International Statistical Review (1995), **63**, 2, 215–232, Printed in Mexico

Bayesian Graphical Models for Discrete Data

David Madigan *[1] **and Jeremy York**** [2]

[1]*University of Washington, Seattle, USA.* [2]*Battelle Pacific Northwest Laboratories.*

Summary

For more than half a century, data analysts have used graphs to represent statistical models. In particular, graphical "conditional independence" models have emerged as a useful class of models. Applications of such models to probabilistic expert systems, image analysis, and pedigree analysis have motivated much of this work, and several expository texts are now available.

Rather less well known is the development of a Bayesian framework for such models. Expert system applications have motivated this work, where the promise of a model that can update itself as data become available, has generated intense interest from the artificial intelligence community. However, the application to a broader range of data problems has been largely overlooked.

The purpose of this article is to show how Bayesian graphical models unify and simplify many standard discrete data problems such as Bayesian log linear modeling with either complete or incomplete data, closed population estimation, and double sampling. Since conventional model selection fails in these applications, we construct posterior distributions for quantities of interest by averaging across models. Specifically we introduce Markov chain Monte Carlo model composition, a Monte Carlo method for Bayesian model averaging.

Key words: Bayesian graphical models; Decomposable models; Directed acyclic graphs; Markov chain Monte Carlo; Model averaging; Log linear models; Closed population estimation; Double sampling

1 Introduction

The use of graphs to represent statistical models dates back to Wright (1921) and has been the focus of considerable activity in recent years. In particular, researchers have directed attention at graphical "conditional independence" models for discrete data and at the application of such models to probabilistic expert systems. The books by Whittaker (1990), Pearl (1988), and Neapolitan (1990), as well as Spiegelhalter *et al.* (1993) summarize these developments. Similar models have found application in image analysis (e.g., Geman & Geman (1984)) and pedigree analysis (e.g., Thompson (1986)).

Rather less well known is the development of a Bayesian framework for such models (Dawid & Lauritzen (1993), Spiegelhalter & Lauritzen (1990)). Expert system applications have motivated this work, where the promise of a model that can update itself as data become available, has generated intense interest from the artificial intelligence community (Charniak (1991), Kornfeld (1991), Heckerman *et al.* (1994)). However, the application to a broader range of discrete data problems has been largely overlooked.

*Department of Statistics, GN-22, University of Washington, Seattle, WA 98195 (madigan@stat.washington.edu). Madigan's research was supported by a grant from the NSF.

**York's research was supported by a NSF graduate fellowship.

The authors are grateful to Julian Besag, David Bradshaw, Jeff Bradshaw, James Carlsen, David Draper, Ivar Heuch, Robert Kass, Augustine Kong, Steffen Lauritzen, Adrian Raftery, and James Zidek for helpful comments and discussions. The authors are also grateful to Denis Allard who translated the Summary into French.

The purpose of this article is to show how Bayesian graphical models unify and simplify many standard discrete data problems such as Bayesian log linear modeling with either complete or incomplete data, model selection and accounting for model uncertainty, closed population estimation, and double sampling. This list by no means exhausts the possible applications. Our objective is to demonstrate the diverse range of potential applications, alert the reader to a new methodology and hopefully stimulate further development.

At the risk of over-simplification we sketch the basic framework for the Bayesian analysis of graphical models with a simple epidemiological example. In Norway, the Medical Birth Registry (MBR) gathers data nationwide on congenital malformations such as Down's syndrome. The primary purpose of the MBR is track prevalences over time and identify abnormal trends. The data, however, are subject to a variety of errors, and epidemiologists have built statistical models to make inference about true prevalences. For Down's syndrome, such a model includes three dichotomous random variables: the reported Down's syndrome status, R, the true Down's syndrome status, S, and the maternal age, A, where age is dichotomized at 40.

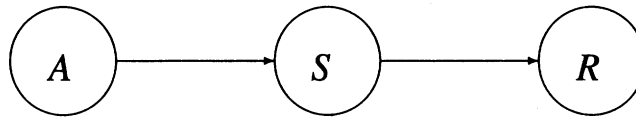


Figure 1. *Down's Syndrome: An Acyclic Directed Graphical Model*

Figure 1 displays a reasonable model for these variables. This directed graph represents the assumption that the reported status and the maternal age are conditionally independent given the true status. The joint distribution of the three variables factors accordingly:

$$\text{pr}(A, S, R) = \text{pr}(A)\text{pr}(S \mid A)\text{pr}(R \mid S). \tag{1}$$

The specification of the joint distribution of A, S, and R, in (1), requires five parameters:

$$\text{pr}(R \mid S), \text{pr}(R \mid \overline{S}), \text{pr}(S \mid A), \text{pr}(S \mid \overline{A}) \quad \text{and} \quad \text{pr}(A) \tag{2}$$

where $\overline{S}$, for example, denotes the absence of Down's syndrome. Once these probabilities are specified, the calculation of specific conditional probabilities such as $\text{pr}(R \mid A)$ can now proceed via a series of local calculations without storing the full joint distribution (Dawid (1992)).

In most of the expert systems applications of these models, it is assumed that an "expert" provides the probabilities. An obvious development of the above framework is to update expert knowledge about the model parameters as data accumulate thereby providing an extension from probabilistic reasoning to statistical analysis. The use of point estimates for the probabilities in (2) precludes the possibility of such updating so instead we elicit prior *distributions* for these quantities. In effect, the probabilities become random variables and can be added to the graph as in Figure 2. Within this framework Spiegelhalter & Lauritzen (1990) show how independent beta distributions placed on these probabilities can be updated locally to form the posterior as data become available. This provides an attractive strategy for Bayesian analysis of discrete data. The graph concisely communicates model assumptions and facilitates the derivation of various model properties. Informative prior distributions can realistically be elicited in terms of probabilities rather than the abstract parameters. Furthermore, the required computations are straightforward.

In later sections we extend and apply this framework. Some common themes will be apparent across these applications:

First, Bayesian graphical models facilitate the implementation of the complete Bayesian paradigm.

