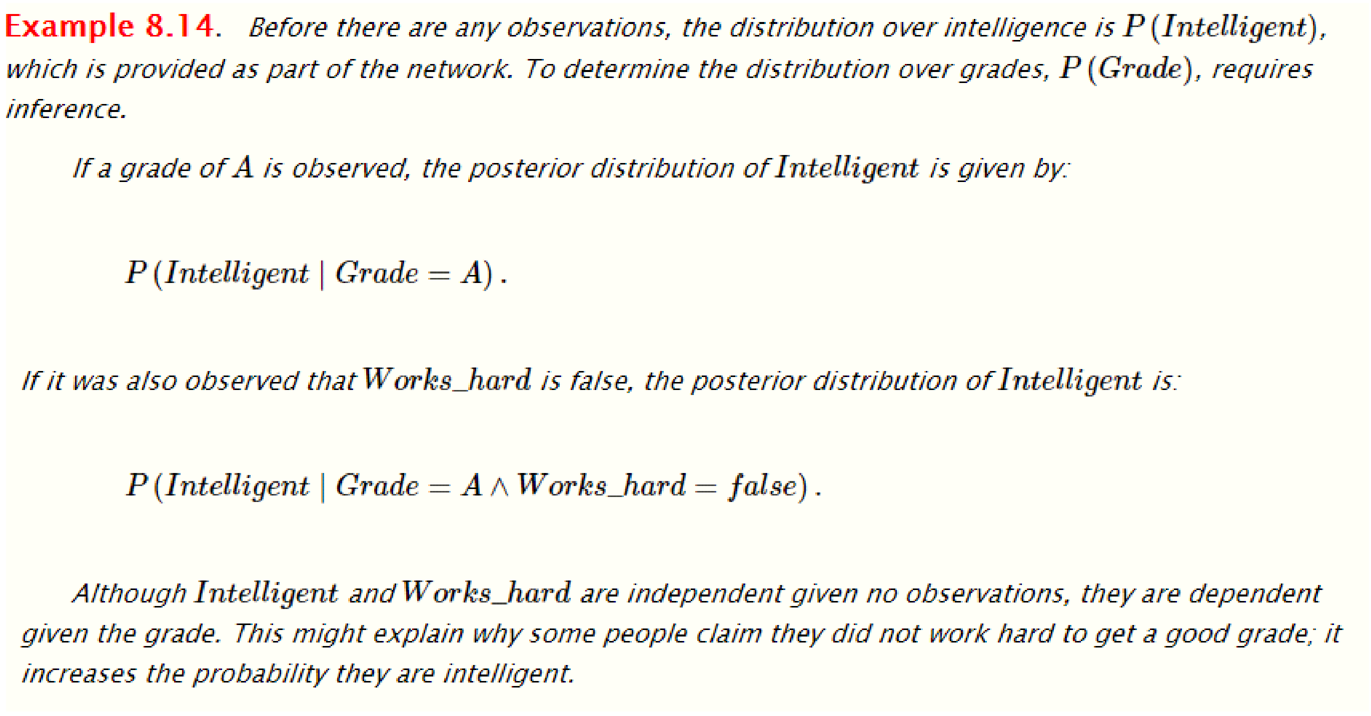
$$\begin{aligned} P(Inte &\quad lligent, Works\_hard, Answers, Grade) \\ = &\quad P\,(Intelligent) * P\,(Works\_hard) * P\,(Answers \mid Intelligent, Works\_hard) \\ &\quad * P\,(Grade \mid Answers) \end{aligned}$$

In the examples below, the domains of the variables are simple, for example the domain of $Answers$ *may be* $\{insightful, clear, superficial, vacuous\}$ *or it could be the actual text answers.*

Figure 8.2: Belief network for exam answering of Example 8.13

Observations and Queries

A belief network specifies a joint probability distribution from which arbitrary conditional probabilities can be derived. The most common probabilistic inference task is to compute the posterior distribution of a query variable, or variables, given some evidence, where the evidence is a conjunction of assignments of values to some of the variables.



