Chapter 3

Building models

Bayesian networks create a very efficient language for building models of domains with inherent uncertainty. However, as can be seen from the calculations in Section 2.4, it is a tedious job to perform evidence transmission even for very simple Bayesian networks. Fortunately, software tools which can do the calculation job for us are available. Several commercial products exist containing both an editor for Bayesian networks and a runtime module which takes care of evidence transmission. In the rest of this book we assume that the reader has access to the HUGIN system provided by the diskette attached to the book, or to any other Bayesian network programming environment.

Therefore we can start by concentrating on how to use Bayesian networks in model building and defer a presentation of the methods for probability updating to Chapter 4.

In Section 3.1 we examine, through three examples, the considerations when determining the structure of a Bayesian network model. Section 3.2 gives examples of estimation of the conditional probabilities. The examples cover theoretically well-founded probabilities as well as probabilities taken from data bases and purely subjective estimates. Section 3.3 gives several modelling tricks to use when the amount of numbers to acquire is overwhelming. In Section 3.4 we touch upon methods for learning structure from a data base and for adapting the conditional probabilities to incoming cases.

Finally we describe the system *Child*.

3.1 Catching the structure

3.1.1 Family out?

> When I go home at night, I want to know if my family is home before I try the doors. (Perhaps the most convenient door to enter is double locked when nobody is home.) Now, often when my wife leaves the house she turns on an outdoor light. However, she sometimes turns on this light if she is expecting a guest. Also, we have a dog. When

nobody is home, the dog is put in the back yard. The same is true if the dog has bowel trouble. Finally, if the dog is in the back yard, I will probably hear her barking, but sometimes I can be confused by other dogs barking.

The first thing to have in mind when organizing a Bayesian model for a decision support system is that its purpose is to give estimates of certainties for events which are *not observable* (or only observable at an unacceptable cost). So, the primary task in model building is to identify these events. We call them *hypothesis events*.

Here we have two hypothesis events, namely *family at home* and *family out*.

Now, the hypothesis events have to be organized into a set of variables. A variable incorporates an exhaustive set of mutually exclusive events. That is, for each variable precisely one of its events is true.

Here it is very easy to organize the events into one variable *F-out?* with states *y* and *n*.

The next thing to have in mind is that in order to come up with a certainty estimate, we should provide some *information channels*. So, the task is to identify the types of achievable information which may reveal something about the state of some hypothesis variable. This is also done by establishing certain variables, *information variables*, such that a piece of information corresponds to a statement about the state of an information variable. Typically, the information will be a statement that a particular information variable is in a particular state; but also more soft statements are allowed.

Here, the information variables are *L-on?* (light on) with states *y* and *n* and *H-bark?* (hear bark) also with states *y* and *n*.

Now it is time to consider the causal structure between the variables. At this stage we need not worry about how information is transmitted through the network. The only thing to worry about is which events have a direct causal impact on other events.

In this example it is clear that *F-out?* has an impact on *L-on?* as well as on *H-bark?*, and that there is no causal relation between *H-bark?* and *L-on?*.

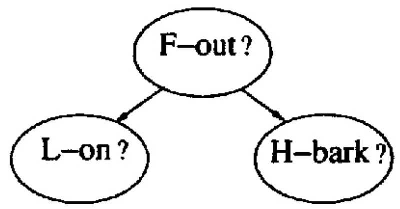

Figure 3.1 A causal structure for *family-out?*

We may stop with the model in Figure 3.1 and start specifying the probabilities *P(F-out)*, *P(H-bark? | F-out?)* and *P(L-on? | F-out?)*. We will defer the remaining treatment of this example to the section on specification of the probabilities (Section 3.2.4).

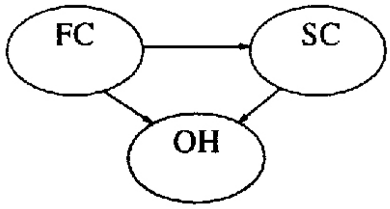

Figure 3.2 An oversimplified structure for the poker game. The variables are *FC* (first change), *SC* (second change), and *OH* (opponent's hand).

