15 BAYESIAN METHODS APPLIED TO THE DATABASE INFERENCE PROBLEM

LiWu Chang and Ira S. Moskowitz

Abstract:

We apply Bayesian estimation and network techniques to the database inference problem. Bayesian analysis permits the realistic estimation of probabilities of missing data as well as insight into how prior knowledge and observed data interact. We urge our community to exploit this powerful tool.

15.1 INTRODUCTION

We take a Bayesian approach to the database inference problem (inference problem for short). Although the database community has analyzed the inference problem in many different ways (e.g., [5], [10], [9], [14], [18], [20], [22], [28]), other researchers have never formally used Bayesian techniques to study the problem (although several papers e.g., [8], [20], [28] have alluded to this method). Our main contribution is to apply standard Bayesian estimation theory to the inference problem. In particular, we use the method of inductive learning with a Bayesian network e.g., [7]. We analyze the type of inference problems to which our method is applicable.

The ability of a low-level user (Low) to infer higher-level information is the MLS database *inference problem.* We assume that the low user has the entire low-view of the database at its disposal. If the low database is, in fact, a high database with certain entries blocked out we have the *missing data* [16] approach to the inference problem (this term is also called *incomplete data* [27]). The data that is missing is the hidden high data. If the low database

S. Jajodia (ed.), *Database Security XII*

Table 15.1 Simple Relational Database with an Unknown Value

Name	*Location*	*Job*
Bond	Russia	Spy
Smart	France	Cook
Poirot	France	Cook
Sandiego	Russia	Spy
Christie	England	Clerk
Goldfinger	Russia	Spy
Holmes	England	Cook
Ames	Russia	?

is not part of a high database but by Low following implications between the database attributes, it is possible for Low to come up with a new relationship between the data that is, in fact, high information then we are dealing a *logical inference* approach [23]. These are two approaches to the inference problem to which Bayesian techniques can be usefully applied. We only study missing data in this paper and will address the later in future work. We are not concerned with other approaches such as the well-studied statistical/query based approach e.g., [10], [22][1].

Let us consider the following missing data example put forth by Marks [20]. We assume that Table 15.1 is the low database. The columns are the attributes. The challenge is to determine which job Ames holds, which is the missing value. That is high information since we are assuming that Table 15.1 with the last entry filled in is the high database. The obvious, though simplistic, answer might be to say that Ames is definitely a spy. In fact, this fact seems to be tacitly assumed in some papers, e.g. [20]. Why should we accept this? Do we have enough data upon which to base our decision? If we flip a coin twice, it comes up heads once, and tails once, can we say that it is a fair coin? If the first time I buy a lottery ticket or play a slot machine I win would you give me your home to gamble with on my second attempt? The problem is that we are dealing with a very small sample that is not statistically significant. Certainly, not everyone in Russia is a spy. There must be cooks and there must be clerks. *A fortiori*, if we know that only a small percentage of the people presently in Russia are spies, then we should not assume that Ames is a spy.

[1] In the statistical/query based approach the low user is not allowed full knowledge of the database. The complete database is considered high information. The low user is allowed to ask certain questions of the database and/or to know certain statistical information (e.g. max, min, mean, variances, etc). From this partial information the low user then tries to glean high information. Of course, there might be some intersection between the different approaches.

Table 15.2 Contingency Table

	Russia	*France*	*England*
Cook	0	2	1
Spy	3	0	0
Clerk	0	0	1

Therefore, we see that deterministic approaches to the inference problem can give skewed results and may be too limited to meet our security concerns. Certainly, a probabilistic leak of high information can be a cause for concern.

We wish to develop a means for analyzing the inference problem when:
Condition 1-The low database may have missing values, and
Condition 2-The low database may have nondeterministic rules and probabilistic relationships between attributes.

The inference problem fits into the bigger scheme of data mining [18], [15], [12]. With this in mind, we use the data mining (learning) technique of Bayesian networks to analyze the inference problem, meeting conditions one and two as above. The example in Table 15.1 is a toy problem that can be expressed as a simple Bayesian network. Before giving formal definitions we will work through the toy problem.

The problem is to determine the job of Ames. We have three choices: Spy, Cook, or Clerk[2]. In Table 15.1 the first column (name) identifies the sample and designates the row (tuple). We take columns two and three and form a contingency table as shown in Table 15.2.

We take a *subjective* or "degree of *belief*" approach to probability. In a belief based approach, we use our prior knowledge of a situation, along with observed data, to arrive at a probability. The prior knowledge is referred to as the *prior* distribution (simply called the prior), the data on hand is referred to as the *observation.* The idea of combining the prior and the observation to give us the *posterior* distribution, by using Bayes' theorem, is called the Bayesian method.

