Representing Complex Distributions*

Daphne Koller
Stanford University

Nir Friedman
Hebrew University

November 7, 2000

Our most basic model is a joint distribution over some finite set of random variables $X_1, \ldots, X_n$. The key issue that we will tackle in this part of the book is the task of representing complex joint distributions compactly, and in a way that will later allow us to construct efficient inference and learning algorithms. The key to accomplishing this goal is the use of certain important structural properties of the joint distribution, properties that seem to hold in a wide variety of important domains.

1 Exploiting independence properties

Consider the problem of representing a joint distribution P over n random variables $X_1, \ldots, X_n$. Even in the simplest case where these variables are binary-valued, a joint distribution requires the specification of 2^n numbers — the probabilities of the 2^n different assignments of values $x_1, \ldots, x_n$. However, some distributions can be represented much more compactly. For example, assume that we know that each X_i represents the outcome of a toss of coin i. In this case, we typically assume that the different coin tosses are independent, so that our distribution P will satisfy $\mathrm{Ind}(X_i; X_j)$ for any i, j. In fact, a much stronger independence statement holds: the i'th toss, X_i, is independent of all other tosses together. That is, $\mathrm{Ind}(X_i; X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for any i. Note that this implies that any pair is independent as well (see Exercise **??**).

If θ_i is the probability with which coin i lands heads, the joint distribution P can be specified using the n parameters $\theta_1, \ldots, \theta_n$. These parameters implicitly specify the 2^n probabilities in the joint distribution. For example, the probability that all of the coin tosses landed heads is simply $\theta_1 \cdot \theta_2 \cdot \ldots \cdot \theta_n$. More generally, letting $\theta_{x_i^1}$ denote θ_i and $\theta_{x_i^0}$ denote $1 - \theta_i$, we can define:

$$P(x_1, \ldots, x_n) = \prod_i \theta_{x_i} \tag{1}$$

Of course, we cannot represent every joint distribution over this space in such a way. This fact is obvious not only from intuition, but also from a somewhat more formal perspective. The space of all joint distributions is a $2^n - 1$ dimensional subspace of $\mathbb{R}^{2^n}$ — the set $\{(p_1, \ldots, p_{2_n}) \in \mathbb{R}^{2^n} : p_1 + \ldots + p_{2^n} = 1\}$. On the other hand, the space of all joint distributions specified in a factorized way as in Eq. (1) is an n-dimensional manifold in $\mathbb{R}^{2^n}$. Therefore, the two spaces cannot be the same.[1]

As this simple example shows, certain families of distributions — in this case, the distributions generated by n independent random variables — permit an alternative parameterization which

*this is based on draft of a text book, some of the references and citations are missing

[1]While this argument might seem trivial in this case, it turns out to be an important tool for comparing the expressive power of different representations.

is substantially more compact than the naive representation as an explicit joint distribution. Of course, in most of the applications that are of interest to us, the random variables are not independent. However, a very similar idea will be the basis for our solution. This representation will also be based on the two ideas of exploiting structural assumptions about the distribution, and using an alternative parameterization that exploits it. Before presenting the general framework, let us begin with a simple example that illustrates the basic intuition.

Consider the problem of determining whether a woman is pregnant or not. The doctor can administer a test, whose result is very informative, but not fully reliable. In this case, our probability space is the joint distribution over these two random variables: whether the woman is pre**g**nant, G, and whether the **d**octor's test comes out positive, D. Assuming that the two variables are binary, the joint distribution has four entries. For example, one possible joint distribution P would be

G	D	$P(G,D)$
g^0	d^0	0.54
g^0	d^1	0.06
g^1	d^0	0.02
g^1	d^1	0.38

(2)

There is, however, an alternative and substantially more natural way of representing the same joint distribution. From the basic definition of conditional probability we have that

$$P(G,D) = P(G)P(D \mid G).$$

This equation leads to the following alternative way of representing the joint distribution. Instead of specifying the various joint entries $P(g,d)$, we would specify it in the form of $P(g)$ and $P(d \mid g)$. Thus, for example, we can represent the joint distribution of Eq. (2) using the following two tables, one representing the prior distribution over G and the other the *conditional probability distribution (CPD)* of D given G:

g^0	g^1
0.6	0.4

G	d^0	d^1
g^0	0.99	0.01
g^1	0.05	0.95‘

(3)

The latter distribution represents the probability that the test will come out positive in the two possible circumstances: in case of pregnancy g^1 and in case of no pregnancy g^0. Note that $P(d^0 \mid g^1)$ is the probability of a *false negative* (5% in this model), whereas $P(d^1 \mid g^0)$ is the probability of a *false positive* (1% here).

It is instructive to consider how we could parameterize this alternative representation. In this case, we need three independent parameters, say θ_{g^1}, $\theta_{d^1|g^1}$, and $\theta_{d^1|g^0}$. Thus, although this alternative representation is more natural than the explicit representation of the joint (as the probabilities of joint events such as g^1, d^1), it is not more compact. In this case, both representations require three independent parameters. However, this alternative parameterization is the basis for our more compact representations below.

Although we will only define Bayesian networks formally in Section 2.2, it is instructive to see how this example would be represented as one. In this case, the Bayesian network would have a node for each of the two random variables G and D, with an edge from G to D representing the direction of the dependence in this model.

Let us now consider a more complicated setting, where the woman can also take a quick and simple home test, which is less reliable than the test given by the doctor. In this case, our probability space is the joint distribution over the three relevant random variables G, D, and the outcome of the

home test, H. Assuming that all three variables are binary, the joint distribution has eight entries. Before we even consider the specific numerical aspects of our distribution P over this domain, we can see that independence does not help us: for any reasonable P, there are no independence that hold. A test result is clearly correlated with whether the woman is pregnant or not. The two test results H and D are also not independent: if we condition on the fact that one of the tests came out positive, the chances that the other test is positive are going to increase. That is, $P(d^1 \mid h^1) > P(d^1)$.[2]

However, it is quite plausible that our distribution P in this case satisfies a conditional independence property. If we are certain that the woman is pregnant, the outcome of one of the two tests no longer gives us any information about the other. Thus, for example,

$$P(d^1 \mid g^1, h^1) = P(d^1 \mid g^1).$$

More generally, we may well assume that

$$P \models \mathrm{Ind}(H; D \mid G) \tag{4}$$

Note that this assumption rests on the fact that the reasons why the tests fail are completely different. If the two tests fail because of the same reason, e.g., because the woman hasn't been pregnant long enough for the hormones to be generated, then information about one test can still influence my probability in the other. We will discuss this point further later on.

As for the case of marginal independence, conditional independence allows us to provide a compact specification of the joint distribution. Again, the compact representation is based on a very natural alternative parameterization. By simple probabilistic reasoning, as above, we have that

$$P(G, H, D) = P(H, D \mid G)P(G).$$

But now, the conditional independence assumption Eq. (4) implies that

$$P(H, D \mid G) = P(H \mid G)P(D \mid G).$$

Hence, we have that

$$P(G, H, D) = P(H \mid G)P(D \mid G)P(G). \tag{5}$$

This factorization immediately leads us to the desired alternative parameterization. In order to fully specify a joint distribution satisfying Eq. (4), we need the following: $P(G)$, $P(D \mid G)$, and $P(H \mid G)$. The first two might be the same as in Eq. (3). The latter might be

G	h^0	h^1
g^0	0.9	0.1
g^1	0.2	0.8

Together, these three CPDs fully specify the joint distribution (assuming the conditional independence of Eq. (4)). For example,

$$P(g^0, d^1, h^0) = 0.6 \cdot 0.01 \cdot 0.9 = 0.0054.$$

[2] These statements are not formally "true"; clearly, it is perfectly legitimate to have beliefs about this domain that would have any set of independencies. Rather, these are statements about "reasonable" probability distributions for this domain.

In this case, the alternative parameterization is more compact than the joint. Five independent parameters suffice to specify any distribution satisfying Eq. (4): θ_{g^1}, $\theta_{d^1|g^1}$, $\theta_{d^1|g^0}$, $\theta_{h^1|g^1}$, and $\theta_{h^1|g^0}$. By contrast, seven independent parameters are required to specify an arbitrary joint distribution.

It is important to note another important advantage of this way of representing the joint: modularity. When we added the new variable H, the joint distribution changed entirely. Had we used the explicit representation of the joint, we would have had to write down eight new numbers. In the factored representation, we could reuse our local probability models for the variables G and D, and specify only the new probability model for H.

This model is an instance of a more general *Naive Bayes model* (also known as *Idiot Bayes model*). The Naive Bayes model assumes that individuals fall into one of a number of mutually exclusive and exhaustive classes. Thus, we have a class variable C that takes on values in some set $\{c_1, \ldots, c_K$. We also have a number of *finding* or *evidence* variables $X_1, \ldots, X_n$ whose values are observable. Our goal is to decide, based on the values of the finding variables for a given instance, how likely the various classes are. This model was first used in medical diagnosis, where the classes are diseases and the findings are symptoms, test results, etc. In the naive Bayes model, we assume that all of these variables are conditionally independent given the class. More formally, we have that

$$\mathrm{Ind}(X_i; X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \mid C) \quad \text{for all } i \tag{6}$$

In the Naive Bayes model, we can describe the distribution using only a very small number of parameters: $P(C)$, specifying the prior probability in the various classes; and for each finding variable X_j, $P(X_j \mid C)$, specifying the conditional probability of the values of the variable in the different classes. For example, if all of the variables are binary, the number of independent parameters required to specify the distribution is $2n + 1$ (two for each finding and one for the class, as above). Thus, the number of parameters is linear in the number of variables, as opposed to exponential for the explicit representation of the joint.