3.1.2 A simplified poker game

> In this poker game each player receives three cards and is allowed two rounds of changing cards. In the first round you may discard any number of cards from your hand and get replacements from the pack of cards. In the second round you may discard at most two cards. After the two rounds of card changing, I am interested in an estimate of my opponent's hand.

The hypothesis events are the various types of hands in the game. They may be classified in the following way (in increasing rank): nothing special, 1 ace, 2 of the same value, 2 aces, flush (3 of a suit), straight (3 of consecutive value), 3 of the same value, straight flush. Ambiguities are resolved according to rank. This is of course a simplification, but you often have to do so when modelling. The hypothesis events are collected into one hypothesis variable *OH* (opponent's hand) with the classes given above as states.

The only information to acquire is the number of cards the player discards in the two rounds. (By saying so, we again are making an approximation. The information on the cards you have seen is relevant for your opponent's hand. If, for example, you have seen three aces then he cannot have two aces.)

So, the information variables are *FC* (first change) with states *0, 1, 2, 3* and *SC* (second change) with states *0, 1, 2*.

A causal structure for the information variables and the hypothesis variable could be as in Figure 3.2.

However, this structure will leave us with no clue as to how to specify the probabilities.

What we need are variables describing the opponent's hands in the process: the initial hand *OH0* and the hand *OH1* after the first change of cards. The causal structure will then be as in Figure 3.3.

To determine the states of *OH0* and *OH1* we have to produce a classification which is relevant for the determination of the states of the children (*FC* and *OH1*, say). We may let *OH0* and *OH1* have the following states: *nothing special, 1 ace, 2 of consecutive value, 2 of a suit, 2 of the same value, 2 of a suit and 2 of consecutive value, 2 of a suit and 2 of the same value, 2 of consecutive value and 2 of the same value, flush, straight, 3 of the same value, straight flush.*

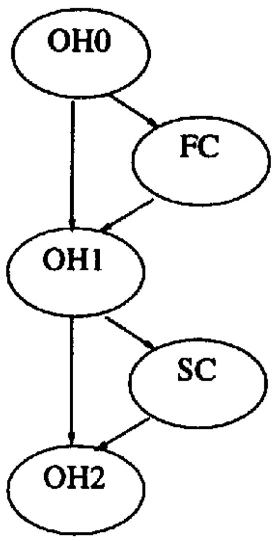

Figure 3.3 A structure for the poker game. The two mediating variables *OH0* and *OH1* are introduced. *OH2* is the variable for my opponent's final hand.

We defer further discussion of the classification to the section on specifying the probabilities (Section 3.2.2).

Variables in a model which are neither hypothesis variables nor information variables are called *mediating variables*. The decision on how to incorporate mediating variables is mainly a question of convenience. Usually mediating variables will ease the acquisition of conditional probabilities and thereby also increase the precision of the model. On the other hand there is a risk of increasing the complexity to a level which may jeopardize performance.

Another point is that it may happen that two variables – A and B – are dependent, but this dependence does not factor through any of the other variables. On the other hand, there is no obvious causal direction on the dependence. This should be taken as an indication that a mediating variable should be introduced as a parent of A and B. The next example illustrates this point.

3.1.3 Insemination

> Six weeks after insemination of a cow there are three tests for the result: blood test (BT), urine test (UT) and scanning (Sc). The results of the blood test and the urine test are mediated through the hormonal state (Ho) which is affected by a possible pregnancy (Pr). (This is a constructed example.)

A model will be like the one shown in Figure 3.4.

For both the blood test and the urine test there is a risk that a pregnancy does not show after six weeks. This is due to the fact that the change in the hormonal state may be too weak. Therefore, given pregnancy, the variables *BT* and *UT* are dependent.

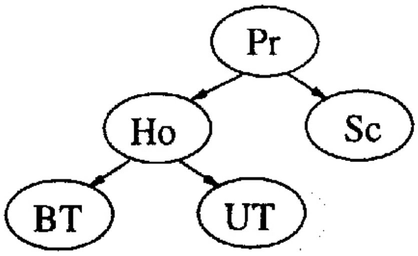

Figure 3.4 A model for test of pregnancy (Pr). Both the blood test (BT) and the urine test (UT) measure the hormonal state (Ho).