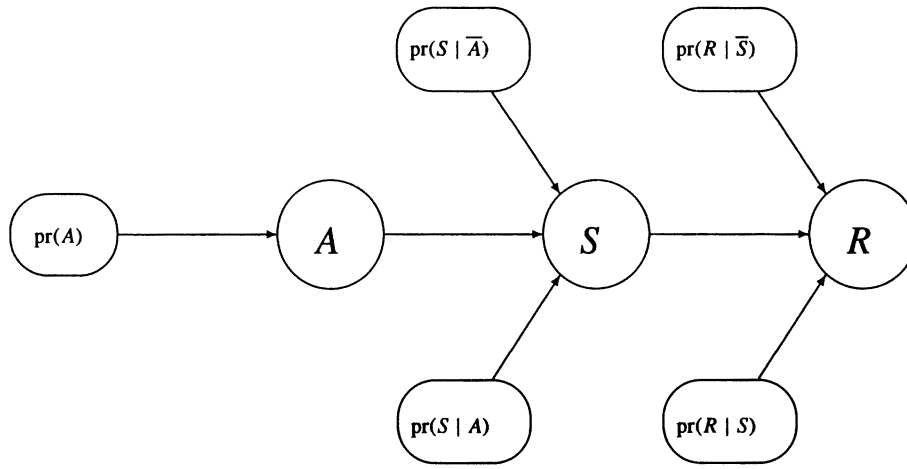


Figure 2. *Down's Syndrome: An Acyclic Directed Bayesian Graphical Model*

The elicitability of *informative* prior distributions motivates many of the constructions we present in later sections.

Second, in several of the applications we consider, missing data and/or latent variables produce ostensibly insurmountable analytic obstacles. Such complexity frequently rules out the consideration of larger models involving many covariates and other generalizations. We show how Bayesian graphical models coupled with Markov chain Monte Carlo techniques provide a conceptually simple approach to such problems.

Finally, the Bayesian framework facilitates accounting for model uncertainty with model averaging.

In summary, there are many advantages to analyzing discrete data with Bayesian graphical models. This paper aims to illustrate these advantages, and point out some potential drawbacks.

In the next section, we define graphical models and describe more fully the Bayesian framework sketched above. Section 3 addresses double sampling problems using directed graphical models. Section 4 addresses a closed population estimation problem using undirected graphical models. Finally we discuss possible extensions of this work and other potential applications.

2 Bayesian Analysis of Discrete Graphical Models

A *graphical model* is a statistical model embodying a set of conditional independence relationships which may be summarized by means of a graph (see Appendix I for a summary of the graph terminology we use). We will only consider graphs that are either *acyclic* and *directed*, or *chordal* and *undirected* in what follows. In a directed graph, *all* the edges are directed and appear as arrows in the figures. A directed graph is acyclic if it contains no directed cycles. In an undirected graph, *all* the edges are undirected and appear as plain lines in the figures. An undirected graph is chordal (or "triangulated") if it contains no cycles of four or more vertices without a chord. In either case, each vertex in the graph will correspond to a random variable X_v, $v \in V$ taking values in a sample space $\mathcal{X}_v$. To simplify notation, we use v in place of X_v in what follows.

In the *directed* case (see for example Figure 1), the parents pa(v) of a vertex v are those vertices from which edges point into v. The *descendants* of a vertex v are the vertices which are reachable

from v along a directed path. The parents are taken to be the only direct influences on v, and thus, v is independent of its non-descendants given its parents. This property implies a factorization of the joint distribution of X_v, $v \in V$, which we denote by $\text{pr}(V)$, given by:

$$\text{pr}(V) = \prod_{v \in V} \text{pr}(v \mid \text{pa}(v)). \tag{3}$$

In the *undirected* case (see for example Figure 3), for sets A, B and $S \subset V$, A and B are conditionally independent given S, whenever A and B are separated by S (that is, all paths in the graph connecting A and B pass through S). For a more detailed exposition of Markov properties for directed and undirected graphs, we refer the reader to Lauritzen *et al.* (1990).

In the case where the random variables (X_v), $v \in V$ are all discrete, the class of models defined by the undirected graphs are a subclass of the hierarchical log linear models where the cliques of the graph correspond to the maximal terms in the log linear model. In what follows we restrict our attention to the "decomposable models" for which the graph is chordal. These are the "closed-form" log linear models for which parameters can be estimated without recourse to iterative methods. Note that this provides a simple method to check if a given log linear model is a closed-form model, in contrast to the intricate rules in, for example, Bishop *et al.* (1975, Section 3.4.3). The simplicity of such decomposable models has been exploited in a number of contexts—see for example Lauritzen & Spiegelhalter (1988), Dawid & Lauritzen (1993), Madigan & Mosurski (1990) and Madigan & Raftery (1994).

2.1 *Bayesian Framework for Directed Graphical Models*

Here we describe the Bayesian framework for directed graphical models. Consider the five parameters, $\text{pr}(R \mid S)$, $\text{pr}(R \mid \overline{S})$, $\text{pr}(S \mid A)$, $\text{pr}(S \mid \overline{A})$, and $\text{pr}(A)$, of the Bayesian graphical model of Figure 2. Spiegelhalter & Lauritzen (1990) and Cooper & Herkovits (1992) make two key assumptions which greatly simplify subsequent analysis.

First, they assume that the parameters are independent *a priori*. Figure 2 embodies this assumption; recall, that vertices are conditionally independent of their non-descendants given their parents. For instance, $\text{pr}(S \mid A)$ in Figure 2 has no parents. Therefore, it is conditionally independent of, for instance, $\text{pr}(A)$, since this is not a descendant of $\text{pr}(S \mid A)$.

Second, they assume that each of the probabilities has a Beta distribution (or Dirichlet distribution for non-dichotomous variables). This prior is conjugate with multinomial sampling and many techniques exist for eliciting beta distributions from experts (see, for example, Morgan & Henrion, 1990).

Calculation of likelihoods and posterior distributions is straightforward. Suppose we have the following prior distributions for three of the parameters: $\text{pr}(A) \sim \text{beta}(1,1)$, $\text{pr}(S \mid A) \sim \text{beta}(5,1)$, and $\text{pr}(R \mid \overline{S}) \sim \text{beta}(1,9)$, and we observe a single case, $d = (A, \overline{S}, R)$, that is, maternal age over 40 and incorrectly recorded by the MBR as Down's syndrome. Then, it is easy to show that $\text{pr}(d) = \frac{1}{1+1} \times \frac{1}{5+1} \times \frac{1}{1+9}$. Conditional on d, the posterior distributions of these parameters become: $\text{pr}(A) \sim \text{beta}(2,1)$, $\text{pr}(S \mid A) \sim \text{beta}(5,2)$, and $\text{pr}(R \mid \overline{S}) \sim \text{beta}(2,9)$. The posterior distributions of the remaining two parameters, $\text{pr}(S \mid \overline{A})$ and $\text{pr}(R \mid S)$, are unchanged. In this manner, we can sequentially calculate the likelihood for a collection of cases, D, conditional on the model of Figure 2. Heckerman *et al.* (1994) provide a closed-form expression for this likelihood.