The Naive Bayes model was used in some of the first probabilistic expert systems (particularly the medical ones) []. It is conceptually simple, supports very efficient reasoning (as we'll discuss later on), and works reasonably well in surprisingly many cases. However, the independence assumptions are clearly unrealistic in most settings. Its obvious limitations, and the fact that it was not clear how to extend the model to richer settings, led many people to abandon the probabilistic framework for expert system design. Bayesian networks, by allowing the specification of probabilistic models that do not make these strong independence assumptions, have made probabilistic reasoning applicable in a much wider variety of settings. We note, however, that the Naive Bayes is still heavily used, with significant success, in machine learning and pattern recognition.

2 Bayesian networks

Bayesian networks build on the same intuitions as the Naive Bayes model by exploiting independence properties of the distribution in order to allow a compact and natural representation. However, they are not restricted to representing distributions satisfying the very strong independence assumptions implicit in the Naive Bayes model. They allow us the flexibility to tailor our representation of the distribution to the independence properties that appear reasonable in the current setting.

The core of the Bayesian network representation is a directed acyclic graph (DAG) G, whose nodes are the random variables in our domain and whose edges correspond, intuitively, to direct influence of one node on another. This graph G can be viewed in two very different ways:

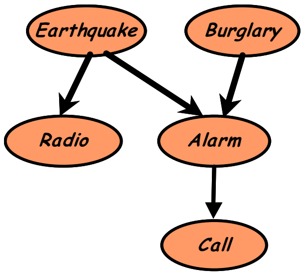


Figure 1: The Alarm DAG.

- as a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way;
- as a compact representation for a set of conditional independence assumptions about a distribution.

As we will see, these two views are, in a strong sense, equivalent.

We will start by providing an example of a Bayesian network, and how it can be used as a data structure for a joint distribution. Then, we will proceed to provide a formal definition of Bayesian networks and the conditional independence statements they encode. We will then show that these two different views of a Bayesian network are equivalent.

2.1 The Alarm example

We begin by showing an example of a Bayesian network, which illustrates how the graphical structure can be used as a basis for representing a complex joint distribution in a factorized way.

Consider a domain where we are trying to reason about an alarm installed in a house. The alarm can go off either because of a burglary or because of a minor earthquake. If there is an alarm, then a neighbor might call. If there is an earthquake, the local radio station may report it. There are five random variables in this domain, all of which are binary valued: A (alarm); B (burglary); E (earthquake); C (phone call); R (radio report). Thus, the joint distribution has 32 entries.

Recall that a Bayesian network is based on a graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another. (As we'll discuss below, it is often convenient to view these edges as representing direct causal influence.) In this domain, perhaps the most obvious network structure is the one presented in Figure 1. The edges encode our intuition about the way the world works:

- There either is or is not a burglary, but that choice does not depend on any of the other variables in the model.
- Similarly, the choice of whether there is or is not an earthquake happens separately and independently of everything else.
- The alarm can be set off either by a burglary or by an earthquake, and therefore it depends on both.
- The neighbor calls based on whether he hears the alarm. His decision to call depends only on that.
- The radio report depends only on whether there was an earthquake.

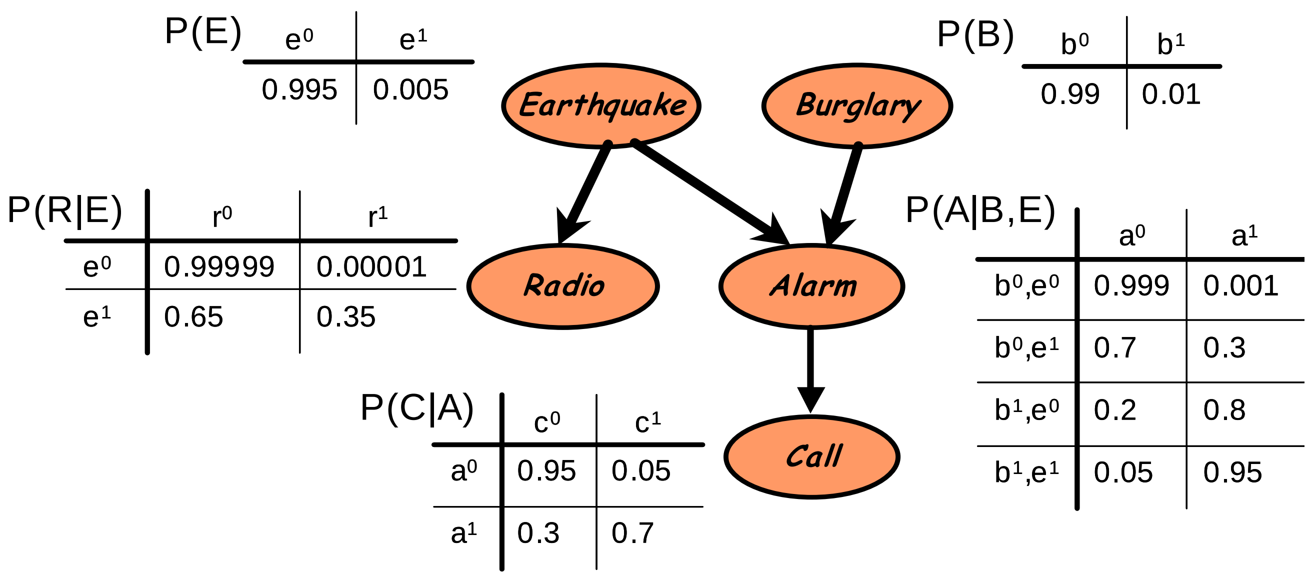


Figure 2: The Alarm Bayesian network with CPTs.

We can now annotate this graph to incorporate the quantitative aspects of the different events — the local probability models associated with the different variables. One such model — $P(A)$ — would specify the prior probability of a burglary. Another — $P(E)$ — specifies the prior probability of a minor earthquake. The probability of the alarm going off is a *conditional* probability distribution (CPD) — $P(A \mid B, E)$; it specifies the probability of the alarm going off in any of the relevant circumstances: a burglary and an earthquake (b^1, e^1); a burglary and no earthquake (b^1, e^0); no burglary and an earthquake (b^0, e^1); no burglary and no earthquake (b^0, e^0). The local probability model of C is also a CPD, $P(C \mid A)$, representing how likely the neighbor is to call if there is an alarm and if there is no alarm. Similarly, we have a CPD for $P(R \mid E)$. One possible choice of CPDs for this domain is shown in Figure 2.[3]

How do we use this data structure to specify the joint distribution? Consider some particular state in this space, e.g., b^1, e^0, a^1, c^1, c^0. Intuitively, the probability of this event can be computed from the probabilities of the basic events that comprise it: the probability that there is a burglary, 0.01; the probability that there is no earthquake, 0.995; the probability that the alarm goes off in this situation, 0.8; the probability that the neighbor calls when the alarm goes off, 0.7; and the probability that the radio does not report an earthquake if there is none, 0.99999. Thus, the total probability of this state is:

$$P(b^1)P(e^0)P(a^1 \mid b^1, e^0)P(c^1 \mid a^1)P(r^0 \mid e^0) = 0.01 \cdot 0.995 \cdot 0.8 \cdot 0.7 \cdot 0.99999 = 0.005566428.$$

Clearly, we can use the same process for any state in the joint probability space. In general, we will have that

$$P(B, E, A, C, R) = P(B)P(E)P(A \mid B, E)P(C \mid A)P(R \mid E) \tag{7}$$

This equation is our first example of the *chain rule for Bayesian networks*, which we will define in a general setting later on.

[3]Our choice of probabilities is somewhat unrealistic, as the probabilities of various unlikely events (e.g., an earthquake) is much higher than is plausible. However, the probabilities we use do illustrate accurately the type of qualitative reasoning patterns that a Bayesian network distribution achieves.

Now that we have defined a joint distribution, it specifies (albeit implicitly) the probability of any event given any observations. Let us go through a scenario, and see how the probabilities in various events change as evidence is obtained.

A priori, the probability of a burglary is 0.01 and the probability of an earthquake is 0.005. The prior probability that our neighbor calls is 0.0568. How reliable is our neighbor at detecting burglaries? The probability that he calls in case of a burglary — $P(c^1 \mid b^1)$ — is 0.5705, substantially higher than the prior, although not an extremely reliable sensor. The rate of false positives — $P(c^1 \mid b^0)$ is 0.0516, around 5%; this error is composed of cases where the alarm goes off without a burglary (whether due to an earthquake or for no reason), and cases where our neighbor calls without the alarm. Queries such as these, where we predict the "downstream" effects of various events (such as a burglary) are instances of *causal reasoning* or *prediction*.

We can also reason in the other direction, reaching conclusions about the most likely cause for an observation. For example, assume that we observe that our neighbor calls. The probability of burglary goes up to 0.3251, and the probability of earthquake goes up to 0.1034. Note that these probabilities are still quite low. Intuitively, this behavior is quite reasonable: the prior probabilities in these events is very low, and neither the alarm nor the neighbor are reliable enough to completely overcome that. Queries such as this are instances of *evidential reasoning* or *explanation*.