If we did not include the mediating variable, the model would be the one shown in Figure 3.5.

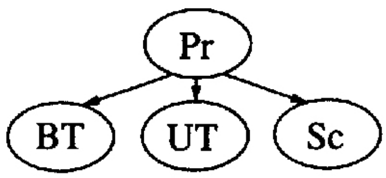

Figure 3.5 The pregnancy model without the *hormonal state* variable.

This model assumes the two tests to be independent given Pr.

If the model in Figure 3.5 is used for diagnosing a possible pregnancy, a negative outcome of both the blood test and the urine test will be counted as two independent pieces of evidence and therefore overestimate the probability for the insemination to have failed. (See Exercise 3.1.)

3.1.4 Simple Bayes models

The first Bayesian diagnostic systems were constructed through the following procedure.

- Let the possible diseases be collected into one hypothesis variable H with prior probability $P(H)$.
- For all information variables I, acquire the conditional probability $P(I \mid H)$ (the likelihood of H given I).
- For any set of findings $f_1, \ldots, f_n$ on the variables $I_1, \ldots, I_n$ calculate the product $L(H \mid f_1, \ldots, f_n) = P(f_1 \mid H)P(f_2 \mid H) \cdot \cdots \cdot P(f_n \mid H)$. This product is called the *likelihood* for H given $f_1, \ldots, f_n$. The posterior probability for H is calculated as $\mu P(H)L(H \mid f_1, \ldots, f_n)$, where μ is a normalization constant.

The calculations above reflect the simple model shown in Figure 3.6. (See Exercise 3.2.)

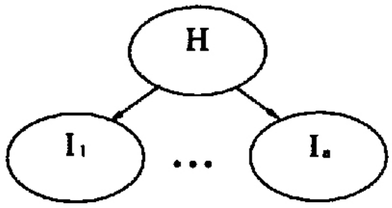

Figure 3.6 A simple Bayes model.

The model assumes that the information variables are independent given the hypothesis variable. As can be seen from the insemination example, the assumption need not hold, and if the model is used anyway, the conclusions may be misleading.

3.1.5 Causality

In the examples presented in the previous section there was no problem in establishing the links and their direction. However, you cannot expect this part of the modelling to always go smoothly.

First of all, causal relations are not always obvious – recall the debate on whether or not smoking causes lung cancer, or whether a person's sex has an impact on their abilities in the technical sciences. Furthermore, causality is not a well understood concept: is a causal relation a property of the real world, or, rather, is it a concept in our minds helping us to organize our perception of the world? We shall, however, not go into the scientific debate on causality and how to discover causal relations.

One point only. Causality has to do with actions where the state of the world is changed: you may, for example, find yourself confronted with two correlated variables A and B, but you cannot determine a direction. If you observe the state of A you will change your belief of B, and vice versa. A good test is then to imagine that some outside agent *fixes* the state of A. If this does not make you change the belief of B, then A is not a cause of B.

On the other hand, if this imagined test indicates a causal arrow in both directions, then you should look for an event which has a causal impact on both A and B. If C is such a candidate, then check whether A and B become independent given C.

3.2 Determining the conditional probabilities

The basis for the conditional probabilities in a Bayesian network can have different epistemological status ranging from well-founded theory over frequencies in a data base to subjective estimates. We shall give examples of each type.

3.2.1 Stud farm

> The stallion Brian has sired Dorothy with the mare Ann and sired Eric with the mare Cecily. Dorothy and Fred are the parents of Henry, and Eric has sired Irene with Gwenn. Ann is the mother of both Fred and Gwenn, but their fathers are in no way related. The colt John with the

Table 3.1 P(child | father, mother) for genetic inheritance. The numbers (α, β, γ) are the child's probabilities for *(aa, aA, AA)*.

	aa	aA	AA
aa	(1, 0, 0)	(0.5, 0.5, 0)	(0, 1, 0)
aA	(0.5, 0.5, 0)	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

parents Henry and Irene has been born recently; unfortunately, it turns out that John suffers from a life threatening hereditary disease carried by a recessive gene. The disease is so serious that John is displaced instantly, and as the stud farm wants the gene out of production, Henry and Irene are taken out of breeding. What are the probabilities for the remaining horses to be carriers of the unwanted gene?