Thm [Bayes] $P(E|F) = \frac{P(F|E)P(E)}{P(F)}$

When attempting to fit observed data to into a probabilistic model (such as what we are doing with the low view of the database), we must be careful not to over-fit the data. Certainly, given two points representing a function, we should not say, in general, that the function is a straight line. Similarly, we can not always assume a deterministic interpretation of the data that is given. The Bayesian approach is very good for this.

[2]Note we are assuming that there are only three choices of jobs. If there were more (but not represented in the observed data) our Bayesian approach would still work by increasing the number of parameters.

We need to determine $P(A = i)$, where A represents the (categorical) random variable "Job of Ames" which can take on the values i = cook, spy, clerk. It is meaningless that the name is Ames. What matters is that there is a Russian with an unknown occupation. The occupations of the French or English are not germane to our analysis. However, we must consider the Russian data. This is why we condition upon it. The event representing the data from the first column of Table 15.2 is represented by D_r. Our goal is to determine $P(A = spy|D_r)$. Our prior distribution is the distribution describing the occupation of a Russian. The prior distribution has three values. For a discrete random variable the range values are the parameters. We let $\theta_1 = P(cook)$, $\theta_2 = P(spy)$, and $\theta_3 = P(clerk)$. Since probabilities must sum to one, we really only have two parameters (θ_3 can be written as: $\theta_3 = 1-(\theta_1+\theta_2)$). This is, in effect, a second order probability analysis—we are assigning probabilities to probabilities. We decompose $P(A = spy)$ as follows (All integration is definite, for notational simplicity, we often do not write out the region of integration): we have

$$P(A = spy|D_r) = \int\int \frac{P(A = spy|D_r, \theta_1, \theta_2)P(D_r|\theta_1, \theta_2)f(\theta_1, \theta_2)}{P(D_r)} d\theta_2\, d\theta_1 \quad [1.1]$$

What is $P(D_r|\theta_1, \theta_2)$? We assume that the occurrences making up Table 15.2 are independent. Therefore, since there are three spies:

$$P(D_r|\theta_1, \theta_2) = {\theta_1}^0{\theta_2}^3(1 - \theta_1 - \theta_2)^0 = {\theta_2}^3 \quad [1.2]$$

Consider the term $P(A = spy|D_r, \theta_1, \theta_2)$. This is the probability of $A = spy$ conditioned on the data D_r and the the priors having the values θ_1 and θ_2. (Note we are abusing notation by sometimes not distinguishing between the random variable describing the prior θ_i and the values that the parameter may assume.) This tells us that we are to assume that the parameters are taking on those particular values. Therefore, since the second parameter is the prior for *spy*, we have no choice but to assign the conditional probability $P(A = spy|\theta_1, \theta_2)$ the value θ_2. Also note that the event D_r does not influence this conditional probability. Therefore, $P(A = spy|D_r, \theta_1, \theta_2) = \theta_2$ also.

15.2 TOY EXAMPLE: NON-INFORMATIVE PRIOR AND TWO DISCRETE PRIOR EXAMPLES

Next, we need to determine the density function $f(\theta_1, \theta_2)$. We have many choices for what the distribution of these parameters should be. For now we take a *non-informative* view of the prior and assume that the parameters are jointly uniformly distributed. Consider the parameter θ_1. It gives the possible values for the probability of $A = cook$. Therefore, θ_1 can take on any value between zero and one. Given a value of θ_1 the parameter θ_2 can be between 0 and $1 - \theta_1$ (of course the third parameter is $1 - \theta_1 - \theta_2$). Therefore, the integration is taken over the region given by the right triangle $0 \leq \theta_1 \leq 1$, $0 \leq \theta_2 \leq 1 - \theta_1$. Since the area of the triangle is $1/2$ we see that $f(\theta_1, \theta_2) = 2$. At this point, we do a standard Bayesian trick and treat $P(D_r)$ as a normalizing

constant k^{-1} and do not calculate it at this time. Therefore Eq. [1.1] simplifies to

$$P(A = spy|D_r) = 2k \int_0^1 \int_0^{1-\theta_1} \theta_2 {\theta_2}^3 d\theta_2 \, d\theta_1 = \frac{k}{15}$$

Now let us consider $P(A = cook|D_r)$. We see that $P(A = cook|D_r, \theta_1, \theta_2) = \theta_1$ and we have

$$P(A = cook|D_r) = \iint \frac{P(A = cook|D_r, \theta_1, \theta_2) P(D_r|\theta_1, \theta_2) f(\theta_1, \theta_2)}{P(D_r)} d\theta_2 \, d\theta_1 = \frac{k}{60}$$