Now, assume that we turn on the radio and hear a report of an earthquake in our neighborhood. How does that affect our beliefs? As expected, the probability of an earthquake goes up substantially — $P(e^1 \mid c^1, r^1) = 0.9993$ as compared to $P(e^1 \mid c^1) = 0.021$ without the additional evidence from the radio. More interestingly, the probability of a burglary goes *down* — $P(b^1 \mid c^1, r^1) = 0.0268$. Note that burglary and earthquake are not mutually exclusive; indeed, they occur independently. Hence, there is nothing about the increased probability of an earthquake that, by itself, reduces our beliefs in a burglary. Rather, the earthquake provides an explanation for the neighbor's phone call; therefore, our basis for believing the alternative explanation — the burglary — is no longer as active. This type of reasoning is called *explaining away*; it is a very basic reasoning pattern, that people use all the time (e.g., in medical diagnosis).

Explaining away is an instance of a general pattern of reasoning called *intercausal reasoning*, where, in the presence of an effect, information about one cause can change our beliefs in the other. Explaining away is a special case that occurs for binary random variables, where, if the effect is true, the presence of one cause decreases the probability in the other. In general, however, the effect can go in any direction.

Note that the probability of burglary does not go down to its original level of 0.01. This is partially due to the fact that the radio report does not conclusively prove the existence of the earthquake. But it still holds even if we were certain that an earthquake occurred: $P(b^1 \mid c^1, e^1) = 0.0176 > 0.01$. The earthquake does not always cause the alarm to go off, and therefore there is still a chance that the alarm was caused by other means. If, however, our model was such that the alarm always (deterministically) goes off whenever there is an earthquake — $P(a^1 \mid b^1, e^1) = P(a^1 \mid b^0, e^1) = 1$ — we would have that observing an earthquake provides a full explanation for the phone call, and that $P(b^1 \mid c^1, e^1) = P(b^1)$.

It is important to emphasize that, although our explanations used intuitive concepts such as cause and evidence, there is nothing mysterious about the probability computations we performed. They can be replicated simply by generating the joint distribution, as defined in Eq. (7), and computing the probabilities of the various events directly from that.

2.2 Basic independencies in Bayesian networks

As we discussed above, a Bayesian network graph G can be viewed in two ways. In the previous section, we showed, by example, how it can be used as a skeleton data structure to which we can attach local probability models that together define a joint distribution. In this section, we provide a formal semantics for a Bayesian network, starting from the perspective that the graph encodes a set of conditional independence assumptions. We begin by understanding, intuitively, the basic conditional independence assumptions that we want a directed graph to encode. We then formalize these desired assumptions in a definition.

2.2.1 The Alarm example

In the Alarm example, we used the intuition that edges represent direct dependence. For example, we made intuitive statements such as "the radio report depends only on whether there was an earthquake"; this statement was encoded in the graph by the fact that there are no direct edges into the R node except from E. This intuition, that "a node depends directly only on its parents", lies at the heart of the semantics of Bayesian networks.

We give formal semantics to this assertion using conditional independence statements. For example, the assertion above can be stated formally as the assumption that R is conditionally independent of all other nodes in the network given its parent E:

$$\mathrm{Ind}(R; B, A, C \mid E). \tag{8}$$

In other words, once we know, for example, that there was no earthquake, our beliefs in the existence of a radio report are not influenced by information about any other variable. Similarly, to formalize our intuition that the neighbor's decision to call depends only on the alarm, we can say that C is conditionally independent of all other nodes in the network given its parent A:

$$\mathrm{Ind}(C; B, E, R \mid A). \tag{9}$$

Now, let us consider the A node. Following the pattern blindly, we may be tempted to assert that A is conditionally independent of all other variables in the network given its parents. However, this assumption is false both at intuitive level and for the specific example distribution we used above. Assume, for example, that we condition on b^0, e^1. In this setting, is A independent of C? Clearly, the answer is no: if we observe c^1, then our probability in a^1 should go up; that is, we would expect

$$P(a^1 \mid b^0, e^1, c^1) > P(a^1 \mid b^0, e^1).$$

Indeed, if we examine our distribution, the latter probability is 0.3 (as specified in the CPD) whereas the former is a much higher 0.8571.

Thus, we see that we do not expect a node to be conditionally independent of all other nodes given its parents. In particular, even given its parents, it can still depend on its descendants. Can it depend on other nodes? In our example, do we expect A to depend on R given B and E? Intuitively, the answer is 'no'. Once we know, say, that there was an earthquake, the radio report gives us no further information about the alarm. Thus, we would want the property that:

$$\mathrm{Ind}(A; R \mid B, E). \tag{10}$$

It remains only to consider the variables B and E. They both have no parents, so we are looking for marginal independencies. As we discussed, we do not expect a variable to be independent of its descendants. For the variable E, its only non-descendant is E. Indeed, we assumed implicitly that the burglary and earthquake occur independently, i.e., one did not cause the other. Thus, we expect that

$$\mathrm{Ind}(E; B). \tag{11}$$

For the variable B, both E and R are non-descendants. If $\mathrm{Ind}(E; B)$ then $\mathrm{Ind}(B; E)$. The variable R increases our beliefs in the existence of an earthquake, but since earthquakes are burglaries are independent, that does not change our beliefs about B. Thus, we have that

$$\mathrm{Ind}(B; E, R). \tag{12}$$

We can see a pattern emerging here. Our intuition tells us that the parents of a variable "shield" it from probabilistic influence that is causal in nature. In other words, once I know the value of the parents, no information relating directly or indirectly to its parents or other ancestors can influence my beliefs about it. However, information about its descendants *can* change my beliefs about it, via an evidential reasoning process.

2.2.2 Formal semantics

We are now ready to provide the formal definition of the semantics of a Bayesian network structure.

Definition 2.1: A *Bayesian network structure* G is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$. Let $\mathbf{Pa}_{X_i}$ denote the parents of X_i in G, and $NonDescendants_{X_i}$ denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional independence assumptions, which we denote by $Markov(G)$:

For each variable X_i, we have that

$$\mathrm{Ind}(X_i; NonDescendants_{X_i} \mid \mathbf{Pa}_{X_i}),$$

i.e., X_i is independent of its non-descendants given its parents.

∎

Reviewing the conditional independence assumptions that our intuition dictated in the Alarm network G_{alarm}, Eq. (8) – Eq. (12), we see that they are precisely the assumptions $Markov(G_{alarm})$.

Recall that this definition is our construct; we get to select which independence statements a DAG represents. We chose these because they are intuitive and useful.

2.2.3 The Genetics example

Let us conclude this discussion with one final example, where the Markov assumptions are particularly intuitive. This example is based on a domain, that of genetic modeling, where Bayesian networks (or a subset of the general concept) was developed in parallel (or even before) to their use in other communities.

In the genetic modeling domain, we are trying to model the transmission of a certain property, say blood type, from parent to child. The basic genetics facts are as follows. Each person carries a genetic information that encodes the makeup of each protein in the body. (This information is encoded by DNA strands that are organized in *chromosomes* .) In the example of blood type, the genetic information is encoded by a specific gene. Each person carries two copies of the gene, one inherited from her mother and one inherited from her father. The particular gene for blood type

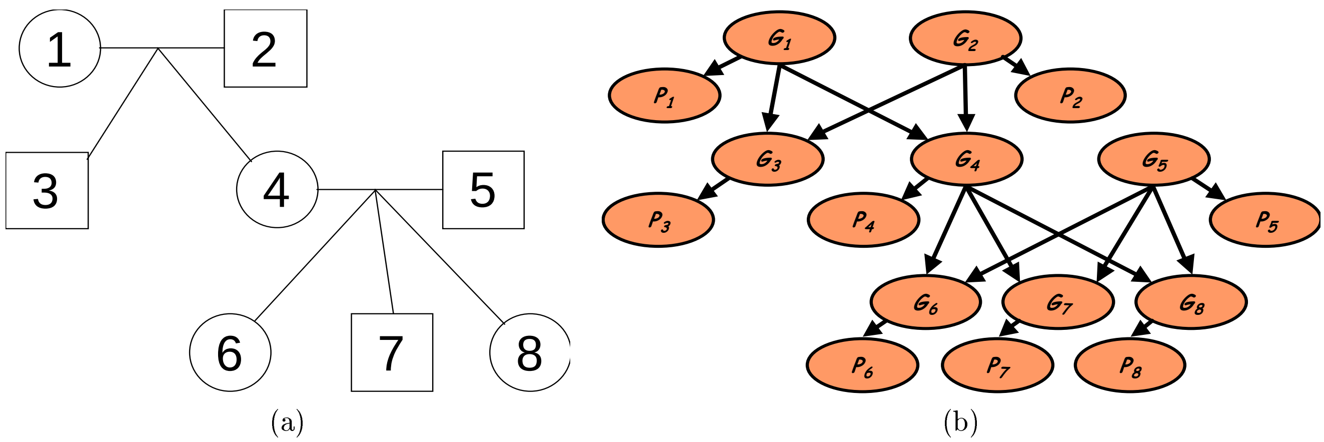


Figure 3: A simple genetics example. (a) an example of a pedigree. (b) the corresponding Bayesian network: G_i denotes the genotype of person i, and P_i denotes the phenotype of person i.

has three possible values (or *alleles* "A", "B", and "O". The *genotype* of a person is a description of the alleles she carries. Thus, in the blood type example, it can be one of the pairs "A/A", "A/B", "A/O", "B/B", "B/O", or "O/O". The observed property, or *phenotype* is a function of the genotype. In the case of blood type, there is a simple function of the genotype: both "A/A" and "A/O" lead to blood type, "A", both "B/B" and "B/O" lead to blood type "B", "A/B" lead to blood type "AB", and "O/O" lead to blood type "O". In more complex cases, the phenotype is a probabilistic function of the genotype. For example, a genotype might make a person more susceptible to a specific form of cancer.