Example 8.14. *Before there are any observations, the distribution over intelligence is* $P(Intelligent)$, *which is provided as part of the network. To determine the distribution over grades,* $P(Grade)$, *requires inference.*

If a grade of A *is observed, the posterior distribution of* $Intelligent$ *is given by:*

$$P(Intelligent \mid Grade = A).$$

If it was also observed that $Works\_hard$ *is false, the posterior distribution of* $Intelligent$ *is:*

$$P(Intelligent \mid Grade = A \wedge Works\_hard = false).$$

Although $Intelligent$ *and* $Works\_hard$ *are independent given no observations, they are dependent given the grade. This might explain why some people claim they did not work hard to get a good grade; it increases the probability they are intelligent.*

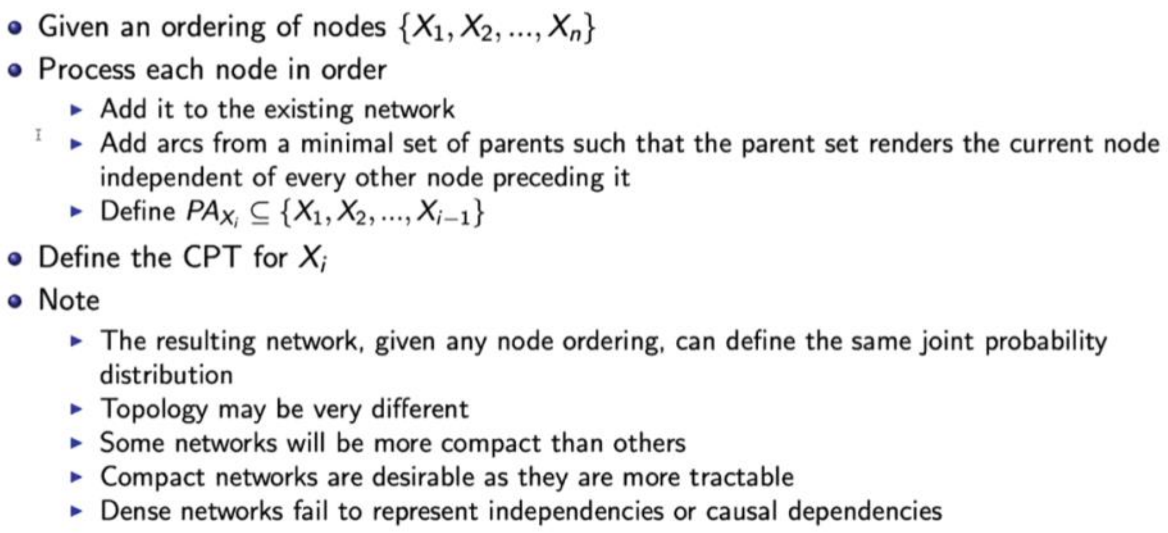
Constructing Belief Networks

Book

To represent a domain in a belief network, the designer must consider the following:

1. What are the relevant variables?
    a. What the agent may *observe in the domain - each feature should be a variable*, because the agent must be able to condition on all of its observations
    b. What information the agent is interested in knowing the **posterior probability** of. Each of these needs to be a *variable that can be queried*
    c. Other hidden variables that will not be observed or queried but make the model simpler. These variables either account for dependencies, reduce the size of the specification of the conditional probabilities, or better model how the world is assumed to work
2. What values should these variables take?
    a. For each variable, the designer should *specify what it means to take each value in its domain*. What must be true in the world for a variable to have a particular value? This must satisfy the **clarity principle**: an omniscient agent should be able to know the value of a variable.
3. What is the r*elationship between he variables*? This should be expressed by adding arcs in the graph to define the parent relation
4. How does the *distribution of a variable depend on its parents*? This is express in terms of the conditional probability distributions.

See examples here.



- Given an ordering of nodes $\{X_1, X_2, ..., X_n\}$
- Process each node in order
  - Add it to the existing network
  - Add arcs from a minimal set of parents such that the parent set renders the current node independent of every other node preceding it
  - Define $PA_{X_i} \subseteq \{X_1, X_2, ..., X_{i-1}\}$
- Define the CPT for X_i
- Note
  - The resulting network, given any node ordering, can define the same joint probability distribution
  - Topology may be very different
  - Some networks will be more compact than others
  - Compact networks are desirable as they are more tractable
  - Dense networks fail to represent independencies or causal dependencies

Representing Conditional Probabilities and Factors

A conditional probability distribution is a function on variables - given an assignment to the values of the variables, it gives a number.