If D is complete, updating each probability individually in this fashion preserves parameter independence. However, if the data have cases with missing values, this is no longer the case. For example, if we observe a case in which the maternal age is missing, then $\text{pr}(D \mid A)$ and $\text{pr}(D \mid \overline{A})$ become dependent *a posteriori*. Spiegelhalter & Lauritzen (1990) suggest ignoring this dependence in subsequent analysis, but this can lead to severely biased estimates (see, for example, Spiegelhalter & Cowell (1991)).

2.2 *Bayesian Framework for Undirected Decomposable Graphical Models*

Dawid & Lauritzen (1993) described the Bayesian framework for decomposable undirected models. Figure 3 shows such a model which also embodies the assumption that A and R are independent given S.

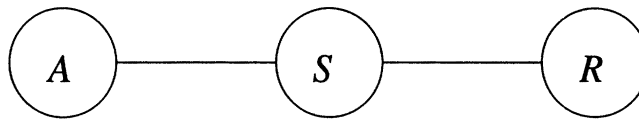


Figure 3. *A Decomposable Undirected Graphical Model*

The cliques of this graph are $\{A, S\}$ and $\{S, R\}$, and there is a single "clique intersection", $\{S\}$. The corresponding factorization of the joint distribution is:

$$\text{pr}(V) = \frac{\text{pr}(A, S)\text{pr}(S, R)}{\text{pr}(S)}.$$

In general, the factorization has a term for each clique in the numerator and a term for each clique intersection in the denominator. The clique intersections are defined by the ordering of the cliques, and the factorization requires that this be a "perfect ordering". Only decomposable models admit perfect clique orderings.

The parameters for the Bayesian graphical model of Figure 3 are the probabilities for each of the cliques:

$\theta_{A,S} = (\text{pr}(A, S), \text{pr}(A, \overline{S}), \text{pr}(\overline{A}, S), \text{pr}(\overline{A}, \overline{S}))$ and

$\theta_{S,R} = (\text{pr}(S, R), \text{pr}(S, \overline{R}), \text{pr}(\overline{S}, R), \text{pr}(\overline{S}, \overline{R}))$.

Following Dawid & Lauritzen (1993), we assume that $\theta_{A,S}$ and $\theta_{S,R}$ have Dirichlet distributions, constrained so that each implies the same marginal distribution for $\text{pr}(S)$. Specifically, if $\theta_{A,S} \sim \text{Dirichlet}(\lambda_{A,S}, \lambda_{A,\overline{S}}, \lambda_{\overline{A},S}, \lambda_{\overline{A},\overline{S}})$ and $\theta_{S,R} \sim \text{Dirichlet}(\lambda_{S,R}, \lambda_{S,\overline{R}}, \lambda_{\overline{S},R}, \lambda_{\overline{S},\overline{R}})$, then:

$$\lambda_{A,S} + \lambda_{\overline{A},S} = \lambda_{S,R} + \lambda_{S,\overline{R}} \quad \text{and} \quad \lambda_{A,\overline{S}} + \lambda_{\overline{A},\overline{S}} = \lambda_{\overline{S},R} + \lambda_{\overline{S},\overline{R}}.$$

Dawid & Lauritzen (1993) showed that there exists a unique "hyper-Dirichlet" distribution with specified marginal Dirichlet distributions on the cliques of a decomposable graphical model and discuss the conditional independence relationships amongst the parameters implied by this prior distribution.

Again, this prior distribution is conjugate with multinomial sampling. If $\theta_{A,S} \sim \mathcal{D}(1, 2, 3, 4)$ and $\theta_{S,R} \sim \mathcal{D}(2, 2, 3, 3)$, the probability of $d = (A, \overline{S}, R)$ is $\frac{2/10 \times 3/10}{4/10}$ and the distributions of the parameters posterior to d are $\theta_{A,S} \sim \mathcal{D}(1, 3, 3, 4)$ and $\theta_{S,R} \sim \mathcal{D}(2, 2, 4, 3)$. In this manner we can sequentially calculate the likelihood for a collection of cases.

2.3 *Which Graphical Model?*

We have described two classes of graphical models: undirected decomposable and acyclic directed. This raises the question of which class is most appropriate for any given problem.

In problems where some variables are obviously determined before others, or cause others, the directed graphs allow a natural representation of these assumptions. For example, if there is a relationship between the maternal age (A) and Down's syndrome (D), it is certainly A which influences or causes or precedes D, and not the other way around; thus, an edge between them should point from A to D.

Undirected models, in contrast, are best suited to problems where the variables are determined simultaneously, or perhaps are both influenced by some variable which is not explicitly modeled. For example, it does not make sense to say that an individual's eye colour influences or causes his or her hair colour, or vice versa, and so a relationship between these variables is better represented as an undirected edge.

Many problems will include both kinds of relationships, motivating the use of "chain graphs" with both directed and undirected edges (Frydenberg (1990)). Currently we are extending the class of Bayesian graphical models to include chain graphs.

In the applications which follow, rather than carry out a model selection exercise leading to a *single* graphical model, we account for model uncertainty by averaging over all the models in the class. Thus, the posterior distribution of a quantity of interest, Δ, conditional on the data, D, is given by:

$$\text{pr}(\Delta \mid D) = \sum_{k=1}^{K} \text{pr}(\Delta \mid M_k, D)\text{pr}(M_k \mid D). \tag{4}$$

This is an average of the posterior distributions under each of the models, weighted by their posterior model probabilities. Madigan & Raftery (1994) and Draper (1995) discuss the implementation of (4) in detail.

In general the number of terms in the summation in (4) can be impractically large and we approximate (4) using "Markov chain Monte Carlo model composition"(MC^3). MC^3 generates a stochastic process which moves through the class of models. Specifically, let $\mathcal{M}$ denote the class of models under consideration. We can construct a Markov chain $\{M(t), t = 1, 2, \ldots\}$ with state space $\mathcal{M}$ and equilibrium distribution $\text{pr}(M_i \mid D)$. If we simulate this Markov chain for $t = 1, \ldots, N$, then under mild regularity conditions, for any function $g(M_i)$ defined on $\mathcal{M}$, the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \tag{5}$$

is a simulation–consistent estimate of $E(g(M))$ (Smith & Roberts, 1993). To compute (4) in this fashion set $g(M) = \text{pr}(\Delta \mid M, D)$.

To construct the Markov chain we define a neighbourhood $\text{nbd}(M)$ for each $M \in \mathcal{M}$ which consists of the model M itself and the set of models with either one edge more or one edge fewer than M. Define a transition matrix q by setting $q(M \to M') = 0$ for all $M' \notin \text{nbd}(M)$ and $q(M \to M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \to M')$. It is then accepted with probability:

$$\min\left\{1, \frac{\#(\text{nbd(M)})\text{pr(M}' \mid \text{D)}}{\#(\text{nbd(M}')\text{)pr(M} \mid \text{D)}}\right\}$$

Otherwise the state stays in state M. The irreducibility of the transition matrix q is obvious in the directed case. For the decomposable case it follows immediately from Lemma 5 of Frydenberg & Lauritzen (1989). Madigan & Raftery (1994) describe how to efficiently compute $\frac{\text{pr}(M' \mid D)}{\text{pr}(M \mid D)}$ through local calculations. York *et al.* (1995) extend MC^3 to also integrate over missing data.