One often needs to reason about the phenotypes of family members. This is important in estimating the risk of an undesirable phenotype based on family history (this estimate might be used for determining the need for more through genetic screen). To perform such reasoning, we need to consider the joint distribution of genotypes and phenotype in the family. Figure 3(a) shows an example of a simple family tree or *pedigree*.

To treat a pedigree we need to consider the dependencies between the phenotypes and genotypes variables. In this example, the independence assumptions arise immediately from the biology. The genotype of a person depends only on the genotypes of her parents. Her phenotype depends only on her own genotype. These are precisely the Markov assumptions in the resulting network structure, shown for a simple family tree in Figure 3(b). The intuition here is clear; for example, a person's genotype is correlated with that of her uncle, but once we know her parents' genotype, the two are independent.

2.3 Graphs and distributions

2.3.1 I-maps

A BN structure G encodes a set of conditional independence assumptions. As for any set of assumptions, some distributions satisfy the assumptions while others do not.

Definition 2.2: We say that a DAG G is an *I-map* (independency mapping) of a distribution P if $P \models Markov(G)$. ▮

Let us illustrate the concept of an I-map on a very simple example. Consider a joint probability space over two independent random variables X and Y. There are three possible graphs over these

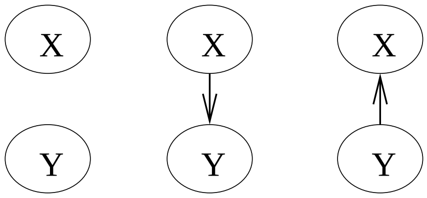


Figure 4: All three possible structures over two random variables

two nodes, as shown in Figure 4. The first encodes the assumption that $\text{Ind}(X;Y)$. The latter two encode no independence assumptions.

Now, consider the following distribution:

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

In this example, X and Y are independent in P, i.e., $P \models \text{Ind}(X;Y)$. Hence, we have that $G_\emptyset$ is an I-map of P. In fact, all three graphs are I-maps of P: $Markov(G_{X\to Y})$ is empty, so that trivially P satisfies all the independencies in it (similarly for $G_{Y\to X}$).

Now, consider the following distribution:

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

In this case, $P \not\models \text{Ind}(X;Y)$, so that $G_\emptyset$ is not an I-map of P. Both other graphs are I-maps of P.

2.3.2 Factorization

A BN structure G encodes a set of conditional independence assumptions; every distribution for which G is an I-map must satisfy these assumptions. This property is the key to allowing the compact factorized representation that we saw in the Alarm example in Section 2.1. The basic principle is the same as the one we used in the Naive Bayes decomposition in Section 1.

Consider an distribution P for which our Alarm BN G_{alarm} is an I-map. We will decompose the joint, and show that it factorizes into local probabilistic models, as in Section 2.1. Consider the joint distribution $P(B,E,A,C,R)$; from the Chain Rule for probabilities we can decompose this joint in the following way:

$$P(B,E,A,C,R) \quad = \quad P(B)P(E \mid B)P(A \mid B,E)P(C \mid A,B,E)P(R \mid C,A,B,E)$$

This transformation relies on no assumptions; it holds for any joint distribution P. However, it is also not very helpful, as the conditional probabilities in the factorization on the right hand side are neither natural nor compact. For example, the last factor requires the specification of 16 conditional probabilities: $P(r^1 \mid c,a,b,e)$ for every assignment of values c,a,b,e.

This form, however, allows us to apply the conditional independence assumptions induced from the BN. Let us assume that G_{alarm} is an I-map for our distribution P. In particular, we have that $P \models \mathrm{Ind}(B; E)$. From that, we can conclude that $P(E \mid B) = P(E)$, allowing us to simplify the second factor on the right hand side. Similarly, we know from Eq. (9) that $P \models \mathrm{Ind}(C; B, E \mid A)$. Hence, $P(C \mid A, B, E) = P(C \mid A)$, allowing us to simplify the second term. Using Eq. (8) in a similar way, we obtain that

$$P(B, E, A, C, R) = P(B)P(E)P(A \mid B, E)P(C \mid A)P(R \mid E). \tag{13}$$

This factorization is precisely the one we used in Section 2.1.

This result tells us that any entry in the joint can be computed as a product of factors, one for each variable. Each factor represents a conditional probability of the variable given it parents in the network. This factorization applies to *any* distribution P for which G_{alarm} is an I-map.

We now state and prove this fundamental result more formally.

Definition 2.3: Let G be a DAG over the variables $X_1, \ldots, X_n$. We say that a distribution P over the same space *factorizes according to* G if P can be expressed as a product

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_{X_i}). \tag{14}$$

This equation is called the *Chain Rule for Bayesian Networks.* The individual factors $P(X_i \mid \mathbf{Pa}_{X_i})$ are called *conditional probability distributions (CPDs)* or *local probabilistic models.* ∎

Definition 2.4: A DAG G and a distribution P that factorizes over G, where P is specified as set of CPDs associated with G's nodes, is called a *Bayesian network.* ∎

We can now prove that the phenomenon we observed for G_{alarm} holds more generally.

Theorem 2.5: *Let G be a DAG over a set of random variables $\mathbf{X}$ and let P be a joint distribution over the same space. If G is an I-map for P, then P factorizes according to G.*

Proof: Assume, without loss of generality, that $X_1, \ldots, X_n$ is an ordering of the variables in $\mathbf{X}$ which is consistent with G, i.e., if X_i is a parent of X_j in G, then $i < j$. As in our example, we first use the chain rule for probabilities:

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

Now, consider one of the factors $P(X_i \mid X_1, \ldots, X_{i-1})$. As G is an I-map for P, we have that $P \models \mathrm{Ind}(X_i; \mathit{NonDescendants}_(X_i) \mid \mathbf{Pa}_{X_i})$. By assumption, all of X_i's parents are in the set $X_1, \ldots, X_{i-1}$. Furthermore, none of X_i's descendants can possibly be in the set. Hence,

$$\{X_1, \ldots, X_{i-1}\} = \mathbf{Pa}_{X_i} \cup \mathbf{Z}$$

where $\mathbf{Z} \subseteq \mathit{NonDescendants}_(X_i)$. From the Markov assumption for X_i and from Exercise **??**, it follows that $\mathrm{Ind}(X_i; \mathbf{Z} \mid \mathbf{Pa}_{X_i})$. Hence, we have that $P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid \mathbf{Pa}_{X_i})$. Applying the same transformation to all of the factors in the chain rule decomposition, the result follows. ∎

Thus, the conditional independence assumptions implied by a BN graph G allow us to factorize the distribution (one that factorized according to G) into small CPDs. The resulting representation can be substantially more compact, particularly for sparse structures. In our Alarm example, the number of independent parameters is ten: one each for $P(B)$ and $P(E)$, four for $P(A \mid B, E)$, and two each for $P(C \mid A)$ and $P(R \mid E)$. The specification of the full joint distribution would require 31 independent parameters. More generally, in a distribution over n binary random variables, the specification of the joint distribution requires $2^n - 1$ independent parameters. If the distribution factorizes according to a graph G where each node has at most k parents, the total number of independent parameters required is less than $n \cdot 2^k$.

Theorem 2.5 shows one direction of the fundamental connection between the conditional independencies encoded by the BN structure and the factorization of the distribution into local probability models: that the conditional independencies imply factorization. The converse also holds: factorization according to G implies the associated conditional independencies.

Theorem 2.6: *Let G be a BN graph over a set of random variables $\mathbf{X}$ and let P be a joint distribution over the same space. If P factorizes according to G, then G is an I-map for P.*

We illustrate this theorem by example, leaving the proof as an exercise (Exercise **??**). Let P be some distribution that factorizes according to G_{alarm}. We need to show that $Markov(G_{alarm})$ hold in P. Consider the Markov independence assumption for the random variable R — $\mathrm{Ind}(R; B, A, C \mid E)$. To prove that it holds for P, we need to show that

$$P(R \mid B, A, C, E) = P(R \mid E).$$

By definition

$$P(R \mid B, A, C, E) = \frac{P(R, B, E, A, C)}{P(B, E, A, C)}$$

By the chain rule for BNs Eq. (13), the numerator is equal to $P(B)P(E)P(A \mid B, E)P(C \mid A)P(R \mid E)$. By the process of marginalizing over a joint distribution, we have that the denominator is:

$$\begin{aligned} P(B, E, A, C) &= \sum_R P(B, E, A, C, R) \end{aligned}$$

$$\sum_R P(B)P(E)P(A \mid B, E)P(C \mid A)P(R \mid E)$$

$$\begin{aligned} &= P(B)P(E)P(A \mid B, E)P(C \mid A) \sum_R P(R \mid E) \\ &= P(B)P(E)P(A \mid B, E)P(C \mid A) \end{aligned}$$

where the last step is because $P(R \mid E)$ is a distribution over values of R, and therefore it sums to 1. We therefore have that

$$\begin{aligned} P(R \mid B, A, C, E) &= \frac{P(R, B, E, A, C)}{P(B, E, A, C)} \\ &= \frac{P(B)P(E)P(A \mid B, E)P(C \mid A)P(R \mid E)}{P(B)P(E)P(A \mid B, E)P(C \mid A)} \\ &= P(R \mid E). \end{aligned}$$

3 Independencies in Bayesian networks

Dependencies and independencies are key properties of a distribution, and are crucial for understanding its behavior. As we will see, independence properties are also important for answering queries: they can be exploited to substantially reduce the computation cost of inference. Therefore, it is important that our representation make these properties clearly visible, both to a user and to algorithms that manipulate the BN data structure. When can we "read off" dependencies and independencies directly from the graph? This question decomposes into two parts:

- Which additional independencies follow from the $Markov(G)$ assumptions that are implied by the definition of the BN semantics? These independencies will hold for *every* distribution P over G.
- Can there be additional independencies that hold for specific distributions P over a BN G but that do not follow from G.