Using $P(A = clerk|D_r, \theta_1, \theta_2) = 1 - \theta_1 - \theta_2$ we see that $P(A = clerk|D_r) = \frac{k}{60}$. Since $P(A = cook|D_r) + P(A = spy|D_r) + P(A = clerk|D_r) = 1$ we have that $P(D_r) = 1/10$ and therefore $k = 10$. Therefore, $P(A = spy|D_r) = 2/3$, $P(A = cook|D_r) = 1/6$, and $P(A = clerk|D_r) = 1/6$. We could have easily calculated $P(D_r)$ directly in this case since

$$P(D_r) = \int \int P(D_r|\theta_1, \theta_2) P(\theta_1, \theta_2) d\theta_2 d\theta_1 \qquad [2.1]$$

What would happen if our data were different? What if there were n spies and no cooks or clerks in first column of the database table? In that case Eq. [1.1] would become

$$P(A = spy|D_r) = 2k \int_0^1 \int_0^{1-\theta_1} {\theta_2}^{n+1} d\theta_2 \, d\theta_1 = \frac{2k}{(n+2)(n+3)} \qquad [2.2]$$

where, as before $k^{-1} = P(D_r)$. Since Eq. [2.1] gives us $P(D_r) = 2/[(n+1)(n+2)]$ we see that Eq. [2.2] reduces to $P(A = spy|D_r) = (n+1)/(n+3)$. This tells us that as $\lim_{n\to\infty} P(A = spy|D_r) = 1$. This agrees with our intuition—the number of spies is getting larger and larger and still no clerks or cooks appear in the data set. The data set lets us adjust our views on the prior and gives us what we hope is a better guess at the posterior distribution $P(A = spy|D_r)$.

What if we used a different prior? What would happen if our prior knowledge was in fact definite knowledge? By this we mean that the prior is given by one (non-trivial) value, $P(\theta_1 = a, \theta_2 = b) = 1$. These statements are so strong that they overrule any influence from D_r. Obviously with such priors, one would not need to calculate anything. We have no choice but to say that $P(A = spy|D_r) = b$, regardless of what D_r is. Let us see if our equations give us this result. Recall that the purpose of playing with this toy example is to give us insight into the Bayesian technique. Since we have probability mass functions $P(\theta_1 = i, \theta_2 = j)$ instead of, as before, probability density functions $P(\theta_1, \theta_2)$, the integration becomes summation. Therefore, we now have

$$P(A = spy|D_r) = \frac{P(A = spy|D_r, \theta_1 = a, \theta_2 = b) P(D_r|\theta_1 = a, \theta_2 = b) \cdot 1}{P(D_r)}$$

In our initial example with three spies, no cooks or clerks, then both $P(D_r) = P(D_r|\theta_1 = a, \theta_2 = b) = b^3$. Therefore, these two terms cancel out. This holds

true as long as $b \neq 0$. In that case, we would be dividing by zero. In this case, it would also be impossible to have a data set with spies in it! So our mathematics and intuition agree (luckily). So we are left with $P(spy|D_r, \theta_1 = a, \theta_2 = b)$ which is just b. So, when there is just one value for the prior, it is also the value for the posterior distribution.

Now let us consider the example where the prior θ_2 can take on two non-trivial values b and d, without loss of generality $b < d$. We take as a probability mass function: $P(\theta_1 = a, \theta_2 = b) = u$, $P(\theta_1 = c, \theta_2 = d) = 1 - u$ where $a \leq 1 - b$ and $c \leq 1 - d$. Therefore, we have that $P(A = spy|D_r) = \frac{1}{P(D_r)} \{P(A = spy|D_r, \theta_1 = a, \theta_2 = b)P(D_r|\theta_1 = a, \theta_2 = b)P(\theta_1 = a, \theta_2 = b) + P(A = spy|D_r, \theta_1 = c, \theta_2 = d)P(D_r|\theta_1 = c, \theta_2 = d)P(\theta_1 = c, \theta_2 = d)\}$. Which gives us $P(A = spy|D_r) = \frac{1}{P(D_r)} \{b \cdot b^3 u + d \cdot d^3(1 - u)\}$ and similarly $P(A = cook|D_r) = \frac{1}{P(D_r)} \{uab^3 + (1 - u)cd^3\}$ and $P(A = clerk|D_r) = \frac{1}{P(D_r)} \{u(1 - a - b)b^3 + (1 - u)(1 - c - d)d^3\}$. Since all three terms must add to one, we have that $P(D_r) = ub^3 + (1 - u)d^3$. Note that, due to the simplicity of the priors, $P(D_r)$ could have been directly calculated as $P(D_r|\theta_1 = a, \theta_2 = b)P(\theta_1 = a, \theta_2 = b) + P(D_r|\theta_1 = c, \theta_2 = d)P(\theta_1 = c, \theta_2 = d)$. In the next section we will calculate $P(D_r)$ in this latter method. We include the discussion on the normalizing constant because this is a very common way of dealing with Bayesian problems [26]. Therefore,

$$P(A = spy|D_r) = \frac{ub^4 + (1 - u)d^4}{ub^3 + (1 - u)d^3} \qquad [2.3]$$

We see by Eq. [2.3] that $P(A = spy|D_r)$ is a function of $u, u \in [0, 1]$ so we will write it as $f(u)$. Obviously $f(0) = d$, and $f(1) = b$. Since $b < d$, we see that the derivative of f w.r.t. u is negative. Therefore, $f(u)$ is a function with a maximum of d and a minimum of b. The value of u determining the priors and the data set D_r determine the probability of $P(A = spy|D_r)$ but we know it must be in the range $[b, d]$.