Madigan *et al.* (1994a) demonstrate that MC^3 provides improved predictive performance over inference based on any single model that might reasonably have been selected. We note that similar approaches are suggested by Cooper & Herskovits (1992).

3 Bayesian Graphical Models for Double Sampling

We now introduce the first of two applications demonstrating the utility of Bayesian graphical models. This application uses acyclic directed graphical models.

3.1 Example : Down's Syndrome in Norway

Consider again the Norwegian Down's syndrome example. Because of growing concerns about incomplete ascertainment, an additional notification system entitled "Melding om Fosterindiserte Aborter og Medfødte Misdannelser" (MIA) was introduced in 1985 in the county of Hordaland covering about 15% of all births in Norway. The MIA registration is based on prenatal diagnostics and pediatric follow-up including results from cytogenetic tests. While it was expected that the MIA registration would be more accurate than the MBR, the MIA registration *is* subject to error. Table 1 presents data concerning Down's syndrome collected between 1985 and 1988 (Lie *et al.*, 1991,1994).

Table 1

Down's syndrome data for 1985–1988 : R_1 represents case ascertainment through the national MBR registry and R_2 through the regional MIA registry. A represents the maternal age in six categories: $\leq$ 24, 25—29, 30—34, 35—39, 40—44, and $\geq$ 45.

Doubly Sampled Data

	A_1	A_2	A_3	A_4	A_5	A_6	Total
R_1, R_2	1	2	0	3	2	0	8
$R_1, \overline{R}_2$	5	2	3	0	2	1	13
$\overline{R}_1, R_2$	1	4	2	1	1	0	9
$\overline{R}_1, \overline{R}_2$	7478	10247	7058	2532	504	28	27847
Total	7485	10255	7063	2536	509	29	27877

Singly Sampled Data

	A_1	A_2	A_3	A_4	A_5	A_6	Total
R_1	32	55	58	62	23	3	233
$\overline{R}_1$	48957	70371	49115	16834	3348	165	188790
Total	48989	70426	49173	16896	3371	168	189023

York *et al.* (1995) presented a Bayesian graphical model analysis of these data which overcomes three substantive difficulties with the analysis of Lie *et al.* (1994). First, Lie *et al.* (1994) did not consider any covariates such as maternal age in their analysis. Because of the strong association with maternal age, a complete study of the prevalence of Down's syndrome should include this covariate (Lie *et al.*, 1991). However, the complexity of the existing analysis, in particular the calculation of asymptotic variances, suggests that such expansions would be difficult. Second, Lie *et al.* (1994) failed to utilize well-established knowledge about Down's syndrome. Third, Lie *et al.* (1994) identified two models, both of which provide a reasonable fit to the data, but lead to Down's syndrome prevalence estimates and corresponding asymptotic standard errors that are quite different. The Bayesian graphical model framework addresses these problems.

Denoting by Δ, the prevalence of Down's syndrome, and by Y, the observed data, we seek to compute $\text{pr}(\Delta \mid Y)$. To account for model uncertainty and integrate over Z, the missing data on the singly sampled cases, we re-express this as:

$$\text{pr}(\Delta \mid Y) = \sum \text{pr}(\Delta \mid M, Y, Z)\text{pr}(M, Z \mid Y)$$

where the summation is over all models, M, and all possible states of the missing data, Z. We compute this using MC^3, where the Markov chain is augmented to include Z. Note that $\text{pr}(\Delta \mid M, Y, Z)$ is just the posterior distribution of Δ conditional on complete data and is calculated using the methods of Section 2.1.

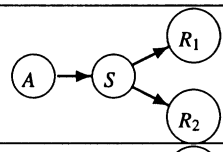
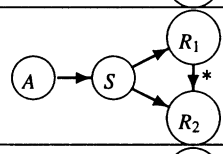
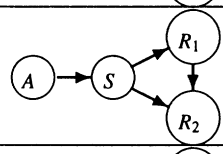
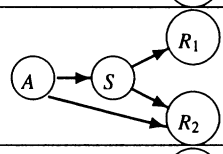
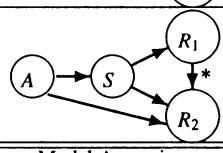
The results of a Bayesian graphical model analysis of the Down's syndrome data are given in Table

2 and Figure 4. In this analysis, all models were assumed equally likely *a priori* and informative prior distributions, based on historical data and expert knowledge were placed on the various probabilities. Specifically, the parameters for the prior on pr(S), the prevalence of Down's syndrome, were based upon raw data from the MBR registry for 1979–1984 without any adjustment for misclassification. During that period, there were 0.97 observed cases per 1,000 births; the beta(0.0097, 9.9903) prior was chosen to have that rate as its expected value and so that our prior knowledge is equivalent to having observed 10 hypothetical births. Choosing the prior to represent 100 hypothetical cases does not have a substantial impact upon the results. The priors for the probability of a case being identified by the registries reflect our belief that the registries are more likely to find cases of the syndrome than not. For the national MBR registry, we placed beta(4,2) priors on pr($R_1 \mid S$) and pr($R_2 \mid S$). For the 6 parameters, pr($S \mid A_i$), we chose prior variances so that Var(pr(S)) is the same for all models, and prior expectations given by historical data for 1979–1984, which are 0.59, 0.59, 0.97, 3.04, 6.88, and 18.50 cases per 1000 for age groups $A_1, \ldots, A_6$, as presented in Lie *et al.* (1991). York *et al.* (1995) describe the remaining prior distributions.

The analysis assumes that there are no false positives, which is reasonable in this context. Models with a '*' on the R_1, R_2 link impose a special kind of dependence where it is assumed that the MIA registry, R_2, will find all cases missed by the national registry, R_1.