3.1 Independencies in G: d-separation

In general, we can ask the following question: Given that a distribution P satisfies some set of independence statements $\mathcal{I}$, what other independence properties is it guaranteed to have? I.e., what can we assert about the distribution knowing only that it satisfies $\mathcal{I}$. As we said, some other independencies follow trivially, e.g., $\mathrm{Ind}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})$ implies $\mathrm{Ind}(\mathbf{Y};\mathbf{X} \mid \mathbf{Z})$. People have spent a lot of effort trying to come up with other axioms for this logic. Unfortunately, the logic of independence statements is not *recursively enumerable*: there is no program that can enumerate the independence properties that are implied by a set of $\mathcal{I}$ of independence statements. In other words, we cannot hope to build a general purpose engine that determines whether an independence statement is implied by a set of independence statement $\mathcal{I}$.

We, however, are interested independencies that are implied by a Bayesian network structure. In other words, we are interested in independencies that are implied by sets of the form $Markov(G)$ for some DAG G. Such sets of independencies are more restricted, and thus we might hope that they allow us to perform reasoning.

3.1.1 Defining d-separation

The intuition for analyzing when an independence $\mathrm{Ind}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})$ must happen in a BN $G$is by considering the complementary property: is it *possible* that $\mathbf{X}$ can influence $\mathbf{Y}$ given $\mathbf{Z}$. I.e., where can probability dependency potentially happen. If there is any way in which $\mathbf{X}$ can influence $\mathbf{Y}$ given $\mathbf{Z}$, then $\mathrm{Ind}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})$ property cannot hold for all distributions. Hence, we will consider the potential flow of influence between variables. When can observing the value of X change our beliefs about Y?

Direct connection One obvious case is when X is a parent of Y. For example, clearly knowing that there was a burglary changes my beliefs about the alarm going off.

Another case is when X is a child of Y. For example, clearly our prior probability $P(b_1)$ is much lower then my posterior probability given the evidence a_1. Thus, B and A are not independent. (Not surprising, since independence is a symmetric notion.)

Thus, two nodes that are directly connected can influence each other.

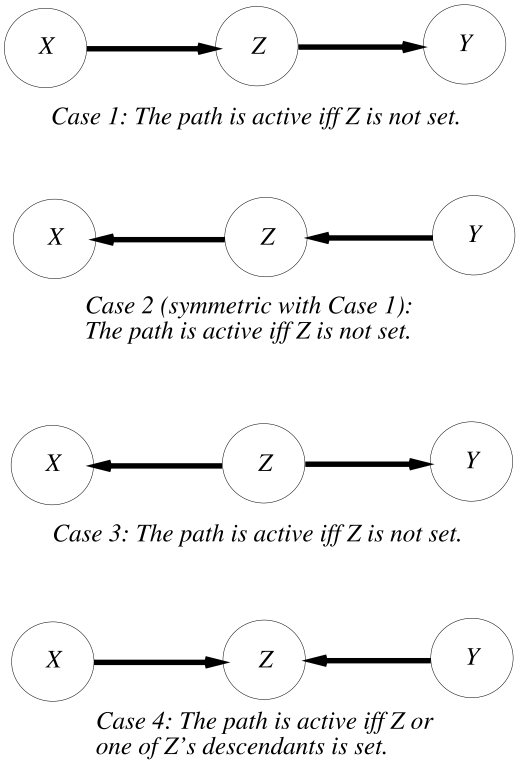


Figure 5: A taxonomy of 2-edge paths

Indirect connection, no evidence Can X influence Y indirectly, via a third node Z? Intuitively, that depends on whether the intervening node Z is observed. Let's begin with the case where no nodes are observed, i.e., we are interested in statements of the form $\text{Ind}(X; Y \mid \emptyset)$. Again, let's check the various cases, see Figure 5:

Case 1: An indirect causal effect. To gain intuition, let's return to the Alarm example. The path $B \rightarrow A \rightarrow C$ is an example of an indirect causal path. Intuitively, if we observe Burglary, the probability of Alarm goes up, and so does the probability of Call. Thus, we would expect influence to be able to flow over a path such as this. Indeed, if compute the probability using the parameterization used in the example, we see that $P(c^1 \mid b^1) > P(c^1)$, as expected.

Case 2: An indirect evidential effect. Clearly, the same effect happens here. This is not surprising, as dependence is symmetrical: if $\text{Ind}(X; Y)$ does not hold then $\text{Ind}(Y; X)$ does not hold either. Intuitively, this is clear. In our Alarm example, we would expect that if someone calls us, then the probability of a burglary go up. Again, this can be verified numerically for our distribution P.

Case 3: A common cause. In the Alarm example, we have such a path in $R \leftarrow E \rightarrow A$. Intuitively, a radio report of an earthquake increases my belief in an earthquake, and so would increase my probability of alarm. Thus, I would expect influence to flow.

Case 4: A common effect. This type of structure is also called a *v-structure*. Does this pattern—influence flowing via a third node—continues to hold in the final case? No. In our example, recall that E and B are marginally independent. Thus, we know that influence cannot flow from E to B via A.

Indirect connection, with evidence How does evidence affect independence? Let's go back over our four cases, and consider the effect of observing Z on the flow of influence. In our Alarm example, we know the answer: the Markov assumptions imply that C is independent of B given its parents, i.e., A. So we have that $\text{Ind}(C; B \mid A)$. Thus, evidence changes the connection between these two variables: without this evidence, one could influence the other, now they cannot. It turns out that this is true in general for case 1 (and case (2)). In case 3, the Markov assumptions again imply that $\text{Ind}(C; B \mid A)$.

Case 4 is more interesting. What happens to my statement of $\text{Ind}(E; B)$ if we observe A? As we discussed above, if we observe the alarm going off, the probability of both burglary and earthquake goes up. I.e., $P(b^1 \mid a^1) > P(b^1)$. But observing e^1 in this context reduces the probability in b^1. In other words, although B and E were independent *a priori*, they are dependent given A. Hence, unlike the previous cases, this is a situation where the presence of evidence serves to enable rather than disable a possible influence path.

Let us consider one final case, which is a variant of case 4. Assume that we do not observe alarm, but we do observe that the neighbor called. Intuitively, the same phenomenon happens. The phone call is indirect evidence for the burglary. Thus, we would expect:

$$P(b^1) < P(b^1 \mid c^1, e^1) < P(b^1 \mid c^1) < P(b^1 \mid a^1)$$

Indeed, this is what the computation would lead to in the example above.

When influence can flow from X to Y via Z, we say that the path $X \rightarrow Z \leftarrow Y$ is *active*. The results of our analysis for active 2-edge paths are summarized in Figure 5.

General case More generally, along longer paths, we have that influence can flow along a path $X_1 — \ldots — X_n$ if it can flow from each X_i to X_{i+2} via X_{i+1}.

We can summarize this intuition in the following definition:

Definition 3.1: Let G be a DAG, and $X_1 — \ldots — X_n$ an undirected path in G. Let $\mathbf{Z}$ be a subset of nodes of G. The path $X_1 — \ldots — X_n$ is *active* given evidence $\mathbf{Z}$ if

- Whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants in in $\mathbf{Z}$;
- no other node along the path is in $\mathbf{Z}$.

Let X and Y be two nodes in G. We say that X and Y are *d-separated given* $\mathbf{Z}$, denoted $d\text{-}sep_G(X; Y \mid \mathbf{Z})$, if there is no active path between X and Y given $\mathbf{Z}$. ∎

As we will see, the notion of d-separation captures the notion of guaranteed independence — independence that is required to hold in any distribution over this network.

Let us look at a few examples in our Alarm example.

- R and B are d-separated, because the v-structure blocks the only path.
- R and B given C are not d-separated. Observing C made the path active.
- R and B given E and C are d-separated. Observing E prevents the influence from flowing.

How do we check d-separation of X and Y given $\mathbf{Z}$? Enumerating all paths is expensive. However, there is a much more efficient algorithm, that requires only linear time in the size of the graph.

- Phase I: We traverse the graph bottom up, from the leaves to the roots. We mark all nodes that are in $\mathbf{Z}$ or that have descendants in $\mathbf{Z}$. Intuitively, these nodes will serve to enable v-structures.
- Phase II: We traverse breadth-first from X to Y, backtracking if we get to a blocked node. A node is blocked if: (a) it the is the "middle" node in a v-structure and unmarked in phase I, or (b) is not such a node and is in $\mathbf{Z}$. If our depth-first search gets us from A to B, then there is an active path between them.

3.1.2 d-Separation and independence

So far, our definition of d-separation has been based on our intuitions regarding flow of influence, and on our one example. As yet, we have no guarantee that this analysis is "correct". Perhaps there is a distribution over the BN where X can influence Y despite the fact that all paths between them are blocked.