We have played with this toy problem enough for now. Our goal with this exercise was to give the intuition behind the Bayesian method before we presented the full theory. That presentation is the next section.

15.3 OUTLINE OF BAYESIAN ANALYSIS

This section is based on work of Anderson [1] and of Heckerman [13].

Bayesian estimation (or prediction) of the value of a random variable X deals with computing the posterior probability of X equaling a certain value, based (conditioned) on the observed data D. The general approach is to derive an estimated probability distribution for the random variable based on the available data, and then to obtain the information about a particular value of interest from this derived probability distribution. The probability distribution is, in general, described by a family of parameters Θ. We assume that X is discrete and denote the parameter set by $\theta_1, \cdots, \theta_{|\Theta|}$, where $|\Theta|$ is the number of non-trivial values X may obtain. In other words, the non-trivial values that

X takes are v_k and the parameter corresponding to that value is θ_k. Note that each θ_i is itself a (usually continuous) random variable. The θ_i are constrained by the equation $\sum_{i=1}^{|\Theta|} \theta_i = 1$, since the θ_i represent probability values. The posterior probability of the k^{th} value v_k of X is

$$P(X = v_k|D) = \int P(X = v_k|\Theta, D)P(\Theta|D)d\Theta \; .$$

The total number of independent parameters is $|\Theta| - 1$, because $\sum_{i=1}^{|\Theta|} \theta_i = 1$. Without loss of generality, we view the last parameter $\theta_{|\Theta|}$ as $1 - \sum_{i=1}^{|\Theta|-1} \theta_i$. Thus the integral is a $(|\Theta| - 1)$-fold integral, and the region of integration is $0 \leq \sum_{i=1}^{|\Theta|-1} \theta_i \leq 1$. D can be dropped out from the first term of the last integral because it no longer affects the probability of v_k once the parameters are known. Thus, we have

$$P(X = v_k|D) = \int P(X = v_k|\Theta)P(\Theta|D)d\Theta \qquad [3.1]$$

To compute the posterior probability of $P(\Theta|D)$, we need $P(D)$, $P(\Theta)$ and $P(D|\Theta)$. This allows us to rewrite Eq. [3.1] as

$$P(X = v_k|D) = \int \frac{P(X = v_k|\Theta)P(D|\Theta)P(\Theta)}{P(D)}d\Theta \qquad [3.2]$$

Under the Bayesian assumption, each datum is independently drawn and the conditional probability of the data, given parameters, obeys the multinomial distribution. Thus, the likelihood of data is given by $P(D|\Theta) = \prod_{k=1}^{|\Theta|} \theta_k^{n_k}$ where n_k is the number of samples in D matching the v_k value (compare to Eq. [1.2]).

We use the Dirichlet distribution for the prior probability $P(\Theta)$. $P(\Theta) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{|\Theta|} \Gamma(\alpha_k)} \prod_{k=1}^{|\Theta|} \theta_k^{\alpha_k - 1}$ where $\alpha_k > 0$, $\alpha = \sum_k^{\Theta} \alpha_k$, and $\Gamma(\cdot)$ is the Gamma function. Note that:

(1) When there are only two parameters this is also called the Beta distribution.

(2) When $\forall k$, $\alpha_k = 1$ this special Dirichlet distribution becomes a uniform distribution (*not* over the unit hypercube, since we must account for the Gamma functions). This is called the non-informative prior (as in the toy problem).

Keep in mind that in our Bayesian analysis the "last" parameter $\theta_{|\Theta|}$ is taken to be $1 - \sum_{i=1}^{|\Theta|-1} \theta_i$.

The use of the Dirichlet distribution (a form "conjugate" to the multinomial distribution) is a standard Bayesian technique for several important reasons [2], [11], [1]. The expected values of the Dirichlet distribution (w.r.t. each parameter) give us a frequentist interpretation of the various coefficients. Our posterior probabilities of X will be in a form that lets us see the influence of the weighting of each prior, via the coefficients θ_k, and the contribution from the observed data. Not all priors are given by the Dirichlet distribution but for this paper, they are.

P(D) is given by (it is basically integration by parts *ad nausea*)

$$P(D) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{|\Theta|} \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \quad [3.3]$$

where $N = \sum_k^{|\Theta|} n_k$.