Table 2

Features of the posterior for Down's syndrome prevalence and the error probabilities of the two registries. Prevalence is given as the rate per thousand. Only models with posterior probability larger than 0.01 *are listed; all models are included in the model averaging results.*

Model	Post. Prob.	$10^3 \times$ pr(S)			pr($\overline{R}_1 \mid S$)		pr($\overline{R}_2 \mid S$)	
		Mode	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
	0.282	1.81	1.92	0.292	0.376	0.085	0.555	0.092
	0.385	1.49	1.51	0.129	0.223	0.053	0.470	0.083
	0.269	1.60	1.70	0.252	0.312	0.088	0.513	0.089
	0.030	1.71	1.78	0.226	0.333	0.076	0.518	0.090
	0.016	1.50	1.52	0.129	0.226	0.054	0.517	0.080
Model Averaging	—	1.54	1.69	0.289	0.292	0.099	0.508	0.095

Except for the inclusion of the age covariate, the first two models in Table 2 correspond respectively

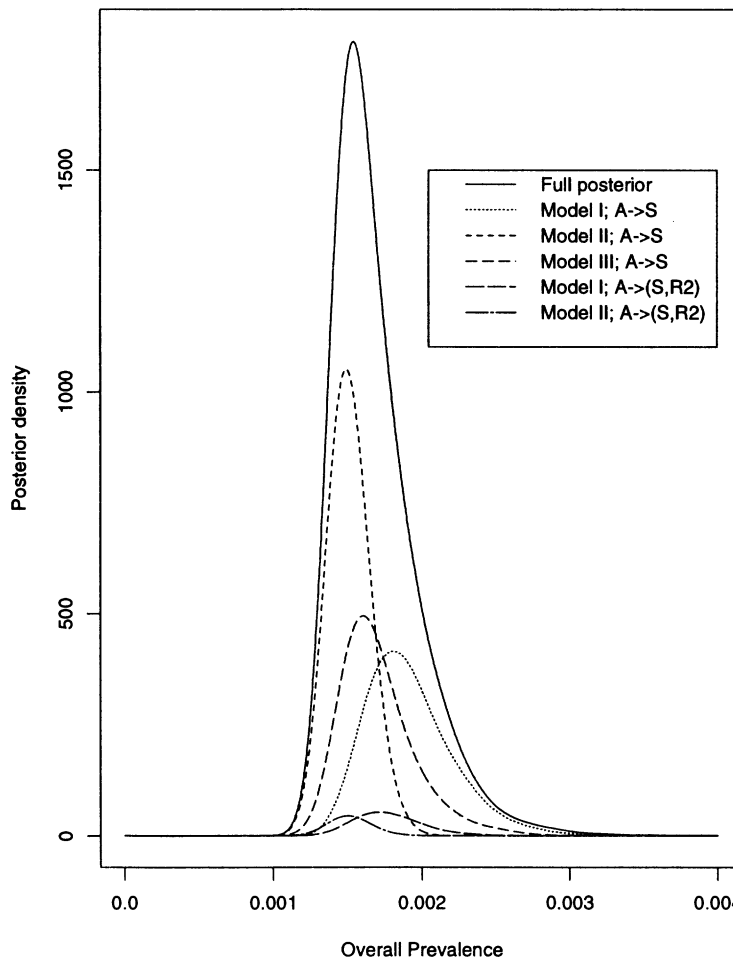


Figure 4. *Overall posterior for Down's syndrome rate per 1000 when the mother's age is included as a covariate, along with the posterior for each individual model scaled according to its posterior probability.*

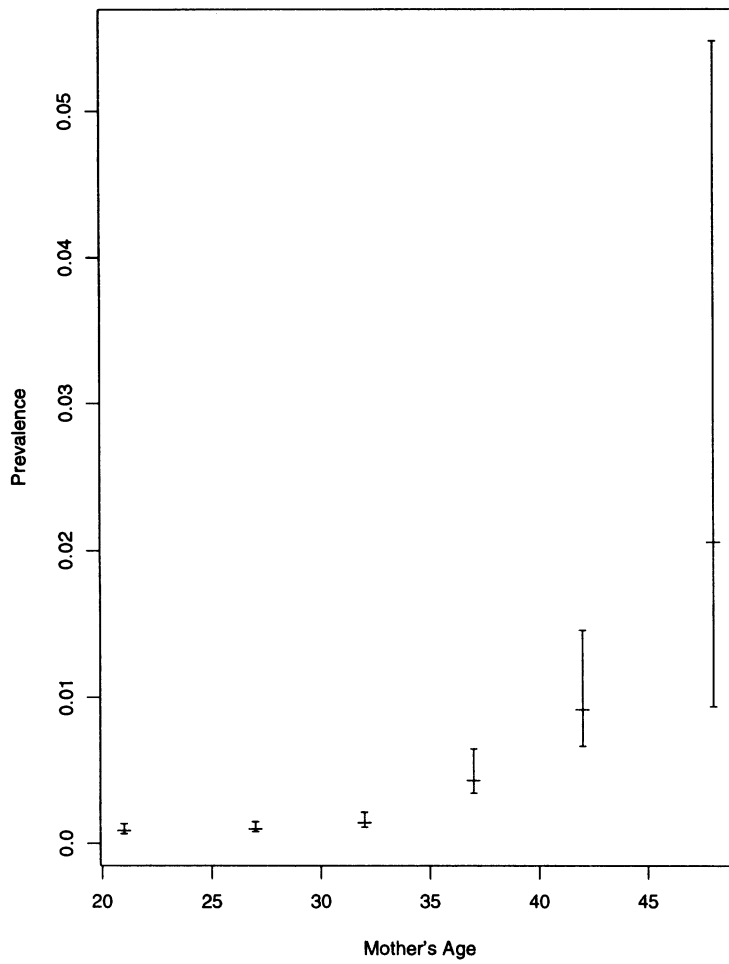


Figure 5. *Mode and 5th and 95th percentiles of the posterior for Down's syndrome prevalence by age of mother, averaged across all models.*

to the two models examined by Lie *et al.* (1994). Their first model produced a maximum likelihood estimate for $10^3 \times \text{pr}(S)$ of 2.02 with a standard deviation of 0.35, while their second model gives 1.49 and 0.13. Our analysis accounts for this model uncertainty, averaging over all the models. Furthermore, incorporation of the maternal age substantially improves model fit and allows for age-specific reporting, such as in Figure 5.

3.2 Why Bayesian Graphical Models?

Bayesian graphical models extend the reach of multiply sampled data analysis into heretofore intractable areas. Expert knowledge can realistically be incorporated and model uncertainty can be accounted for. Models of considerable complexity can be considered and posterior distributions for a variety of quantities of interest derived. In contrast, the analyses of Ekholm & Palmgren (1987) and Lie *et al.* (1994) are complex and difficult to extend.

4 Bayesian Graphical Models for Closed Population Estimation

Here we introduce the second of two applications demonstrating the utility of Bayesian graphical models. This application uses decomposable undirected graphical models.

4.1 Introduction

One approach to estimating the size of a closed population is to use several methods to "capture" individuals in the population; we will consider a capture to be the occurrence of an individual's name in an administrative list. If it is possible to uniquely identify the individuals or their capture histories, then the data can be represented as a contingency table with one dimension for each list. The count for the cell in which individuals were not found by any of the lists is unknown, and must be estimated. It is assumed that the population is closed—there is no migration.