Hence, the first property we want to ensure is *soundness*: the d-separation procedure only states independencies if they really hold. Fortunately, this property holds:

Theorem 3.2: *If a G is an I-map of P and* $\text{d-sep}_G(X;Y \mid \mathbf{Z})$*, then P satisfies* $\text{Ind}(X;Y \mid \mathbf{Z})$.

In other words, any independence reported by d-separation is satisfied by the underlying distribution.

Let us show that this theorem holds for a simple example. Consider our Alarm example, and let us show that B and E are, indeed, independent. Consider any distribution P that factorizes over G.

$$
\begin{aligned}
P(B,E) &= \sum_{A,C,R} P(B,E,A,C,R) \\
&= \sum_{A,C,R} P(B)P(E)P(A \mid B,E)P(C \mid A)P(R \mid E) \\
&= P(B)P(E) \sum_{A,C,R} P(A \mid B,E)P(C \mid A)P(R \mid E) \\
&= P(B)P(E) \sum_{A} [P(A \mid B,E)(\sum_{C} P(C \mid A))(\sum_{R} P(R \mid E)) \\
&= P(B)P(E) \sum_{A} [P(A \mid B,E)(1)(1)] \\
&= P(B)P(E)(1) \\
&= P(B)P(E)
\end{aligned}
$$

3.1.3 Minimal I-maps

As a first attempt, let us consider the strongest converse we can imagine: for any graph G which is an I-map for P, and any independence $\text{Ind}(X;Y \mid \mathbf{Z})$ that holds in P, the assertion $d\text{-}sep_G(X;Y \mid \mathbf{Z})$ must hold in G. This statement is clearly false, because we can represent distributions where there are "accidental" independencies. For example, consider a distribution P over two variables A and B, where A and B are independent. One possible I-map for P is the network $A \to B$. For example, we can set the CPD for B to be

B	$P(B \mid a^0)$	$P(B \mid a^1)$
b^0	0.4	0.4
b^1	0.6	0.6

This clearly violates the proposed "theorem", because the graph G is an I-map distribution P. Nevertheless, there are independencies that hold for this distribution that do not follow for d-separation. These are not independencies, however, that we can read off the network structure. Clearly, this is a general phenomenon: the fully connected graph implies no conditional independence assumptions, and therefore is an I-map for every distribution.

Of course, these counter-examples to our proposed theorem are not very interesting. The graphs that we proposed are clearly unrepresentative of the distribution, in that there are trivially redundant edges. This intuition leads to the following definition:

Definition 3.3: A graph G is a *minimal I-map* for P if it is an I-map for P, and if the removal of even a single edge from G renders it not an I-map. ∎

Recall that Definition 2.4 defines a Bayesian network to be a distribution P that factorizes over G, thereby implying that G is an I-map for P. It is standard to restrict the definition even further, by requiring that G be a minimal I-map for P.

How do we construct a minimal I-map for a given distribution P? The proof of the factorization theorem, Theorem 2.5, gives us an answer. We begin by ordering the nodes somehow, $\{X_1, \ldots, X_n\}$. The construction, as in the proof of the theorem, is such that X_i can only be a parent of X_j in G if $i < j$. We insert the nodes into G one by one, in order, so as to guarantee that the resulting graph is an I-map, but also a minimal one.

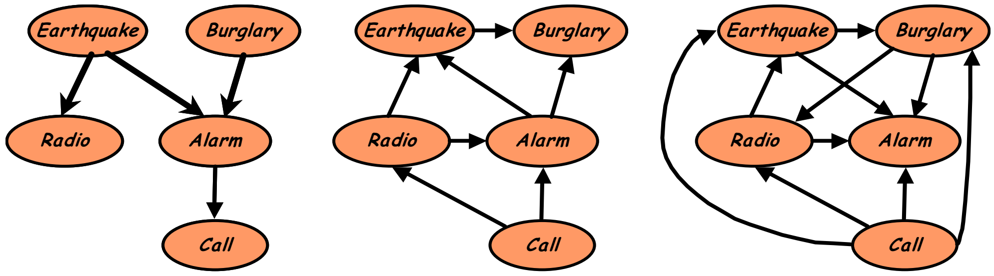
Specifically, consider the insertion of X_i into G. We know that it can depend only $X_1, \ldots, X_{i-1}$, which are already in the G. We pick some minimal subset $\mathbf{U}$ of $X_1, \ldots, X_{i-1}$ to be X_i's parents. More precisely, we require that $\mathbf{U}$ satisfy $\mathrm{Ind}(X_i; \{X_1, \ldots, X_{i-1}\} \mid \mathbf{U})$, and that no node can be removed from $\mathbf{U}$ without violating this property. We then introduce an edge from each node in $\mathbf{U}$ to X_i, and continue to X_{i+1}. Then we connect X_i to its newly selected parents and continue. It is easy to verify that this process yields a minimal I-map.

Is the minimal I-map unique? Not in general. Consider, for example, a case where two variables A and B are logically equivalent, i.e., our distribution P only gives positive probability to instantiations where A and B have the same value. Now, consider a node C that is correlated with A. Clearly, we can choose either A or B to be a parent of C, but having chosen the one, we cannot choose the other without violating minimality. Hence, the minimal parent set $\mathbf{U}$ in our construction is not necessarily unique. However, one can show that, if the distribution is *positive*, i.e., if for any instantiation $\mathbf{x}$ to all the network variables $\mathbf{X}$ we have that $P(\mathbf{x}) > 0$, then the choice of parent set is unique. This means that the I-map is unique *relative to the chosen ordering of the nodes.*

Still, a different ordering could result in a different minimal I-map. The economy and intelligibility of the belief network resulting from the above process can depend crucially on the ordering of the nodes.

Example 3.4: To see this phenomenon, let's return to the Alarm example, and the distribution P we defined in that example. Consider three orderings:

- E, B, A, R, C. In this case, we get our original network G_{alarm}. See Figure 6(a).
- C, R, A, E, B. We now apply the procedure. Clearly, C must be root, since it is first in the ordering. Next, we consider R. As we can check that C and R are *not independent*, we add an arc $C \rightarrow R$. Next, we consider A given C and R. Again, it is clear that A depends on both variables, and that neither of them renders the other independent of A. Thus, we set both C and R as parents of A. Next, we consider E. In this case we can prove that C is



E, B, A, R, C
(a)

C, R, A, E, B
(b)

C, B, R, E, A
(c)

Figure 6: Three minimal I-maps for the Alarm domain.

independent of E given A (and R). Thus, we set A and R to be the parents of E. Finally, when we consider B, we see that A and E separate it from the rest of the variables.

The resulting network is shown in Figure 6(b). As we can see the resulting graph is not as sparse.

- C, B, R, E, A. In this case, the resulting graph is much more complicated, see Figure 6(c).

∎

We can extract two conclusions from this discussion. First, to generate a good I-map for a target distribution, we cannot simply apply the construction algorithm naively. We will return to this issue when we discuss our two main methods for constructing Bayesian networks: knowledge engineering and learning.

The second conclusion is that, even when we have a minimal I-map for a distribution P, we are not guaranteed that all independencies present in the distribution are reflected in the graph structure. Indeed, in the last example, we saw two minimal I-map for the original distribution P, but in both cases, there are a lot of independencies in P that are not reflected in the structure. The minimality of the I-map does not help, because these independencies are not ones that we can get at by eliminating edges.

3.2 Completeness of d-separation

The above discussion shows that the minimality of the I-map is not enough to validate our proposed completeness "theorem". However, a weaker notion of completeness holds.

The d-separation procedure is complete in the following sense: there are no independencies that are *guaranteed* to be true that d-separation fails to detect. In other words, if there is some independence assertion that d-separation in G does *not* conclude, then there is at least one distribution P, for which G is an I-map and in which the dependence holds.

If we consider the examples of Figure 6, although there are independencies in our alarm distribution P that d-separation in the DAG (b) will never detect, there is some other distribution P' such that this DAG is an I-map of P' and it reflects the independencies derived from d-separation over the DAG.

In fact, this statement holds for almost all distributions that factorize over a particular DAG. More generally:

Theorem 3.5: *Suppose that* $\text{d-sep}_G(X; Y \mid \mathbf{Z})$ *does not hold. Then, the set* $\{P : G$ *is an I-map of* $P, P \not\models \text{Ind}(X; Y \mid \mathbf{Z})\}$ *has measure zero in* $\{P : G$ *is an I-map of* $P\}$.

3.2.1 Perfect maps

So we have seen that, for a given distribution, there are some minimal I-maps that are much smaller than others. Indeed, some minimal I-maps miss out on representing some independencies entirely. But, for our distribution, there was a graph (G_0) that captured the independencies perfectly.

Definition 3.6: A DAG G is a *perfect map (P-map)* for a distribution P when $\mathrm{Ind}(X; Y \mid \mathbf{Z})$ holds for P if and only if $\textit{d-sep}_G(X; Y \mid \mathbf{Z})$. ∎

Does every distribution have a perfect map? Unfortunately, the answer is no, and for several reasons.