Finally, using the fact that $P(X = v_k|\Theta) = \theta_k$, we arrive at:

$$P(X = v_k|D) = \frac{\alpha_k + n_k}{\alpha + N} \quad [3.4]$$

Eq. [3.4] is very interesting and one of the reasons that the Dirichlet distribution is used (of course we are not advocating using a certain method/technique simply because it results in the "correct" answer). Eq. [3.4] combines two ratios. One ratio is that of the weighting, via the coefficient α_k, of the given prior parameter against the sum of the coefficients α. The second ratio is that of the number of occurrences of the value in question in the observed data against the total number of observations. Applying Eq. [3.4] to our toy problem with the non-informative priors, we have that $P(A = spy|D_r) = \frac{1+3}{3+3}$ (which is also what we got before!). What if, instead of a uniform prior, we had a generalized Beta distribution? If, when assigning our priors, we have more confidence in "spy" we can show this in the priors by letting α_k be large. As α_k increases we see that $P(A = spy|D_r) \rightarrow 1$. Similarly, even with the non-informative (uniform) priors if we had 300 data elements and all of them were spies (in the Russian column) we would have that $P(A = spy|D_r) = \frac{1+300}{3+300} \approx 1$. In addition, if the observed data has a small amount of non-spy observations, but the number of spies is large, we see that the non-spy observations have a small influence on our posterior probability. Therefore, Eq. [3.4] provides a good intuitive interplay between our assumptions about the prior distribution of the value parameters and the observed data set.

15.4 NETWORK MODEL

We now examine scenarios that are more complicated. We start with a motivating example similar to our toy problem (see Table 15.3).

15.4.1 Example

Now we do not know that Ames is Russian, nor do we know what job he has. We use the random variable J which represents the jobs that a person from any country (Russian, England, or France) can have. Now we must use the entire contingency table (Table 15.2) since we do not know the country. We want to know the probability of Ames being a spy based on the available data and our estimation techniques. Unfortunately, we now have two missing attributes—Location and Job. Therefore, we must generalize our Bayesian technique from the previous section. This generalization takes us into the area of Bayesian networks [26]. We will work this example and then give the full theory. Our approach is based upon the work done by Cooper [7] in Bayesian learning.

Table 15.3 Simple Relational Database with Two Unknown Values

Name	*Location*	*Job*
Bond	Russia	Spy
Smart	France	Cook
Poirot	France	Cook
Sandiego	Russia	Spy
Christie	England	Clerk
Goldfinger	Russia	Spy
Holmes	England	Cook
Ames	?	?

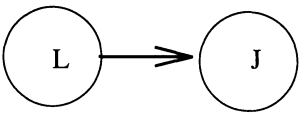

Figure 15.1 A Simple Bayesian Network

The term $P(J = spy|D)$ is what we want. Now D is the data: three Russian spies, two French cooks, one English clerk and one English cook. We decompose $P(J = spy|D)$ as $P(J = spy|D) =$
$P(J = spy, L = R|D) + P(J = spy, L = F|D) + P(J = spy, L = E|D)$ where L is the random variable representing country. Note we do not expect this answer to be the same as $P(A = spy|D_r)$ because we are not sure of the country and are basing our estimations upon the entire observed data.

We form a Bayesian net B_n as shown in Fig. 15.1. The arrow going from L to J tells us that if we change our belief in L, we must change our belief in J.

Let us start out by calculating $P(J = spy, L = R|D)$. According to our Bayesian technique, we write this as $P(J = spy, L = R|D, B_n)$. This emphasizes the net that we are using. The parameter set Θ now consists of nine parameters. Note that we will (subtly) choose our parameters to be consistent with our underlying net topology. Also, all integration is only over the relevant independent parameters. Unfortunately, no notation seems to capture this in general, so all parameters appear. We have the parameter set consisting of the parameters θ_R, θ_E, and θ_F for the three different locations. We only concern ourselves with the set Θ_L made up of θ_R and θ_E ($\theta_F = 1 - (\theta_R + \theta_E)$). Given a location only two parameters are needed to describe the priors for occupation we have the set $\Theta_{J|L}$ made up of the six parameters $\theta_{spy|R}$, $\theta_{cook|R}$, $\theta_{spy|E}$, $\theta_{cook|E}$, $\theta_{spy|F}$, and $\theta_{cook|F}$.

$$P(J = spy, L = R|D, B_n) = \int P(J = spy, L = R, |\Theta, D, B_n) P(\Theta|D, B_n) d\Theta$$