The geneological structure for the horses is given in Figure 3.7.

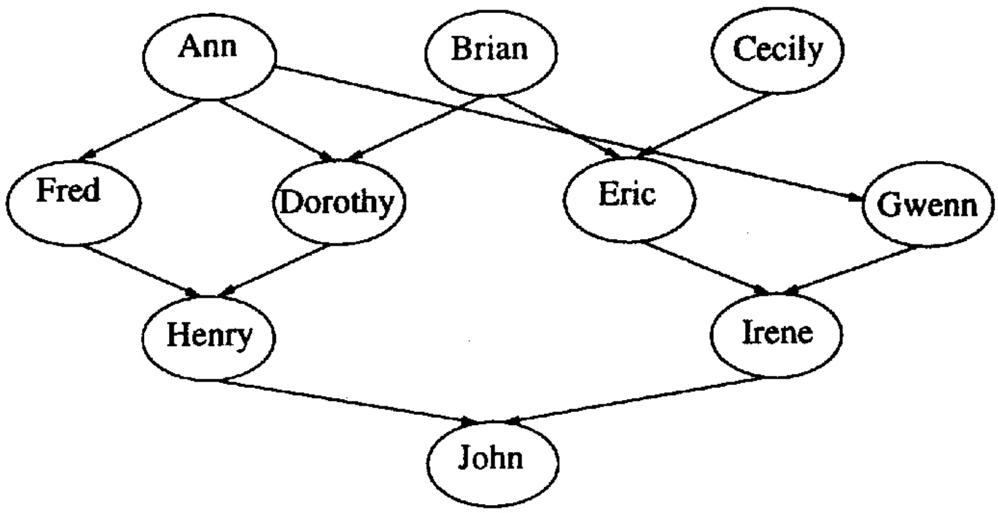

Figure 3.7 Geneological structure for the horses in the stud farm.

The only information variable is John. Before the information on John is acquired he may have three genotypes: he may be sick (aa), a carrier (aA), or he may be pure (AA). The hypothesis events are the genotypes of all other horses in the stud farm.

The conditional probabilities for inheritance are both empirically and theoretically well studied, and the probabilities are as shown in Table 3.1. The inheritance tables could be as Table 3.1. However, for all horses except John we have additional knowledge. Since they are in production they cannot be of type aa. A way to incorporate this would be to build a Bayesian network where all inheritance is modelled in the same way and afterwards enter the findings that all horses but John are not aa. It is also possible to calculate the conditional probabilities directly. If