Example 3.7: Consider a distribution P over 3 random variables X,Y,Z. Assume that X and Z are independent and uniform, i.e., $P(x^0) = P(x^1) = 0.5$. Assume that

$$P(z \mid x, y) = \begin{cases} 2/3 & z = x \oplus y \\ 1/3 & z \neq x \oplus y \end{cases}$$

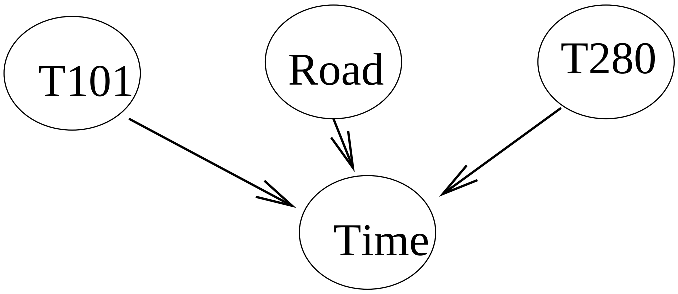
A minimal I-map for this distribution is the DAG $X \rightarrow Z \leftarrow Y$.

However, a simple calculation shows that the joint distribution is

$$P(x, y, z) = \begin{cases} 2/12 & x \oplus y \oplus z = 0 \\ 1/12 & x \oplus y \oplus z = 1 \end{cases}$$

Thus, by symmetry, we have that $X \rightarrow Y \leftarrow Z$ and $Z \rightarrow X \leftarrow Y$ are also I-maps for this distribution. In fact, in this example P satisfies all possible pairwise independencies $\mathrm{Ind}(X; Y)$, $\mathrm{Ind}(X; Z)$, $\mathrm{Ind}(Y; Z)$. Thus none of the three minimal I-maps are perfect maps. ∎

Example 3.8: Consider the following example: I can either take 101 or 280 to go from home to the office; the decision corresponds to a random variable *Road*. There is a random variable for traffic on both roads: *T101* and *T280*. My travel time depends on which road I take, and the travel on that road. What is an I-map?



Is this a perfect map? Clearly, not for independencies that are not at the variable level. In particular, we have that $\mathrm{Ind}(\textit{T101}; \textit{Time} \mid \textit{Road} = 280)$. Clearly, a BN cannot capture independencies of this type, since it only captures independence at the level of variables. (We'll come back to this issue later.) However, this is not the only problem: we also have that $\mathrm{Ind}(\textit{T101}; \textit{T280} \mid \textit{Road}, \textit{Time})$! First, note that this independence is not implied by d-separation, because the v-structure is enabled. However, we can do the reasoning by cases. If $\textit{Road} = 101$, then there is no influence from *T280* to *Time* — intuitively, the edge disappears; thus, there is no influence between *T101* and *T280*. The symmetric case happens in the other direction. This is an example, like above, of structure at the CPD-level that doesn't translate into the graphical structure. ∎

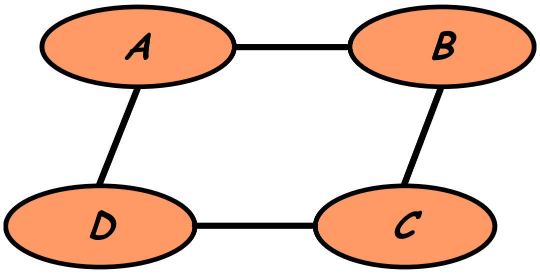


Figure 7: A simple Markov network.

3.3 Equivalent maps

Even if there exists a perfect map for a distribution, there may be more than one. Consider, for example, a distribution P which satisfies only the conditional independence $\text{Ind}(X;Y \mid Z)$. In this case, we have three possible structures, all of which are perfect maps: the first three in Figure 5.

Definition 3.9: Let $\mathbf{X}$ be a set of random variables. Two graphs G_1 and G_2 over $\mathbf{X}$ are *I-equivalent* if they induce precisely the same set of d-separation properties. ∎

Thus, the set of all graphs over $\mathbf{X}$ can be partitioned into a set of mutually exclusive and exhaustive *I-equivalence classes*, which are the set of equivalence classes induced by the I-equivalence relation.

[ADD DISCUSSION OF EQUIVALENCE RELATIONS]

4 Markov Networks

So far we have focused on Bayesian networks (directed graphical models). We now consider an alternative representation that is based on undirected graphs. This representation is useful both on its own, and as an alternative way of understanding inference in directed graphical models.

4.1 Independencies in Markov networks

Example 4.1: Consider the following example. We have a network of computers, such that computer i and j can directly communicated if there is a physical communication channel between them. If the two computers are not directly adjacent, then communications between them are routed through computers some path between them. Now suppose there is a computer virus that infects routers. In some situations the virus can infect a packet, and then it infects the first computers that handles the packet. Thus, the virus can propagate by hoping from node in the network to adjacent ones. Moreover, once a computer is infected by a virus it will remain so. Finally, each computer can be randomly infected by its users. However, since the users of each computer are different, these events are independent.

Consider the network topology shown in Figure 7. In this topology the all packets from A to C must be transmitted by either B and D. Thus, if we know that neither B nor D are infected, then the probability that A is infected and C is infected are independent of each other—since neither of them could have infected the other. Now suppose that we know that B is infected and D is not. In this case, the probability that A was infected by B is same whether C infected B or not. Thus, we conclude that in this computer network we have $\text{Ind}(A;C \mid B,D)$. Similar arguments show that $\text{Ind}(B;D \mid A,C)$ should hold in this network. ∎

We now generalize this example to arbitrary undirected graphs.

Definition 4.2: A *Markov network structure* H is an undirected graph whose nodes represent random variables $X_1, \ldots, X_n$. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint subsets of nodes in the graph. We say that $\mathbf{Z}$ *separates* $\mathbf{X}$ and $\mathbf{Y}$ in H, denoted $sep_H(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$, if every path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$ must pass through a node in $\mathbf{Z}$. ∎

Intuitively, the Markov network H encodes the following set of independence assumptions: For each $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ such that $sep_H(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ holds, we have the assumption that $\text{Ind}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$. More formally,

Definition 4.3: We say that a Markov network H is an *I-map (independency mapping)* for a distribution P if whenever $sep_H(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ holds, we have that $P \models \text{Ind}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$. ∎

As usual, the complete graph implies no independence assumptions, and is therefore an I-map for any distribution. We can therefore define, as for Bayesian networks, the notion of a minimal I-map:

Definition 4.4: We say that a Markov network H is a *minimal I-map* for a distribution P if it is an I-map, and if the removal of any edge from H renders is not an I-map. ∎

How can we construct a minimal I-map for a distribution P? Some edges clearly have to be in H. If we do not connect a pair of nodes A and B via a direct edge, then we will be able to separate them somehow, e.g., by instantiating every other node in the graph. Hence, if $P \not\models \text{Ind}(A; B \mid \mathbf{X} - \{A, B\})$ (where $\mathbf{X}$ is the set of nodes in H), then we must connect A and B using a direct edge. This process adds edges it would clearly be bad to omit; any I-map must contain at least these edges. Surprisingly, in most cases, this is enough. If we put only these edges in the graph, we get an I-map.

Theorem 4.5: *Let P be a* positive *distribution over $\mathbf{X}$, i.e., one where for all $\mathbf{x} \in Val(\mathbf{X})$, $P(\mathbf{x}) > 0$. Consider a Markov network H whose vertices are $\mathbf{X}$ and whose edges $\mathcal{E}$ are as follows:*

$$(A, B) \in \mathcal{E} \quad \Leftrightarrow \quad P \not\models \text{Ind}(A; B \mid \mathbf{X} - \{A, B\})$$

Then H is the unique minimal I-map for P.

If we show that this graph is, in fact, an I-map, the minimality is ensured: we have already shown that any I-map must contain these edges. Similarly, we are also guaranteed that this minimal I-map is unique.

The positivity assumption is critical for this result. Consider, for example, a distribution over the four binary variables A, B, C, D, which has $P(a^1, b^1, c^1, d^1) = P(a^0, b^0, c^0, d^0) = 1/2$, and P is 0 everywhere else. In this case, this algorithm returns the empty network: for example, when selecting edges for A, it is the case that $P \models \text{Ind}(A; B \mid C, D)$ and similarly for the other two nodes. Clearly, the resulting network is not an I-map for P.

Definition 4.6: Let P be a positive distribution. The set of nodes that are the immediate neighbors of a node X_i in the minimal I-map H for P is called the *Markov blanket* of X_i. ∎

It is clear that a node is independent of the rest of the network given its Markov blanket.

What independencies does a Markov network capture exactly. If H is a minimal I-map for P, does it necessarily reflect its independencies exactly? More formally:

Definition 4.7: We say that H is a *perfect map* for P if for any $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, we have that

$$sep_H(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) \Longleftrightarrow P \models \mathrm{Ind}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$$

∎

Does every distribution have a perfect map? Clearly, the answer is no. Consider a distribution arising from a three-node Bayesian network with a v-structure, e.g., the distribution induced in the Alarm example over the nodes *Burglary*, *Earthquake*, and *Alarm*. In the Markov network for this distribution, we must clearly have an edge between B and A and between E and A. Can we omit the edge between B and E? No, because we do not have that $\mathrm{Ind}(B; E \mid A)$ holds for the distribution; rather, we have the opposite: B and E are *dependent* given A. Therefore, the only minimal I-map for this P is the fully connected graph, which does not capture the marginal independence $\mathrm{Ind}(B; E)$ that holds in P.

This example illustrates what happens when we construct the minimal Markov network I-map for a Bayesian network distribution. Consider any pair of nodes X and Y that are parents of the same node Z. When we condition on Z, X and Y become correlated; hence, it is not the case that $\mathrm{Ind}(X; Y \mid \mathbf{X} - \{X, Y\})$. Thus, in the minimal Markov network I-map, there must be an edge between X and Y.