This follows by the rules of conditioning upon Θ. Recall that when we condition upon both D and Θ that Θ subsumes D. Thus, we see that the above $= \int P(J = spy, L = R, |\Theta, B_n)P(\Theta|D, B_n)d\Theta$. Write $P(J = spy, L = R|\Theta, B_n)$ as $P(J = spy|L = R, \Theta, B_n)P(L = R|\Theta, B_n)$. Now we make the local network structure assumption [27] which gives us
$P(J = spy|L = R, \Theta, B_n)P(L = R|\Theta, B_n)$
$= P(J = spy|L = R, \Theta_{J|L}, B_n)P(L = R|\Theta_L, B_n)$. Actually we only use the two parameters $\theta_{spy|R}$ and $\theta_{cook|R}$, we do not change our terms for notational simplicity. This assumption makes sense because in each node only the local parameters matter. We can now write the above as

$$\int P(L = R, |\Theta_L, B_n)P(J = spy|L = R, \Theta_{J|L}, B_n)P(\Theta_L, \Theta_{J|L}|D, B_n)d\Theta_L \, d\Theta_{J|L}$$

Now we make use of the parameter assumption [27]
$P(\Theta_L, \Theta_{J|L}|D, B_n) = P(\Theta_L|D, B_n)P(\Theta_{J|L}|D, B_n)$, this gives us that the above simplifies to

$$= \int P(L = R|\Theta_L, B_n)P(\Theta_L|D, B_n)d\Theta_L$$
$$\cdot \int P(J = spy|L = R, \Theta_{J|L}, B_n)P(\Theta_{J|L}|D, B_n)d\Theta_{J|L}$$

Thus, we see that the Bayesian network boils down to the non-network formulas that we had before. We make the same assumptions about multinomial sampling and the Dirichlet distribution. The B_n terms have just really come along for the ride. Let us consider the first term
$\int P(L = R|\Theta_L, B_n)P(\Theta_L|D, B_n)d\Theta_L = \frac{\alpha_R + n_R}{\alpha + N}$. The coefficient α_R corresponds to the prior for location Russia, which we assume to be one (non-informative prior). The sum of the location coefficients is α, since we are assuming they are all one, this sum is three. N is the total number of observations (7) and n_R is the number of Russians (3). Hence the first integral is 4/10. The second integral follows as before from Eq. [3.4] and is 4/6. So $P(J = spy, L = R|D) = 8/30$. Similarly we find that $P(J = spy, L = F|D) = 3/50$, and $P(J = spy, L = E|D) = 3/50$. Therefore, $P(J = spy|D) = 58/150 < 2/3$. It is not surprising that it is less than 2/3 because this was $P(A = spy|D_r)$ and the D data has more non-spies than the D_r data (which has none).

We now present the complete development of Bayesian networks, following [7], [13], [19].

15.5 BAYESIAN NETWORK THEORY

A Bayesian network model is used to deal with samples drawn from an m-dimension sample space, with the dimensions corresponding to the attributes. Let $\vec{X} \stackrel{\text{def}}{=} (X_1, \cdots, X_m)$. A sample from a database is an instantiation of the set of attributes and is denoted by the bold face vector $\vec{\mathbf{X}}$. This is done to simplify the notation; the value of the random vector is implicitly assumed. A

Bayesian network can be viewed as a collection of local networks. Each consists of a child node X_i and its (immediate) parent nodes, pa_i. It is an acyclic graph with each node corresponding to an attribute and a (directed) link indicating the conditioning probability, $P(X_i = x|pa_i^j)$. Note that an instantiation of the parent variables, pa_i^j, defines a (conditional) probability distribution for X_i. (In our previous example, X_i was the random variable J with values $x \in \{R, E, F\}$, and pa_i was the random variable L with the values spy, cook or clerk corresponding to the values of j in the term pa_i^j.) An n-node network has n local networks. The posterior probability of the entire network is simply the multiplication of posterior probabilities of those local networks.

The Bayesian network model requires two layers of evaluation. At one layer, we evaluate the parameters associated with each local network. The second layer selects the best topological structure for the network B_n. We compute the probability distribution for each local component. Θ is used to denote all parameters. θ_i is the collection of the parameters associated with the local network that has child node X_i. θ_{ij} is the set of parameters with parent variables of X_i taking the j^{th} instantiation (this is the set of all possible values of the parent nodes). Keep in mind that the set $\{j\}$ depends on which node X_i we are using. To keep the notation at a minimum, we do not write j as a function of i, but it should always be kept in mind. θ_{ijk} is the parameter which associates with the variable X_i that takes its k^{th} value, given that its parent variables are at the j^{th} instantiation. We have that θ_{ijk} is the distribution modeling $P(X_i^k|pa_i^j, \theta_{ij}, B_n)$. For a local network, we have $\sum_k^{K_i} \theta_{ijk}$=1. Notations J_i and K_i stand for the number of instantiations associated with parent variables and the number of different values of variable X_i, respectively. So we now examine $P(\vec{\mathbf{X}}|D, B_n)$ because we must also condition upon the underlying net structure,

$$P(\vec{\mathbf{X}}|D, B_n) = \int P(\vec{\mathbf{X}}|\Theta, D, B_n)P(\Theta|D, B_n)d\Theta \qquad [5.1]$$