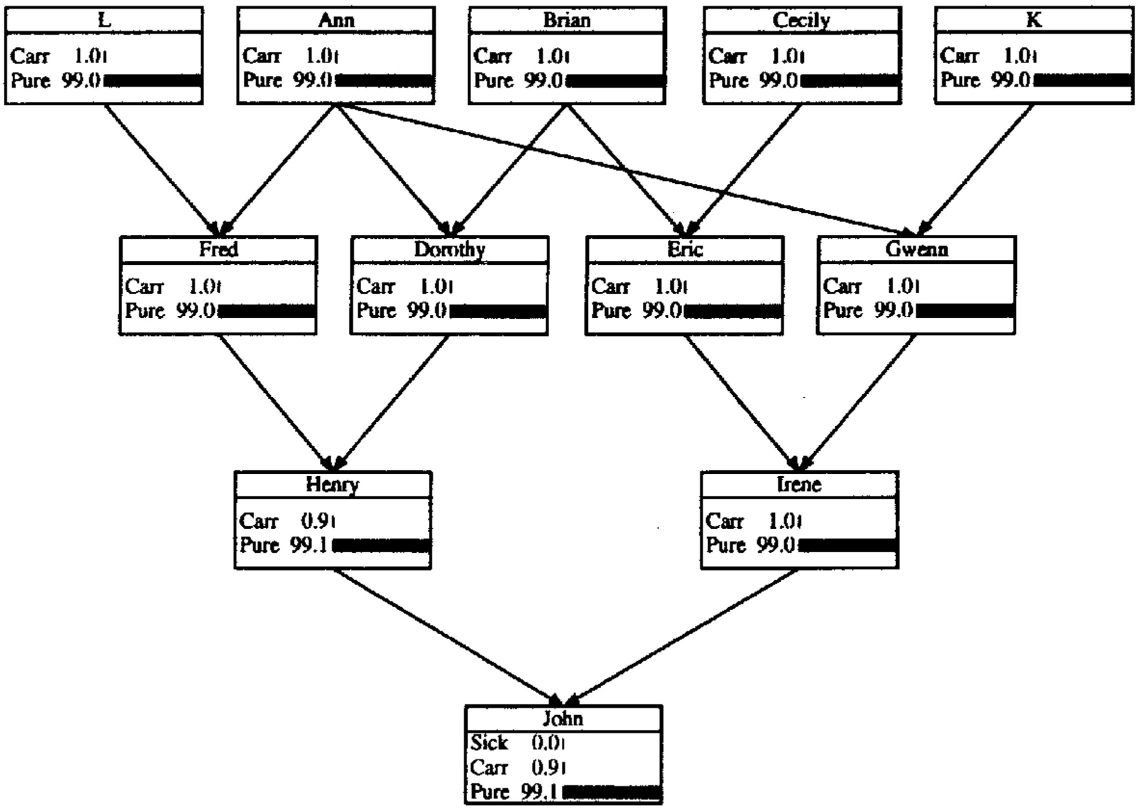

Figure 3.8 The *stud farm* model with initial probabilities. (HUGIN dump.)

we first consider inheritance from parents which may only be of genotype aA or AA, we get Table 3.2.

Table 3.2 $P(\text{child} \mid \text{father, mother})$ when the parents are not sick.

	aA	AA
aA	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 0.5, 0.5)	(0, 0, 1)

The table for John is the same as in Table 3.2. For the other horses we know that aa is impossible. This is taken care of by removing the state aa from the distribution and normalizing the remaining distribution. For example $P(\text{child} \mid aA, aA) = (0.25, 0.5, 0.25)$, but since aa is impossible we get the distribution $(0, 0.5, 0.25)$ which is normalized to $(0, 0.67, 0.33)$. The final result is shown in Table 3.3.

In order to deal with Fred and Gwenn we introduce the two unknown fathers, I and K, as mediating variables and assume that they are not sick. For the horses at the top of the network we shall specify prior probabilities. This will be an estimate of the frequency of the unwanted gene, and there is no theoretical way to come up with it. Let us assume that the frequency is so that the prior belief of a horse being a carrier is 0.01.

Table 3.3 P(child | father, mother) with aa removed.

	aA	AA
aA	(0.67, 0.33)	(0.5, 0.5)
AA	(0.5, 0.5)	(0, 1)