Definition 4.8: The *moral graph* $H[G]$ of a Bayesian network structure G is the undirected graph over the same set of nodes that contains an undirected edge between X and Y if: (a) there is a directed edge between them (in either direction), or (b) X and Y are both parents of the same node.[4] ∎

It is easy to see that this construction has the desired property:

Proposition 4.9: *The moral graph $H[G]$ of a Bayesian network structure G is the minimal Markov network that satisfies all of the independence assumptions that are induced by G.*

This transformation allows us to define the Markov blanket of a variable in the Bayesian network. Recall that the Markov blanket of X_i is the minimal set of variables that render X_i conditionally independent from the rest of the variables. Since the moral graph is the minimal Markov network that is an I-map of the distribution, we see have that the Markov blanket in the moral graph is the also the Markov blanket in the Bayesian network. It is easy to verify that this blanket includes X_i parents, children, and "spouses" (the parents of X_i children).

4.2 Factorization for Markov networks

Like in a Bayesian network, the independence assumptions induced by a Markov network allow us to factorize the distribution into smaller pieces. In this case, the factorization is somewhat different.

We define a *complete subgraph* of H to be a subset of the vertices of H where each vertex is connected by an edge to all other vertices in the subgraph. Naturally, a complete subgraph has no independencies within the subgraph. We define a *clique* to be a maximal complete subgraph, i.e., one to which no additional vertices can be added.[5] Any graph can be decomposed into cliques (which are not necessarily disjoint). For example, the Markov network of Figure 7 as four cliques: $\{A, B\}$, $\{B, C\}$, $\{C, D\}$, $\{A, D\}$.

The decomposition of a Markov network into cliques induces a similar decomposition on the distribution.

[4]The name *moral graph* originated because of the supposed "morality" of marrying the parents of a node...

[5]Note that this is nonstandard usage of the word clique. Usually, a clique is any complete subgraph, not necessarily a maximal one. However, it is the standard interpretation of the word within the graphical models community.

Definition 4.10: Let $\mathbf{Y}$ be a set of random variables. We define a *factor* to be a function from $Val(\mathbf{Y})$ to $I\!R$. ∎

Definition 4.11: Let H be a Markov network, and let $\mathbf{C}_1, \ldots, \mathbf{C}_k$ be the cliques in H. A distribution P *factorizes* over H if there exists factors $f_1, \ldots, f_k$ over $\mathbf{C}_1, \ldots, \mathbf{C}_k$ such that

$$P(X_1, \ldots, X_n) = \alpha f_1(\mathbf{C}_1) \times f_2(\mathbf{C}_2) \times \cdots \times f_k(\mathbf{C}_k)$$

where α is a normalizing constant. The factors f_i are called *clique potentials.* ∎

We can now state an analogous theorem to the one we stated for Bayesian networks:

Theorem 4.12: *Let P be a positive distribution over $X_1, \ldots, X_n$ and H a Markov network over the same variables. Then P factorizes over H if and only if H is an I-map for P.*

Markov networks are very useful for vision applications. There, for example, we have nodes associated with pixels in the image, and edges that connect neighboring pixels. The clique potentials are used to represent the correlations between the intensity values of neighboring pixels. In general, however, the parameterization of Markov networks is not very intuitive, as they do not correspond either to probabilities or to conditional probabilities. As a consequence, the numbers are hard for people to understand, making them hard to elicit; furthermore, it turns out that, for the same reason, these parameters are harder to estimate from data. As a consequence, directed graphical models are more commonly used in most applications. Nevertheless, undirected graphical models will turn out to be a useful internal data structure in certain inference algorithms.

5 Knowledge engineering

Above, we interpreted the process of constructing a BN in as the process of going from a given distribution to a BN. Real life is not like that. We have a vague model of the world, and we need to crystallize it into a BN.

5.1 Picking variables

We must define our variables very precisely. The *clarity test* is a good way of figuring out whether they are well-defined enough: Assume that we are a million years after the events described in the BN; can an omniscient being, one who saw everything, determine the value of the variable? For example, consider a *Weather* variable, with a value *sunny*. To be absolutely precise, we must define: where do we check the weather, at what time, what fraction of the sky must be clear in order for it to be sunny. For a variable such as *Heart-attack*, we must specify how large does the heart attack have to be, in what period of time does it have to happen, ...By contrast, a variable such as *Risk-of-heart-attack* is meaningless.

If we are not careful, we'll have a hard time making sure that evidence observed and conclusions made are coherent.

It is also important to pick reasonable values for our variables. In particular, if our partition is not fine enough, conditional independence assumptions may be false. For example, we may want to construct a model where we have a person's cholesterol level, and two tests that are conditionally independent given that. We define "normal" to be up to 200, and "high" to be above that. But it may be the case that both tests are more likely to fail if the person's cholesterol is marginal

(200–240). In this case, the assumption of conditional independence given the value (high/normal) of the cholesterol test is false. It is only true if we add a "marginal" value.

Generally speaking, we want our model to contain variables that we can potentially observe or that we may want to query. However, sometimes we want to put in *hidden variables* that are neither. Why would we want to do that? Let's go back to the cholesterol test. In order for the answers to be accurate, the person has to not eat since 10pm the following evening. If the person eats (having no willpower), the results are consistently off. We don't really care about this variable, nor can we observe it. However, without it, all of the different cholesterol tests become correlated. In order to avoid graphs where all the tests are correlated, it's better to put in this additional hidden variable, rendering them conditionally independent given the true cholesterol level and the person's willpower.

On the other hand, it's not necessary to add every variable that might be relevant. For example, in our Alarm example, perhaps our neighbor is more likely to call when an ambulance with a siren passes by. Is this fact important to represent? The probabilities already account for the fact that he might be unreliable. It is not worthwhile including this variable if I cannot observe it or something related to it.

5.2 Picking structure

Causal structures tend to work well. Either because of some real locality of influence in the world, or because of the way people perceive the world, causal connection graphs tend to be sparser. Note: causal is in the world, not in our inference process. For example, in an automobile insurance network, it is tempting to put *Previous-accident* as a parent of *Good-driver*, because that is how the insurance company thinks about the problem. This is not the causal order in the world, because being a bad driver causes previous (and future) accidents. In principle, there is nothing to prevent us from directing the edges in this way. However, as we saw in the minimal I-maps for the Alarm example, a non-causal ordering often requires that we introduce a lot of additional edges to account for induced dependencies.

Another issue is that, in real life, we often have to make compromises. Models with all of the variables and edges may be too complicated to reason with. For example, is MSNBC's web-based system for pregnancy and child care, they originally began with a model that included a node for every possible disease. The network turned out to be too complex to run in real-time on a shared web-server. Therefore, they compromised by aggregating all of the disease nodes into a single *Primary-disease* node. The first one is much more realistic, but the second is not too bad an approximation in practice, and is much faster computationally.

In general, our model is always an approximation. Everything is dependent on everything else (butterfly in China). For example, in our alarm example, we assumed that the neighbor calling was unrelated to the earthquake. This assumption might not be true: if there is an earthquake, the neighbor might be less likely to call. In general, we can put a whole bunch of weak but existing influences into the model, but we end up with a useless one. So we make reasonable approximations.

5.3 Picking numbers

Numbers are very hard to elicit from people. They really dislike committing to an exact estimate. There have been several approaches developed for doing this process, e.g., something called a probability wheel, asking people how they would compare the probability in question to certain predefined lotteries or to other things, etc. Nevertheless, it is a very long, drawn out process.

The good news: empirical evidence shows that the answers given by a BN are not all that sensitive to many aspects of the probabilities. In general, 0.53 versus 0.54 doesn't make much difference. What does matter?

- zeros: you simply can't condition away a zero probability event no matter how hard you try. In the first lecture, I mentioned the Pathfinder system, the first real BN-based expert system. In one of the first versions of the system (one based on naive Bayes), 10% of the cases were misdiagnosed because the expert had given a probability of zero to an unlikely but possible event. Realistically, very few things (except definitions) have probability zero, and we must be careful in assigning zeros.
- relative values: the qualitative behavior of the system will be affected by the relative sizes of $P(x \mid \mathbf{y})$ for different values $\mathbf{y}$ of $\mathbf{Pa}_X$. For example, it's important that my neighbor is more likely to call given an alarm than not given an alarm. It's also important that the alarm is more likely to be set off by a burglary than by an earthquake.
- orders of magnitude: an extreme case of relative sizes is when one thing is an order of magnitude more likely than another. Differences like this can radically change the behavior of the system.

Fortunately, BNs give you a good tool for deciding how much a given number affects the outcome. This process, called sensitivity analysis, can be applied as a byproduct of standard inference techniques.

Finally, we note that causal models make it easier to elicit CPDs, since they tend represent very natural quantities. For example, the probability that the alarm will go off in case of a burglary. The probabilities of burglary, earthquake can be obtained from the neighborhood police station and the USGS. In general, the causal direction allows us to isolate the sensor, using its specs to determine the CPTs. In fact, one of the standard ways of doing knowledge engineering is to build the network structure in a way that makes it easy to elicit probabilities. This process is called "extending the conversation". You start out with a variable such as "lung tumor". You ask the expert if he can give beliefs over that. He says "it depends", you ask on what. The appropriate parent variables, e.g., lung cancer are added. This continues until root nodes are reached, i.e., until expert can give you probabilities. E.g., add smoking and exposure to toxics, then add age and gender.