For example, Hook *et al.* (1980) gathered records on persons born in upstate New York between 1969 and 1974 with the defect *spina bifida* from birth certificates (B), death certificates (D), and medical rehabilitation records (R). The records were compared, providing the data in Table 3 (a 0 indicates that an individual was not found in a list, and a 1 indicates that he or she was found). 626 individuals were found in the three lists. We seek to estimate the number of individuals missed by all three lists and thence the population size.

Table 3

Spina Bifida data

	$R=0$	$R=1$
$B=0, D=0$	?	60
$B=0, D=1$	49	4
$B=1, D=0$	247	112
$B=1, D=1$	142	12

Fienberg (1972) and Bishop *et al.* (1975) present a methodology for these problems using log linear models. Unfortunately, Regal & Hook (1991) demonstrated that two different log linear models fit the data of Table 3 very well, but give very different confidence intervals for the population size. Other shortcomings of the technique are that it is difficult to take advantage of information on the size of the population which may be available *a priori*, the incorporation of covariate information is cumbersome, especially with missing values, and simultaneous modeling of multiple diseases is impractically complex.

Decomposable undirected Bayesian graphical models, as described in Section 2.2 provide a flexible

model class for this problem, facilitating the incorporation of prior expert knowledge and accounting for model uncertainty.

Denoting the total population size by N, and following the notation of Section 2, our objective is to evaluate the posterior distribution of N, given the observed data, D:

$$\text{pr}(N \mid D) = \sum_{k=1}^{K} \text{pr}(N \mid M_k, D)\text{pr}(M_k \mid D). \tag{6}$$

We assume *a priori* that N is independent of the model M_k, and thus (6) can be written as:

$$\text{pr}(N \mid D) = \sum_{k=1}^{K} \text{pr}(D \mid M_k, N)\text{pr}(M_k)\text{pr}(N)/\text{pr}(D), \tag{7}$$

where:

$$\text{pr}(D \mid M_k, N) = \int \text{pr}(D \mid \theta, M_k, N)\text{pr}(\theta \mid M_k)d\theta. \tag{8}$$

Here, θ is the vector parameter of probabilities which define M_k and is assumed to be independent of N. York & Madigan (1992) provide formulae for $\text{pr}(D \mid \theta, M_k, N)$ and $\text{pr}(D)$, and explore the consequences of various prior assumptions for $\text{pr}(M_k)$ and $\text{pr}(N)$ and for $\text{pr}(\theta \mid M_k)$, in the context of several examples.

The component distributions of $\text{pr}(\theta \mid M_k)$ all involve the probability of capture on one or more lists. One practical difficulty that arises is that the structure of this prior distribution depends on M_k; recall that a Dirichlet distribution is required for each clique in the graph of M_k. This necessitates the elicitation of a different prior distribution for every model. York & Madigan (1992) describe a pragmatic solution to this problem. Essentially, our approach is to elicit a prior distribution for $\text{pr}(\theta \mid M_{k'})$, where $M_{k'}$ is a model with high prior probability, chosen for convenient elicitation. Prior distributions for θ under all the other models are then derived from $\text{pr}(\theta \mid M_{k'})$, via a simple information theoretic argument. Heckerman *et al.* (1994) provide a detailed examination of this problem.

Elicitation for undirected graphs can be difficult. In the context of the spina bifida example above, a link from say, B to D, would require the elicitation of a prior distribution over the 2×2 table spanned by B and D. Madigan & Raftery (1991) describe in detail an alternative approach whereby the prior distribution is elicited in the context of a directed graph and subsequently transformed into a prior distribution for the undirected model. Madigan *et al.* (1995) describe yet another approach and assess its predictive performance.

The key point is that the elicitation of informative prior distributions, while not without its difficulties, *is* possible. This is not the case for the equivalent distributions in the conventional log linear framework, where the elicitation of prior distributions for the "u"-parameters is, in our experience, impossible.

4.2 *Example : Spina Bifida*

The results of a Bayesian graphical model analysis of the spina bifida example of Table 3 are given in Table 4 and Figure 6. In the figure, the value of

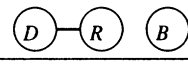
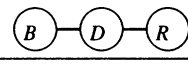
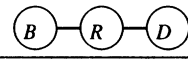
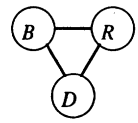
$$P(N \mid D, M_k)P(M_k \mid D)$$

is plotted for any model M_k with non-negligible posterior probability. These curves show both the shape of the posterior distribution of N for particular models, and, by the area beneath them, their relative contribution to the overall posterior distribution. The sum of these curves gives the full posterior, averaged over all models.

In this analysis, uniform prior distributions were adopted for the components of θ under the largest

Table 4

Summaries of the posterior distributions of N for the spina bifida data for all models with posterior probability greater than 0.01. $\hat{N}$ is a Bayes estimate, minimizing a relative squared error loss function

Model	Posterior Prob.	$\hat{N}$	2.5%, 97.5%
	0.373	731	(701, 767)
	0.301	756	(714,811)
	0.281	712	(681,751)
	0.036	697	(628,934)
Model Averaging	—	731	(682,797)

model, and all models were assumed equally likely *a priori*. An informative prior distribution, based on historical data, was adopted for N. According to the Center for Disease Control (Human Services Research Institute, 1985), the incidence of non-anencephalic spina bifida in the US was 7.2 per 10,000 births for the period of 1970–73; additionally, the incidence of encephalocele was 1.2 per 10,000. Combining this with number of live births for the period given by Hook *et al.* (1980), we get an estimate of 725 expected cases. We let pr(N) be Poisson(λ), where λ has a gamma distribution with mean 725 and a standard deviation of 200, to reflect our prior uncertainty. For further details we refer the reader to York & Madigan (1992).

York & Madigan (1992) describe a detailed coverage analysis which shows that our analysis provides prediction intervals which are better calibrated than those based on the standard approach.

The posterior means and standard deviations for the probability that an individual will be found via any particular list are given in Table 5. It is awkward to come up with such "efficiency" estimates in the conventional log linear modeling framework, and even more complex to derive their standard errors. In contrast, the methods described here, directly and easily produce efficiency estimates for each list.