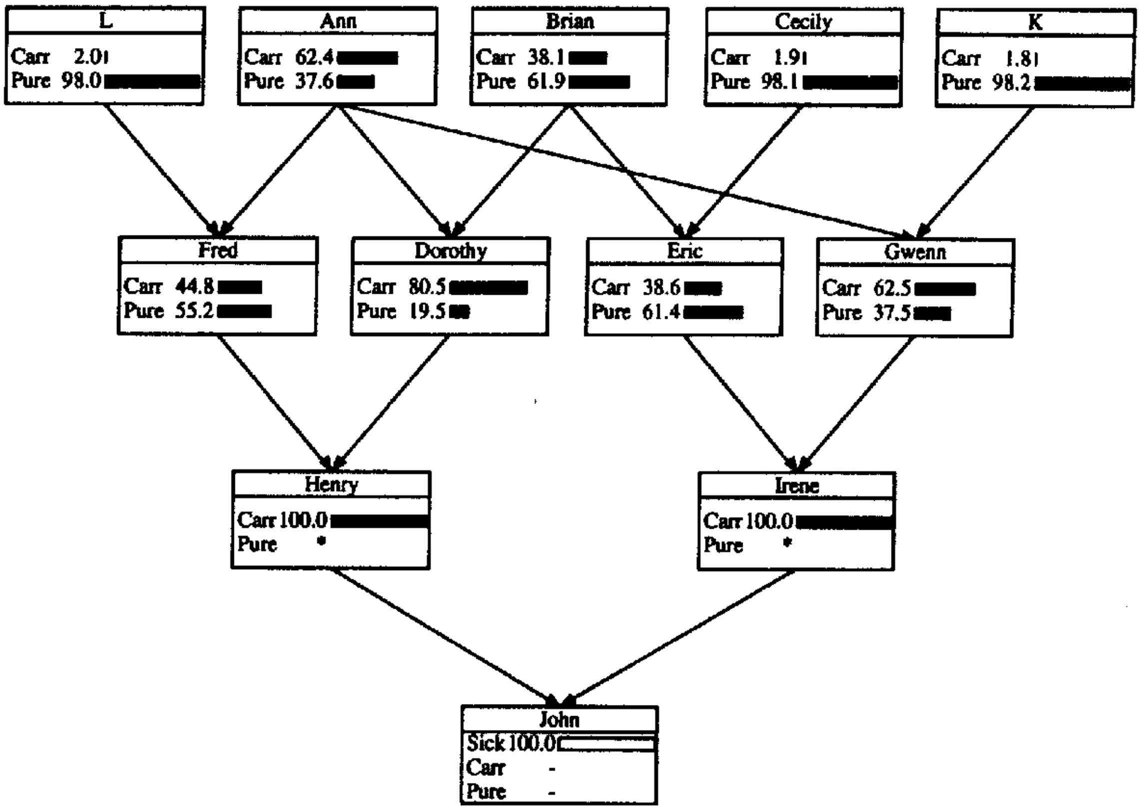

Figure 3.9 *Stud farm* probabilities given that *John* is sick. (HUGIN dump.)

In Figure 3.8 the final model with initial probabilities is shown, Figure 3.9 gives the posterior probabilities given John is aa, and in Figure 3.10 you can see the posterior probabilities with the prior beliefs at the top changed to 0.0001. Note that the sensitivity to the prior beliefs is very small for the horses where the posterior probability for *carrier* is well beyond zero, e.g. Ann and Brian.

3.2.2 Conditional probabilities for the poker game

In the *stud farm* example the conditional probabilities were mainly established through theoretical considerations. This should also be attempted for the model of the poker game developed in Section 3.1.2, but it cannot be carried through entirely.

Consider, for example, $P(FC \mid OH0)$. It is not possible to give probabilities which are valid for any opponent. It is heavily dependent on the opponent's insight, psychology and game strategies. We shall assume the following strategy.