Even with a small number of nodes, this conditional probability table can grow very large.

The relationships between parents and child nodes usually fall into one of several categories that have canonical distributions - i.e. they fit some standard pattern.

The simplest example is provided by deterministic nodes who depend solely on their parent nodes.

Uncertain relationships can be characterized by noisy logical relationships.

Noisy-OR

A generalisation of the logical OR. The noisy-or adds some uncertainty to the standard logical-OR.

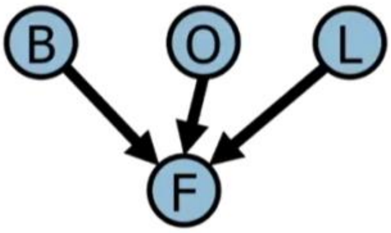
The model makes 3 assumptions:

- each cause has an independent chance of causing the effect
- all the possible causes are listed (we can add a leak-node to catch all miscellaneous causes)
- it assumes whatever inhibits the cause from creating the effect is independent of what inhibits another cause from causing the effect. These inhibitors are not represented as nodes but as "noise parameters".

e.g. if P(Fever|Cold) = 0.4, then the noise parameter is 0.6.

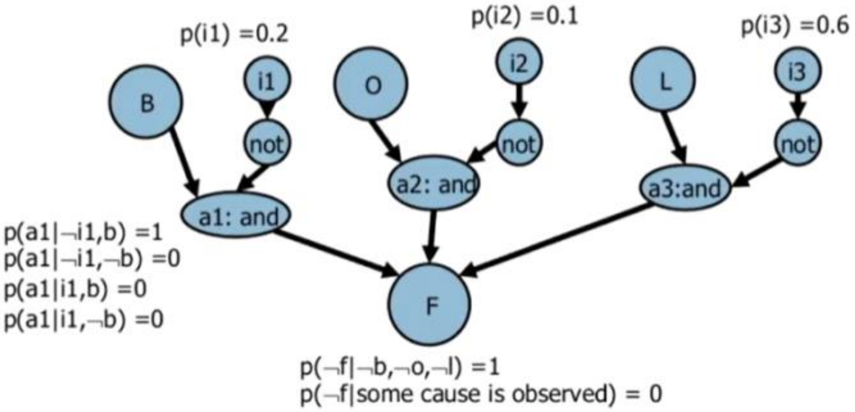
If exactly one parent is true, then the output is false with probability equal to the noise parameter for that node. In general, the probability that the output node is False is just the product of the noise parameters for all the input nodes that are true.

- Local probability distributions can get large as they are $O(2^n)$
- We can approximate these distributions by using canonical interaction models that require fewer parameters
- Noisy-OR:
  - Describes a set of n causes $(x_i' s)$ and their common effect (y)
  - Assumes each x_i is sufficient to cause the effect, y, in the absence all other causes.
  - The ability of x_i to cause y is *independent* of the presence of the other causes
- Only need k parameters:
  - $p_i = p(y|\neg x_1, ..., \neg x_{i-1}, x_i, \neg x_{i+1}, ..., x_k)$



- Consider the BBN representing the relationship between Fatigue, Lung Cancer, Bronchitis and Other Causes
- Causal Inhibition: Each cause has an inhibitor, that inhibits the expression of the cause
  - The effect is observed if and only if the inhibitor is disable
  - Bronchitis will result in Fatigue if and only if the mechanism that inhibits Bronchitis from causing Fatigue is not present
- The inhibiting mechanism of one cause is independent of the mechanism of other causes (Exception independence)
- The effect can happen only if at least one of its causes is present and *not* being inhibited (Accountability)
  - $p(\neg B, \neg O, \neg L, F) = 0$

- Nodes whose value is exactly specified by the parent nodes are called deterministic nodes, i.e. a_1, a_2 and a_3 in the BBN below
- An inhibitor has a probability of being "observed"


p(i1) =0.2
p(i2) =0.1
p(i3) =0.6
B
i1
O
i2
L
i3
not
not
not
a1: and
a2: and
a3:and
F
p(a1|¬i1,b) =1
p(a1|¬i1,¬b) =0
p(a1|i1,b) =0
p(a1|i1,¬b) =0
p(¬f|¬b,¬o,¬l) =1
p(¬f|some cause is observed) = 0