We assume that at the local network of X_i, the probability $P(\theta_{ij}|D, B_n)$ has the Dirichlet distribution. By the parameter independence assumption $P(\Theta|D, B_n)$ is equal to $\prod_i \prod_j^{J_i} P(\theta_{ij}|D, B_n)$. The first term of the integral in Eq. [5.1] can be simplified, with the local network structure independence assumption so that Eq. [5.1] can be written as follows (As discussed in the example all integration is only over independent parameter sets. Also the parameters have been chosen to be consistent with B_n. If they were not, we would still get the same answer but our integration would be over a different region and the Jacobian from the change of variables would normalize things out.): $P(\vec{\mathbf{X}}|D, B_n) = \prod_{i=1} \int \cdots \int P(X_i^k|pa_i^j, \theta_{ij}, B_n)P(\theta_{ij}|D, B_n)d\theta_{ij}$, therefore

$$\begin{aligned} P(\vec{\mathbf{X}}|D, B_n) &= \prod_{i=1} \int \cdots \int \theta_{ijk} P(\theta_{ij1}, \cdots, \theta_{ij|K_i|}|D, B_n)d\theta_{ij1}, \cdots, d\theta_{ij|K_i|} \\ &= \prod_{i=1} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \end{aligned}$$

Table 15.4 Missing values over multiple tuples

Name	*Location*	*Job*
Bond	Russia	Spy
Smart	France	Cook
Poirot	France	Cook
Sandiego	Russia	Spy
Christie	England	Clerk
Goldfinger	Russia	Spy
Holmes	?	Cook
Ames	Russia	?

15.6 MISSING VALUES OVER MULTIPLE TUPLES

Consider the following from Table 15.4 which shows a more complicated situation than what we have previously discussed because we have data missing in more than one tuple (row).

Our approach is largely based on the Gibbs sampling Monte Carlo [25],[3] method where a missing value is repeatedly assigned with a new estimation conditioned on the current values of all other data. The estimation of missing values in our approach includes the following steps: 1-Initialize missing values and the network model. 2-For each attribute, estimate its value with a sample if its value is missing and replace the initial value with this new estimate.
3- Evaluate the network model. 4-Repeat the last two steps for all attributes. 5-Stop when reassignments are not required.

Let D^I and D^C denote the original incomplete database and the database with missing values assigned. The above approach is summarized by:

$$C_l^{u_i} = ?|D^I, B_n \quad \Leftarrow max_{x_i'} P(D^C_{x_i'}|B_n)$$

where $C_l^{u_i}$ means that the value of attribute X_i in sample C_l is missing, and $D^C_{x_i'}$ means that x_i' is assigned to X_i at the ith attribute for the sample C_l. Details can be found in the presentation version of this paper (available on the web).

15.7 CONCLUSIONS AND FUTURE WORK

Our Bayesian method is a double-edged sword (both sides useful). If Low is attempting to determine (probabilistically) high information, it can use Bayesian techniques to guesstimate the correct probability. On the other hand, with knowledge of Bayesian methods, High can introduce spurious data to confound Low's estimation techniques. The idea of padding data is not new. By using the Bayesian formulas, however, we can develop a framework on how to introduce this padded data judiciously, in order to conserve resources. In particular,

we see how the data can influence the final expressions (e.g., the n_k terms) of the posterior probabilities.

We wish to emphasize that our Bayesian techniques certainly call into question assumptions that the given data implies with *certainty* the occurrence of an event. Of course, the more data observed (low-view) and the more confidence we have in our prior probability distributions, the more we can *accurately* predict the missing high data. This is best summed up as *Small amounts of data imply questionable decisions, while large amounts of data imply better (but still not perfect) predictions.* This is certainly not a new thought but one that can be lost amidst elaborate new theories and notations. In this paper (as in [24]), we try to show the advantages of using powerful, well-analyzed methods that already exist in other fields. Bayesian techniques are well-studied and have been successfully used in software debugging and information retrieval [4], [21]. This is close in spirit to the inference problem in MLS database design. Our method is useful when dealing with multiple missing attribute values, and our future work will deal with more complicated databases.

We wish to continue our work by studying more complicated network topologies and determining under what conditions the Bayesian technique may fail. We also wish to explore the High-padding countermeasures discussed above.

We also believe that Bayesian analysis can complement the database search algorithm (e.g., [14]), where a path connecting one entity (e.g., the company table) to another one (e.g., the project table) can be constructed from multiple tables. Once a plausible path is found, Bayesian analysis can carry out inferencing by determining the causal dependency relationships among attributes. This is a logical approach which we feel can complement our *statistical approach.* We also feel that our approach can also complement the rough sets approach put forth by others [17] [28]. We also feel that a decision tree analysis can be useful in analyzing database inferences (see [6]).