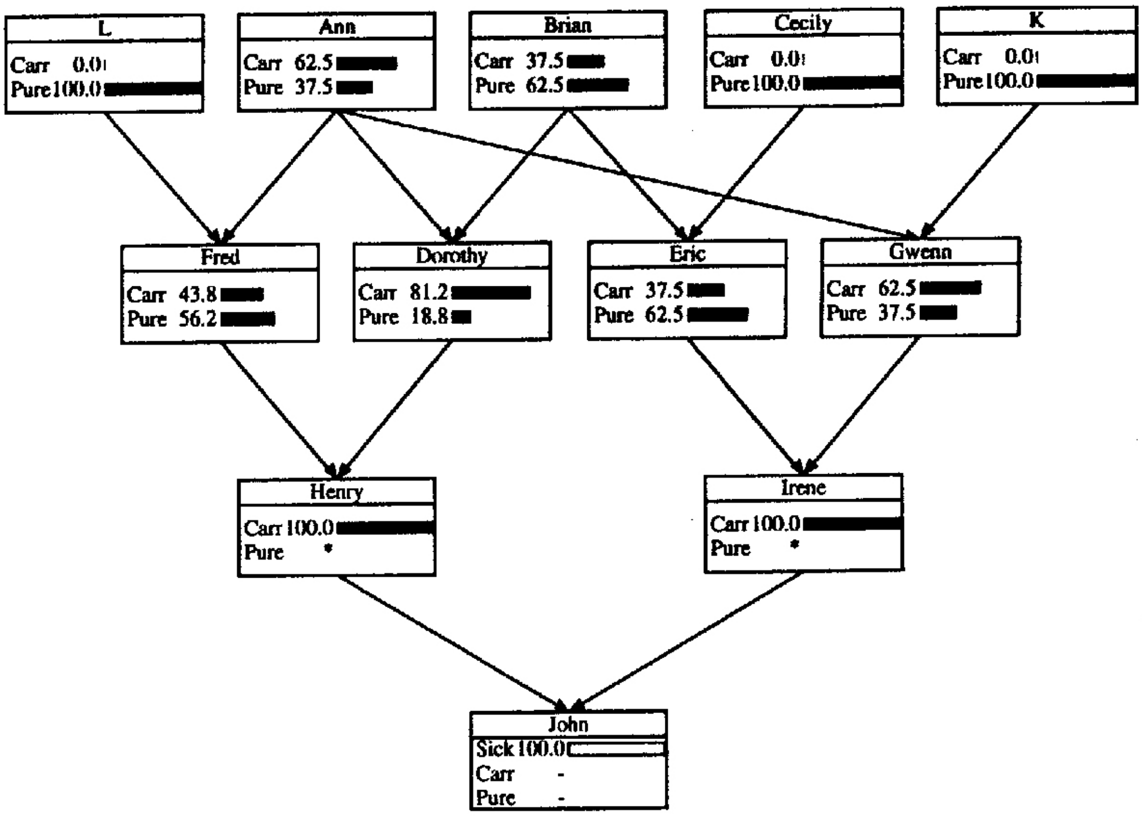

Figure 3.10 *Stud farm* probabilities with prior probabilities for top variables changed to (0.0001, 0.9999). (HUGIN dump.)

If nothing special (*no*), then change 3.

If 1 ace (*1 a*), then keep the ace.

If 2 of consecutive value (*2 cons*) or 2 of a suit (*2 s*) or 2 of the same value (*2 v*) then discard the third card.

If 2 of a suit and 2 of consecutive value, then keep 2 of a suit. (This strategy could be substituted by a random strategy for either keeping 2 of a suit or 2 of consecutive value.)

If 2 of a suit and 2 of the same value or 2 of consecutive value and 2 of the same value, then keep the 2 of the same value,.

If flush (*fl*), straight (*st*), 3 of the same value (*3 v*) or straight flush (*sfl*), then keep it.

Based on the strategy above, a logical link between FC and $OH0$ is established. Note that the strategy makes the states for combined hands redundant. They play no role, and therefore we remove them.

The strategy for $P(SC \mid OH1)$ is the same except that in the case of *no*, only 2 cards are discarded.

The remaining probabilities to specify are $P(OH0), P(OH1 \mid OH0, FC)$ and $P(OH2 \mid OH1, SC)$.

Table 3.4 $P(OH1 \mid OH0, FC)$ for the non-obvious parent configurations.

		(OH0, FC)				
		(no, 3)	(1 a, 2)	(2 cons, 1)	(2 s, 1)	(2 v, 1)
	no	0.1583	0	0	0	0
	1 a	0.0534	0.1814	0	0	0
	2 cons	0.0635	0.0681	0.3470	0	0
	2 s	0.4659	0.4796	0.3674	0.6224	0
OH1	*2 v*	0.1694	0.1738	0.1224	0.1224	0.9592
	fl	0.0494	0.0536	0	0.2143	0
	st	0.0353	0.0383	0.1632	0.0307	0
	3 v	0.0024	0.0026	0	0	0.0408
	sfl	0.0024	0.0026	0	0.0102	0

P(*OH0*). The states are (*no, 1 a, 2 cons, 2 s, 2 v, fl, st, 3 v, sfl*).