Table 5

Posterior mean and standard deviation for the probability that a given list will correctly identify an individual with spina bifida

List	Posterior Mean	Posterior Std. Dev
Birth Certificate	0.699	0.032
Death Certificate	0.284	0.020
Medical Records	0.258	0.019

If we use our estimate of N to compute a prevalence rate for spina bifida for the population of all live births, we arrive at an estimate of 0.847 per 1000 births, with 2.5th and 97.5th posterior percentiles being 0.790, 0.923. In comparison to the estimate of 0.725 per thousand if we assume that no cases were missed, there is substantial evidence that more than one case per ten thousand is

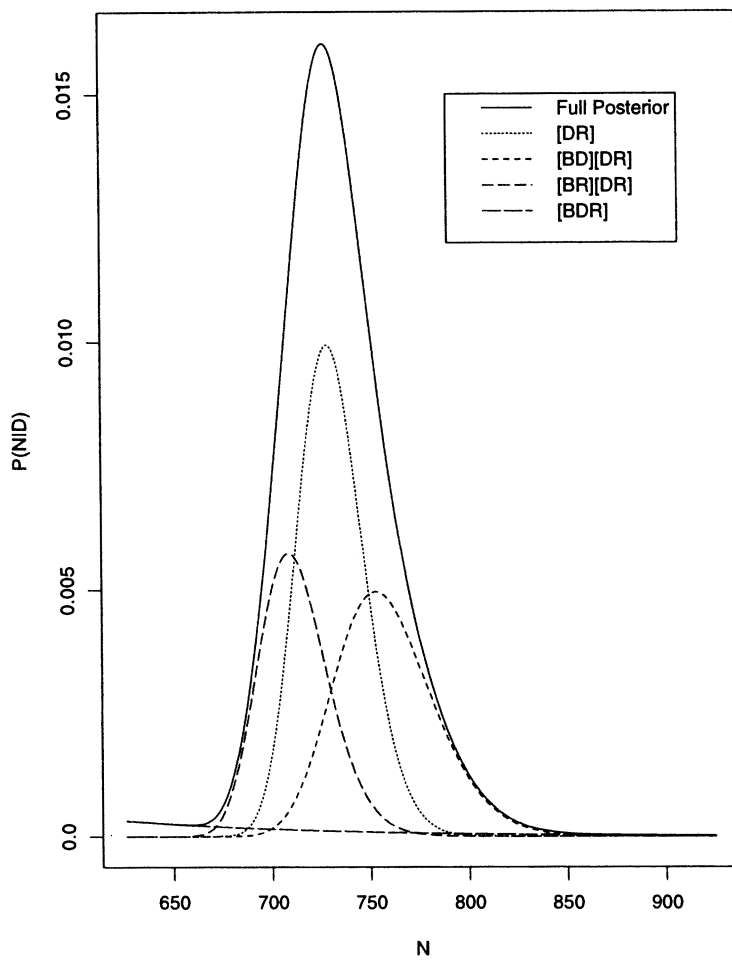


Figure 6. *Posterior distribution for the number of cases of spina bifida for different models. "Full Posterior" shows the posterior distribution averaged over all the decomposable Bayesian graphical models*

missed; and this is for a population count based on three separate lists. The estimate of 0.699 for the efficiency of birth certificates alone indicates that around 30% of the total cases would be overlooked if that registry were the sole source of information.

4.3 Why Bayesian graphical models?

Bayesian graphical models allow for flexible modeling of inter-list dependencies together with an effective medium to communicate these dependencies, i.e., a graph. Furthermore, informative expert knowledge can be expressed directly in terms of well-understood quantities, and distributions for other quantities of interest such as list efficiencies are easily computed. York & Madigan (1992) consider the spina bifida example with the addition of a covariate, namely race. Their analysis is similar to that presented above, despite the missing covariate values for five cases. A similar analysis with log linear models is possible but is analytically challenging.

The restriction to decomposable models may be a real concern for some applications. For the spina bifida example, York & Madigan (1992) show that inclusion of the non-decomposable no-third-order-interaction log linear model has little impact on the results.

5 Discussion

We have attempted to show that Bayesian graphical models represent a powerful unified framework for a variety of discrete data problems. Modeling assumptions are entirely transparent and computations are simple to program. Expert knowledge can easily be incorporated. We describe further applications in Madigan & Raftery (1994), Madigan *et al.* (1994a,b), and Carlsen *et al.* (1994).

One note of caution: the Markov chain Monte Carlo model averaging method outlined here may run into some practical difficulties in the analysis of *very* large datasets with missing values. The essential problem is that possible realizations of the missing data convey considerable information about the best models. Consequently, their joint distribution, $\mathrm{pr}(M, Z \mid Y)$ can be highly multimodal. Philips & Smith (1994) present an alternative approach which we are investigating.

The methods we discuss can readily be extended in two particular directions. First, graphical Gaussian models could be included. These were introduced as covariance selection models by Dempster (1972) and are discussed in Whittaker (1990). The variables being modeled in a graphical Gaussian model have a multivariate normal distribution. Conditional independences, which correspond to zeroes in the inverse variance, are represented by an undirected graph. The Bayesian framework for these models has been developed by Dawid & Lauritzen (1993). Recent extensions to this model class described by Cox & Wermuth (1993) are of considerable interest in this context.

Second, the graphs we consider here are either undirected or fully directed. The methods could be extended to include chain graphs (Frydenberg, 1990) which include acyclic directed and decomposable undirected graphs as special cases. A very valuable development would be to also include the mixed discrete/continuous models of Wermuth & Lauritzen (1990) and Edwards (1990).

Appendix I: Graph Theoretic Terminology

The terminology we use is largely adapted from Lauritzen *et al.* (1990).

A *graph* is a pair $G = (V, E)$ where V is a finite set of vertices and the set of edges, E, is a subset of $V \times V$ of ordered pairs of distinct vertices. Edges $(\alpha, \beta) \in E$ with both (α, β) and (β, α) in E are called *undirected*, whereas an edge (α, β) with its opposite (β, α) not in E is called *directed*.

If the graph has only undirected edges it is *undirected* and if all the edges are directed, the graph is said to be *directed.* Our graphs are either directed or undirected.

If $A \subseteq V$ is a subset of the vertex set, it induces a subgraph $G_A = (A, E_A)$, where the edge set $E_A = E \cap (A \times A)$ is obtained from G by keeping edges with both endpoints in A.

A graph is *complete* if all vertices are joined by an edge. A subset is *complete* if it induces a complete subgraph. A complete subset that is not contained within another complete subset is called a *clique.*

In a directed graph, if $(\alpha, \beta) \in E$, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted by $\text{pa}(\beta)$ and the set of children by $\text{ch}(\beta)$.

In an undirected graph, if $(\alpha, \beta) \in E$, α and β are said to be *adjacent* or *neighbours.* The *boundary*, $\text{bd}(A)$, of a subset A of vertices is the set of vertices in $V \backslash A$ that are neighbours to vertices in A.