Acknowledgments

We thank Ruth Heilizer along with Judy Froscher, Myong Kang, Carl Landwehr, and Cathy Meadows. We also thank the anonymous referees and workshop participants for their helpful comments. Research supported by the Office of Naval Research.

References

[1] Anderson, J. (1990) *The Adaptive Character of Thought.* NJ, Erlbaum.

[2] Berger, J. (1985) *Statistical Decision Theory and Bayesian Analyses.* NY: Springer-Verlag.

[3] Buntine, W. (1994) "Operations for Learning with Graphical Models," J. Artificial Intelligence Research, Vol.2, 159–225.

[4] Burnell, L. & Horvitz, E. (1995) "Melding Logic and Probability for Software Debugging," Comm. ACM Vol. 38, No. 3, 31–41.

[5] Buszkowski, W. & Orlowska, E. (1986) "On the Logic of Database Dependencies," Bull. Polish Academy of Science Mathematics, 34/5-6, 345–354.

[6] Chang, L. & Moskowitz, I. (1998) "Parsimonious Downgrading and Decision Trees Applied to the Inference Problem," Proc. New Security Paradigms Workshop 1998.

[7] Cooper, G. & Herskovits, E. (1992) "A Bayesian Method for the Induction of Probabilistic Networks from Data," J. Machine Learning, 9/4, 309–347.

[8] Delugach, H. & Hinke, T. (1994) "Using Conceptual Graphs to Represent Database Inference Security Analysis," J. Computing and Information Technology, V. 2, No. 4, 291–307

[9] Delugach, H. & Hinke, T. (1996) "Wizard: A Database Interface Analysis and Detection System," IEEE Trans. KDE, Vol 8, No 1, 56–66.

[10] Denning, D., Jones, A. & Lipton, R. (1979) "Secure databases: protection against user influence," ACM Trans. Database Syst. 4, 97–106.

[11] Geiger, D. & Heckerman, D. (1994) "A Characterization of The Dirichlet Distribution Applicable to Learning Bayesian Networks," MSR-TR-94-16.

[12] Hale, J. & Shenoi, S. (1997) "Catalytic Inference Analysis: Detecting Inference Threats due to Knowledge Discovery," Proc. IEEE Symp. Security & Privacy, 188–199.

[13] Heckerman D. (1996) "Bayesian Networks for Knowledge Discovery," Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 273–305.

[14] Hinke, T., Delugach, H. & Wolf, R. (1997) "Protecting Databases from Inference Attack," Computers & Security, Vol. 16, No. 8, 687–708.

[15] Hinke, T., Delugach, H. & Wolf, R. (1997) "A Framework for Inference-Directed Data Mining," *Database Security Vol. X,* IFIP, 229–239.

[16] Kong, A., Liu, J. & Wong, W. (1994) "Sequential Imputation and Bayesian Missing Data Problems," Journal of ASA, Vol. 89, No. 425, pp 278-288.

[17] Lin, T. Y. (1993) "Rough Patterns in Data-Rough Sets and Intrusion Detection Systems," J. of Foundation of Computer Science and Decision Support, Vol. 18, No. 3-4, 225–241.

[18] Lin, T.Y., Hinke, T.H., Marks, D.G., & Thuraisingham, B. (1996) "Security and Data Mining," *Database Security Vol. IX,* IFIP, 391–399.

[19] Madigan, D. & Raftery, A. (1994) "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window" J. of American Statistical Association, Vol. 89, No. 428, 1535–1546.

[20] Marks, D. (1996) "Inference in MLS Database Systems," IEEE Trans. Knowledge and Data Engineering, Vol 8, No. 1, 46–55.

[21] Maron, M. & Kuhns, J. (1960) "On Relevance, Probabilistic Indexing, and Information Retrieval," J. ACM 7, 216–244.

[22] Matloff, N. (1988) "Inference Control Via Query Restriction Vs. Data Modification: A Perspective," *Database Security: Status and Prospects I,* IFIP, 159–166.

[23] Morgenstern, M. (1988) "Controlling Logical Inference in Multilevel Database systems," Proc. IEEE Symp. on Security & Privacy, 245–255.

[24] Moskowitz, I. & Costich, O. (1992) "A Classical Automata Approach to Noninterference Type Problems," Proc. Computer Security Foundations Workshop 5, 2–8.

[25] Neal, R. (1993) "Probabilistic Inference Using Markov Chain Monte Carlo Methods" Technical Report, University of Toronto, TR CRG-TR-93-1.

[26] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.

[27] Spiegelhalter, D. & Lauritzen, S. (1990) *Sequential Updating of Conditional Probabilities on Directed Graphical Structures*, Networks, V. 20, 579–605.

[28] Zhang, K. (1997) "IRI: A Quantitative Approach to Inference Analysis in Relational Database," *Database Security Vol. X,* IFIP, 214–221.