Through various (approximated) combinatorial calculations the prior probability distribution is found to be

$P(OH0) = (0.1672, 0.0445, 0.0635, 0.4659, 0.1694, 0.0494, 0.0353, 0.0024, 0.0024)$

P(*OH1* | *OH0*, *FC*). Due to the logical links between *OH0* and *FC* it is sufficient to consider only nine out of the possible 36 parent configurations, namely (*no, 3*), (*1 a, 2*), (*2 cons, 1*), (*2 s, 1*), (*2 v, 1*), (*fl, 0*), (*st, 0*), (*3 v, 0*), (*sfl, 0*). The last four are obvious. In Table 3.4 the results of approximate combinatorial calculations are given.

The probabilities for the remaining parent configurations may be whatever convenient. So, put, for example, $P(OH1 \mid 3\ v,\ 1) = (1, 0, 0, 0, 0, 0, 0, 0, 0)$.

P(*OH2* | *OH1*, *SC*). First a table $P(OH2' \mid OH1, SC)$ similar (but not identical in the numbers) to Table 3.4 can be calculated. However, the states of OH2′ are not the ones we are interested in. We are interested in the *value* of the hand and a state like *2 cons* is of no value unless one of them is an ace. Therefore, the probabilities for the states of *OH2′* are transformed to probabilities for *OH2*. For the transformation, the following rules are used:

$$1\ a = 1\ a + \frac{1}{6}(2\ cons + 2\ s)$$

$$no = no + \frac{5}{6}(2\ cons + 2\ s).$$

The probabilities of *2 a* are calculated specifically. The resulting probabilities are given in Table 3.5.

Using a model like the one in Figure 3.3 and with the conditional probability tables specified in this section, we have established a model for assisting a (novice) poker player. However, if my opponent knows that I use the system he may choose

Table 3.5 $P(OH2 \mid OH1, SC)$ for the non-obvious configurations.

		(OH1, Sc)				
		(no, 2)	(1 a, 2)	(2 cons, 1)	(2 s, 1)	(2 v, 1)
	no	0.5613	0	0.5903	0.5121	0
	1 a	0.1570	0.7183	0.1181	0.1024	0
	2 v	0.1757	0.0667	0.1154	0.1154	0.8838
OH2	2 a	0.0055	0.1145	0.0096	0.0096	0.0736
	fl	0.0559	0.0559	0	0.2188	0
	st	0.0392	0.0392	0.1666	0.0313	0
	3 v	0.0027	0.0027	0	0	0.0426
	sfl	0.0027	0.0027	0	0.0104	0

to change his strategies. His goal is to win rather than to obtain good hands, and he therefore may choose a strategy that makes me overestimate his hand. For instance, it seems a good strategy to discard two cards instead of three in the case of *no*. I will be convinced that he has an ace, and his chances for a good hand are not substantially reduced. We will return to this point in Chapter 6 on decision making.

3.2.3 Transmission of symbol strings

A language L over 2 symbols (**a**, **b**) is transmitted through a channel. Each word is surrounded by the delimiter symbol c. In the transmission some characters may be corrupted by noise and be confused with others.

A five-letter word is transmitted. Give a model which can determine the probabilities for the transmitted symbols given the received symbols.

There are five hypothesis variables $T_1, \ldots, T_5$ with states a and b and five information variables $R_1, \ldots, R_5$ with states a, b, c. Besides, mediating variables for the delimiters before and after the word may be considered. There is a causal relation from T_i to R_i. Furthermore, there may also be a relation from T_i to $T_{i+1} (i = 1, \ldots, 4)$. You could also consider more involved relations from pairs of symbols to symbols, but for now we refrain from that. The structure is given in Figure 3.11.

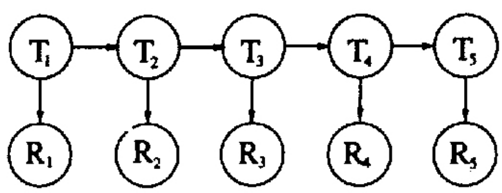

Figure 3.11 A model for symbol transmission. T_i are the symbols transmitted, R_i are the symbols received.

The conditional probabilities can be established through experience. The proba-