A *path* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. If there is a path from α to β we say that α *leads to* β and write $\alpha \mapsto \beta$. The *descendants* $\text{de}(\alpha)$ of α are all the vertices β such that α *leads to* β. The *nondescendants* are $\text{nd}(\alpha) = V \backslash (\text{de}(\alpha) \cup \{\alpha\})$.

A subset S is said to *separate* A *from* B if all chains from vertices $\alpha \in A$ to $\beta \in B$ intersect S.

A *cycle* is a path with the modification that $\alpha = \beta$, i.e. it begins and ends at the same point. A directed graph is *acyclic* if it contains no cycles. An undirected graph is *chordal* if it contains no cycles of length ≥ 4 without a chord (i.e two non-consecutive vertices that are neighbours).

An ordering of the cliques of an undirected graph, say $(C_1, \ldots, C_n)$ is said to be *perfect* if the vertices of each clique C_i also contained in previous cliques $(C_1, \ldots, C_{i-1})$ are all members of *one* previous clique. These sets $S_i = C_i \cap (\cup_{j=1}^{i-1} C_j)$ are called *clique separators.* An undirected graph admits a perfect ordering of its cliques if and only if it is chordal.

References

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*, Cambridge, Mass.: MIT Press.

Carlsen, J.C., Bradshaw, D.H., Madigan, D. & Unyk, A.M. (1994). Music Expectancy and its measurement. *Second International Conference on Music Perception and Cognition*, Los Angeles (also submitted for publication).

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, **12:4**, 50–63.

Cooper, G.F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Cox, D.R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, **8**, 204–283.

Dawid, A.P. & Lauritzen, S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models *Annals of Statistics*, **21**, 1272–1317.

Dawid, A.P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, **2**, 25–36.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society (Series B)*, **57**, 45–97.

Edwards, D. (1990). Hierarchical mixed interaction models. *Journal of the Royal Statistical Society (Series B)*, **52**, 3–20.

Ekholm, A. & Palmgren, J. (1987). Correction for misclassification using doubly sampled data. *Journal of Official Statistics*, **3**, 419–429.

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 591–603.

Frydenberg, M. & Lauritzen, S.L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, **76**, 539–555.

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–353.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Heckerman, D., Geiger, D. & Chickering, D.M. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference* (R. Lopez de Mantaras and D. Poole, eds.), San Francisco: Morgan Kaufmann, pp. 293–301.

Hook, E.B., Albright, S.G. & Cross, P.K. (1980). Use of Bernoulli census and log-linear methods for estimating the prevalence of Spina Bifida in livebirths and the completeness of vital record reports in New York state. *American Journal of Epidemiology*, **112**, 750–758.

Kornfeld, A. (1991). Causal networks: Clarifying uncertainty. *AI Expert*, November, 42–49.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. & Leimer, H-G. (1990). Independence properties of directed markov fields. *Networks*, **20**, 491–505.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion.) *Journal of the Royal Statistical Society (Series B)*, **50**, 157–224.

Lauritzen, S.L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, **17**, 31–57.

Lie, R.T., Heuch, I. & Irgens, L.M. (1991). A temporary increase of Down syndrome among births of young mothers in Norway: An effect of risk unrelated to maternal age?. *Genetic Epidemiology*, **8**, 217–230.

Lie, R.T., Heuch, I. & Irgens, L.M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration schemes. *Biometrics*, **50**, 433–444.

Madigan, D. & Mosurski, K.R. (1990). An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika*, **77**, 315–319.

Madigan, D. & Raftery, A.E. (1991). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Technical Report 213*, Department of Statistics, University of Washington.

Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M. & Almond, R.G. (1994a). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IV* (P. Cheeseman and R.W. Oldford, eds.), New York: Springer-Verlag, pp. 91–100.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M. & Almond, R.G. (1994b). Bayesian graphical models for predicting errors in databases. In *Selecting Models from Data: AI and Statistics IV* (P. Cheeseman and R.W. Oldford, eds.), New York: Springer-Verlag, pp. 123–132.

Madigan, D., Gavrin, J. & Raftery, A.E. (1995). Eliciting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models. *Communications in Statistics: Theory and Methods*, **24**, to appear.

Morgan, M.G. & Herion, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*, Cambridge University Press.

Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert Systems*, New York: Wiley.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Mateo, California.

Philips, D.B. & Smith, A.F.M. (1994). Bayesian model comparison via jump diffusions. *Technical Report 94-20*, Department of Mathematics, Imperial College.

Regal, R.R. & Hook, E.B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, **10**, 717–721.

Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society (Series B)*, **55**, 3–23.

Spiegelhalter, D.J. & Lauritzen, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Spiegelhalter, D.J. & Cowell, R.G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) pp. 447–466. Oxford: Clarendon Press.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–283.

Thompson, E.A. (1986). *Pedigree analysis in human genetics*, Baltimore: The Johns Hopkins University Press.

Wermuth, N. & Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society (Series B)*, **52**, 21–50.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

York, J. & Madigan, D. (1992). Bayesian methods for estimating the size of a closed population. Technical Report 234, Department of Statistics, University of Washington. Under revision at *Biometrika*.

York, J., Madigan, D., Heuch, I. & Lie, R.T. (1995). Birth defects registered by double sampling: A Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, **44**, 227–242.

Résumé

Pendant plus d'un demi-siècle, les graphes ont été utilisés pour représenter des modéles statistiques en analyse de donneés. En particulier, les graphes décrivant l'indépendance conditionnelle sont apparus comme une classe importante de ces modèles. Des applications en analyse d'image, analyse de pedigree, ou encore en système expert sont à l'origine de leur développement, et plusieurs livres de synthèse ont déjà été publiés à ce sujet.

Le développement d'un cadre Bayésien de ces modèles est en revanche moins connu, et les applications en systèmes experts ont motivé la recherche dans ce domaine. La possibilité de construire des modèles capables de se remettre à jour au fur et à mesure que de nouvelles données sont disponibles est à l'origine d'un intérêt intense de la part de la communauté travaillant en intelligence artificielle. Cependent, leur application a une classe plus vaste d'analyse de données a été largement

négligée.

L'objet de cet article est de montrer comment les modèles Bayésiens de graphes permettent d'unifier et de simplifier des problèmes standards, tels que les modèles log-linéaires Bayésiens (avec des données complètes ou non), l'estimation d'une population fermée ou le double échantillonnage. Dans la mesure ou le choix d'un modèle conventionel unique échoue dans ce type de situation, nous construisons des distributions *a posteriori* des quantités d'intérêt en moyennant sur les modèles possibles. Plus particulièrement, nous introduisons la composition de chaînes de Markov–Monte Carlo, une méthode de Monte–Carlo permettant de moyenner sur les modèles retenus.

[*Received May 1993, accepted January 1995*]